Société statistique du Canada
Statistical Society of Canada

51st Annual Meeting of the Statistical Society of Canada

51e Congrès annuel de la Société statistique du Canada

June 2 – June 5, 2024
2 juin au 5 juin 2024

Memorial University of Newfoundland, St. John's

Table of Contents • Table des matières

Welcome to St. John's and to Memorial University • Bienvenue à St. John's et à l'Université Memorial

The department of Mathematics and Statistics at Memorial University welcomes you to St. John's! We are thrilled to be hosting the SSC annual meeting once again.

Land Acknowledgement

We acknowledge that the lands on which Memorial University's campuses are situated are in the traditional territories of diverse Indigenous groups, and we respectfully reflect upon the diverse histories and cultures of the Beothuk, Mi'kmaq, Innu, and Inuit of this province.

Stellar Location

St. John's is the capital and largest city of the Canadian province of Newfoundland and Labrador. It is located on the eastern tip of the Avalon Peninsula on the island of Newfoundland. St. John's, the oldest city in North America, is a tourist city with its vibrant cultural and social life, great coastal walking trails, the scene of fabulous folk and rock festivals and home to an annual Regatta that dates to 1816. Every year, May –June is well known as iceberg season. Roughly 90% of icebergs seen off Newfoundland and Labrador come from the glaciers of western Greenland, while the rest come from glaciers in Canada's Arctic. NL is one of the most spectacular whales watching places on Earth. There are lot of tourist attractions around St. John's, including Signal Hill National Historical Site and Cape Spear Light House Historical Site. You may get more information from our help desk at registration.

About Memorial University

Memorial University of Newfoundland (MUN) is one of Atlantic Canada's top universities in terms of teaching and research. As Newfoundland and Labrador's only university, Memorial has a special obligation to the people of this province. Established as a memorial to the Newfoundlanders who lost their lives on active service during the First World War and subsequent conflicts, Memorial University draws inspiration from these sacrifices of the past as we help to build a better future for our province, our country and our world.

About Department of Mathematics and Statistics

The Department of Mathematics and Statistics has 40 faculty members, half of them have been hired since 2005, and be proud that 20% of us hold the university's highest rank, University Research Professor. We offer courses leading to gen-

Le Département de mathématiques et de statistique de l'Université Memorial vous souhaite la bienvenue à Saint-Jean de Terre-Neuve ! Nous sommes ravis d'accueillir à nouveau le congrès annuel de la SSC.

Reconnaissance territoriale

Nous reconnaissons que les terres sur lesquelles sont situés les campus de l'Université Memorial se trouvent sur les territoires traditionnels de divers groupes autochtones, et nous réfléchissons avec respect aux diverses histoires et cultures des Béothuks, des Mi'kmaqs, des Innus et des Inuits de cette province.

Une situation exceptionnelle

St John's est la capitale et la plus grande ville de la province canadienne de Terre-Neuve-et-Labrador. Elle est située à l'extrémité est de la péninsule d'Avalon, sur l'île de Terre-Neuve. Saint-Jean de Terre-Neuve, la plus ancienne ville d'Amérique du Nord, est une ville touristique avec une vie culturelle et sociale animée, de superbes sentiers de randonnée côtière, de fabuleux festivals de folk et de rock et une régate annuelle qui remonte à 1816. Chaque année, les mois de mai et juin sont connus pour être la saison des icebergs. Environ 90 % des icebergs observés au large de Terre-Neuve-et-Labrador proviennent des glaciers de l'ouest du Groenland, tandis que le reste provient des glaciers de l'Arctique canadien. Terre-Neuve-et-Labrador est l'un des endroits les plus spectaculaires au monde pour l'observation des baleines. Il existe de nombreuses attractions touristiques autour de Saint-Jean, dont le lieu historique national de Signal Hill et le Lieu historique national du Phare-de-Cap-Spear. Vous pouvez obtenir plus d'informations auprès du bureau des inscriptions.

À propos de l'Université Memorial

L'Université Memorial de Terre-Neuve (MUN) est l'une des meilleures universités du Canada atlantique en termes d'enseignement et de recherche. En tant que seule université de Terre-Neuve-et-Labrador, Memorial a une obligation particulière envers les habitants de cette province. Créée en mémoire des Terre-Neuviens qui ont perdu la vie en service actif pendant la Première Guerre mondiale et les conflits ultérieurs, l'Université Memorial s'inspire de ces sacrifices du passé pour contribuer à bâtir un avenir meilleur pour notre province, notre pays et notre monde.

À propos du Département de mathématiques et de statistique

Le Département de mathématiques et de statistique compte 40 professeurs, dont la moitié a été recrutée depuis 2005, et vous pouvez être fiers que 20 % d'entre nous occupent le rang le

eral and honors undergraduate degrees in both the Faculty of Science and the Faculty of Arts, with concentrations in applied mathematics, pure mathematics and statistics. We also offer a variety of graduate degrees in both mathematics and statistics.

The Department of Mathematics and Statistics has 40 faculty members of which 25% belong to Statistics group. We offer courses leading to general and honors undergraduate degrees in both the Faculty of Science and the Faculty of Arts, with concentrations in applied mathematics, pure mathematics and statistics. We also offer a variety of graduate degrees in both mathematics and statistics including Two Year Research based MSc in Mathematics and Statistics, One year Master of Applied Statistics, One Year Master of Data Science (jointly with Computer Science) and Ph.D Mathematics and Statistics. We also offer major and minor in undergraduate degree programs in Mathematics and Statistics.

plus élevé de l'université, celui de professeur de recherche universitaire. Nous proposons des cours menant à des diplômes de premier cycle généraux et spécialisés dans les Facultés des sciences et des lettres, avec des concentrations en mathématiques appliquées, mathématiques pures et statistique. Nous proposons également une variété de diplômes de troisième cycle en mathématiques et en statistique.

Le Département de mathématiques et de statistique compte 40 professeurs, dont 25 % appartiennent au groupe de statistique. Nous proposons des cours menant à des diplômes de premier cycle généraux et spécialisés dans les Facultés des sciences et des lettres, avec des concentrations en mathématiques appliquées, mathématiques pures et statistique. Nous proposons également une variété de diplômes d'études supérieures en mathématiques et en statistique, notamment une maîtrise de recherche de deux ans en mathématiques et en statistique, une maîtrise en statistique appliquée d'un an, une maîtrise en science des données d'un an (conjointement avec le Département d'informatique) et des doctorats en mathématiques et en statistique. Nous proposons également des majeures et des mineures en mathématiques et en statistique dans les programmes de premier cycle.

Message from the SSC President • Message de la Presidente de la SSC

Dear Colleagues, students, friends and participants:

On behalf of the Program Committee and the Local Arrangements Committee, I am delighted to welcome you to the annual meeting of the Statistical Society of Canada. I believe that everyone is excited to reconnect with colleagues and friends, to learn about new ideas and research methods, and to meet new people. There will be a lot of social and scientific activities: we have more than 110 sessions and more than 400 speakers, in addition to 7 workshops, 2 case studies, social events, a banquet, and 2 award ceremonies. As usual, we will have plenary talks by the Presidential Invited Speaker and many award winners. On behalf of the SSC, I would like to thank the many volunteers and workers without whom this annual meeting would not be possible. I would also like to thank the sponsors who supported the conference. Finally, I wish you all an exciting and successful meeting, both socially and scientifically.

Shirley Mills
SSC President

Chers collègues, étudiants, amis et participants :

Au nom du comité du programme et du comité des arrangements locaux, je suis ravi de vous accueillir à la réunion annuelle de la Société statistique du Canada. Je suis certain que tout le monde est excité de renouer avec des collègues et amis, d'apprendre de nouvelles idées et méthodes de recherche et faire de nouvelles connaissances. Il y aura beaucoup d'activités sociales et scientifiques : nous avons plus de 110 séances et plus de 400 conférenciers, en plus de 7 ateliers, 2 études de cas, et des événements sociaux, un banquet, et 2 cérémonies de remise des prix. Comme d'habitude, nous aurons une conférence plénière par l'invité de la présidente ainsi que de nombreux lauréats. Au nom de la SSC, je tiens à remercier les nombreux bénévoles et travailleurs sans qui cette assemblée annuelle ne serait pas possible. Je tiens également à remercier les commanditaires qui ont soutenu cette conférence. Enfin, je vous souhaite à tous une réunion passionnante et fructueuse, tant sur le plan social que scientifique.

Shirley Mills
Présidente de la SSC

Sponsors • Commanditaires

The Statistical Society of Canada would like to thank each of the sponsors, whose generous contributions have madethis conference possible:

La Société statistique du Canada désire remercier chacun de ses commanditaires dont les généreuses contributions ont rendu possible la tenue de ce congrès :

Platinum Sponsors • Commanditaires platine

- Memorial University Conference and Event Services

- Memorial University Faculty of Science

- Canadian Statistical Sciences Institute • Institut canadien des sciences statistiques

- Natural Sciences and Engineering Research Council of Canada/Conseil de recherches en sciences naturelles et en génie du Canada

Silver Sponsors • Commanditaires d'argent

- Werklund School of Education, University of Calgary

- Department of Critical Care Medicine, University of Calgary

Organizers • Organisateurs

Scientific Program Committee • Comité du programme scientifique

- Tessema Astatkie, (Chair • Président) Dalhousie University
- Milena Kurtinecz, Bayer Pharmaceuticals
- Luke Hagar, University of Waterloo
- Alexandru Badescu, University of Calgary
- Jinko Graham, Simon Fraser University
- Farouk Nathoo, University of Victoria
- Matthew Greenberg, University of Calgary
- Thomas Salisbury, York University
- Yildiz Yilmaz, Memorial University of Newfoundland
- Éric Gagnon, Statistics Canada

Local Arrangements Committee @ Memorial • Comité des arrangements locaux @ Memorial

- Asokan Mulayath Variyath (Co-Chair • Co-Présidente)
- Zhaozhi Fan (Co-Chair • Co-Président)
- J C Loredo-Osti
- Veeresh Gadag
- Armin Hatefi
- Yanqing Yi
- Alwell Oyet
- Hensley Hubert
- Kunasekaran Nirmalkanna
- Nan Zhang
- Haiyan Yang

It is impossible to organize an event of the size of the Annual Meeting of the SSC without the help of several individuals and organizations. The local arrangements committee would like to thank all those who helped pull this event together. Since it is not easy to name everyone, we sincerely thank all members of the department / group who helped us. They are,

- AV and IT support from Centre for Innovation and Teaching, MUN
- Administrative Staff and Faculty Members of Department of Mathematics and Statistics, MUN
- Entire team of Dean's Office of Faculty of Science, MUN
- Facilities Management at Memorial
- Team members of MUN Conference and Event Services
- MUN Dining Services
- Conference Services at Sheraton Hotel
- MUN Student Residences
- Room Reservation and services at Sheraton Hotel and Holiday Inn
- MUN Campus Enforcement & Patrol
- MUN Parking Services
- City Wide Cab Services
- SSC Office Staff at Ottawa
- SSC Board
- WHOVA
- Sponsors of SSC 2024

Il est impossible d'organiser un événement de l'ampleur du congrès annuel de la SSC sans l'aide de nombreuses personnes et organisations. Le comité local d'organisation tient à remercier tous ceux qui ont contribué à l'organisation de cet événement. Comme il n'est pas facile de nommer tout le monde, nous remercions sincèrement tous les membres du Département / groupe qui nous ont aidés. Il s'agit de

- Soutien audiovisuel et informatique du Centre pour l'innovation et l'enseignement, MUN
- Personnel administratif et membres du corps enseignant du département de mathématiques et de statistique, MUN
- Toute l'équipe du bureau du doyen de la faculté des sciences, MUN
- Gestion des installations à Memorial
- Membres de l'équipe des services de conférence et d'événements de MUN
- Services de restauration de MUN
- Services de conférence à l'hôtel Sheraton
- Résidences des étudiants de MUN
- Réservation de chambres et services à l'hôtel Sheraton et à l'hôtel Holiday Inn
- Application de la loi et patrouille sur le campus de MUN
- Services de stationnement de MUN
- City Wide Cab Services
- Personnel du bureau de la SSC à Ottawa
- Conseil d'administration de la SSC
- WHOVA
- Commanditaires du congrès 2024 de la SSC

General Information • Informations générales

All Attendees at the SSC Annual Meeting are reminded that in attending an SSC event, they agree to adhere to the SSC Code of Conduct.

Il est rappelé à tous les participants au congrès annuel de la SSC qu'en participant à un événement de la SSC, ils acceptent d'adhérer au Code de conduite de la SSC.

Programs and the Whova App • Programmes et l'application Whova

NOTE – To be environmentally conscious, the SSC Board decided to no longer provide printed programs.

We suggest you download the free WHOVA app from either the Google Play Store or the Apple App Store. This app permits you to view the program, set up your personal agenda, and connect and communicate with fellow attendees. Depending on whether your device is set to English or French, you will be able to view the program in either language. You can use TRACKS to sort through the program. All registrants should receive an email about using WHOVA for SSC2024.

Failing that, the SSC meeting website contains the entire program with and without abstracts and you have the option of printing your program from there.

NOTE - Dans un souci de respect de l'environnement, le Conseil d'administration de la SSC a décidé de ne plus fournir de programmes imprimés.

Nous vous suggérons de télécharger l'application gratuite WHOVA à partir du Play Store de Google ou de l'App Store d'Apple. Cette application vous permet de consulter le programme, d'établir votre programme personnel, de vous connecter et de communiquer avec les autres participants. Selon que votre appareil est réglé sur l'anglais ou le français, vous pourrez consulter le programme dans l'une ou l'autre langue. Vous pouvez utiliser TRACKS pour trier le programme. Toutes les personnes inscrites devraient recevoir un courriel concernant l'utilisation de WHOVA pour le congrès SSC2023.

A défaut, le site web du congrès de la SSC contient le programme complet avec et sans les résumés ; vous pouvez y imprimer votre programme.

Directions • Directions

Finding your way at MUN • Trouver sa voie à MUN
https://map.concept3d.com/?id=219#!ct/15331?sbc/

MUN Residence Location • Lieu de la résidence MUN
https://www.mun.ca/stay/location/

University campus parking map • Plan de stationnement sur le campus universitaire
https://www.mun.ca/cep/parking/st-johns-campus-parking-maps/

Destination St. Johns 2024 Visitor Guide • Guide touristique Destination St. Johns 2024 (anglais uniquement)
https://destinationstjohns.com/wp-content/uploads/2024/05/2024-DSJ-Visitors-Guide-%E2%80%94-WEB2.pdf

Registration • Inscription

Registered participants must pick up their badges and food / banquet vouchers at the Registration Desk at Bruneau Centre for Research and Innovation during the times indicated below. Walk-ins must register online and present their receipt

Les participants inscrits doivent retirer leurs badges et leurs billets de restauration / banquet au bureau des inscriptions du Centre de recherche et d'innovation Bruneau aux heures indiquées ci-dessous. Les personnes qui se présentent sans préinscription

at the registration desk. If you have already registered for the conference but wish to add a workshop, please contact es-admin@ssc.ca for instructions or come to the registration desk before the workshop on Sunday morning.

The Registration Desk is located in Bruneau Centre for Research and Innovation Atrium. Hours are:

Sunday, June 2: 8:00 a.m. – 5 p.m.

Monday June 3: 8:00 a.m. – 5 p.m.

Tuesday June 4: 8:00 a.m. – 5 p.m.

Wednesday June 5: 8:00 a.m. – 12:00 p.m.

doivent s'inscrire en ligne et présenter leur reçu au bureau des inscriptions. Si vous êtes déjà inscrit au congrès mais que vous souhaitez ajouter un atelier, veuillez contacter es-admin@ssc.ca pour obtenir des instructions, ou vous présenter au bureau des inscriptions avant votre atelier, le dimanche matin.

Dimanche 2 juin : 8h00 – 17h00

Lundi 3 juin : 8h00 – 17h00

Mardi 4 juin : 8h00 – 17h00

Mercredi 5 juin : 8h00 – 12h00

Parking and Transportation • Stationnement sur le campus et transports

There are bus services to the Memorial campus (located near University Centre) operated by Metro Bus but it will be running on a spring schedule.

You may wish to carpool, take a taxi, Uber

On campus parking (pay-per-use) is available following parking lots

- Area 27: Parking Garage - Level 1 (all spaces) and Level 2 (designated P&D spaces on east and west perimeter rows) located on Arctic Avenue.
- Area 61: Earth Sciences Garage (ground floor) located beside the University Centre located on Arctic Avenue.

Location of all pay per use spaces can be viewed at

Le campus Memorial (situé près du centre universitaire) est desservi par Metro Bus, mais les horaires seront ceux du printemps.

Vous pouvez faire du covoiturage, prendre un taxi ou Uber.

Le stationnement sur le campus (payant) est disponible dans les parkings suivants :

- Zone 27 : Garage de stationnement – Niveau 1 (toutes les places) et niveau 2 (places P&D désignées sur les rangées périmétriques est et ouest) situé sur l'avenue Arctic.
- Zone 61 : Garage des sciences de la terre (rez-de-chaussée) situé à côté du centre universitaire sur l'avenue Arctic.

L'emplacement de toutes les places de stationnement payant peut être consulté sur

https://www.mun.ca/cep/media/production/memorial/administrative/campus-enforcement-and-patrol/media-library/parking/User_Pay_Parking.pdf

Campus Security • Sécurité sur le campus

- Fire/Police/Ambulance: 911
- Campus Safety Services:
  - EMERGENCIES: 709-864-4100 (ext. 4100 from any campus phone)
  - Non-Emergencies: 709-864-8561

- Pompiers/Police/Ambulance : 911
- Services de sécurité du campus :
  - URGENCE : 709-864-4100 (poste 4100 à partir d'un téléphone du campus)
  - Sans urgence : 709-864-8561

Internet Access • Accès internet

Memorial University is a member of eduroam (EDUcation ROAMing), an authentication service allowing users (researchers, teachers, students, staff) from participating educational institutions to securely access the wireless network of any eduroam-enabled institution by using the same credentials they would use at their home institution. Connecting through eduroam provides basic network connectivity for web browsing (HTTP), secure shell (SSH) and VPN access. Visitors to Memorial University from eduroam participating institutions can access basic wireless services using their university related credentials. All participants can also access a wireless network @Memorial-Guest and there is no password required.

L'Université Memorial est membre d'eduroam (EDUcation ROAMing), un service d'authentification permettant aux utilisateurs (chercheurs, enseignants, étudiants, personnel) des établissements d'enseignement participants d'accéder en toute sécurité au réseau sans fil de n'importe quel établissement équipé d'eduroam en utilisant les mêmes informations d'identification que celles qu'ils utiliseraient dans leur établissement d'origine. La connexion via eduroam fournit une connectivité réseau de base pour la navigation sur le web (HTTP), le shell sécurisé (SSH) et l'accès VPN. Les visiteurs à l'Université Memorial provenant d'établissements participant à eduroam peuvent accéder aux services sans fil de base en utilisant les informations d'identification de leur université. Tous les participants peuvent également accéder à un réseau sans fil @Memorial-Guest et aucun mot de passe n'est requis.

Food on Campus and nearby • Nourriture sur le campus et en ville

Coffee breaks will take place at 9:50–10:20a.m. and 2:30–3:00p.m. on Monday, Tuesday and Wednesday. Coffee will be available in two locations - Bruneau Centre for Research and Innovation (IIC) Atrium and the Arts & Administration Building (A) Atrium.

Lunch breaks will be at 11:50a.m.–1:30p.m. on Monday, Tuesday and Wednesday. Due to the scarcity of restaurants near the Memorial University, lunch is provided daily as part of your registration at R Gushue Hall (see daily vouchers in your registration package); boxed lunches will be available in Bruneau Centre for Research and Innovation only for those attending lunchtime committee meetings. You may also be able to get food at some of the restaurants on campus at University Centre.

Les pauses café auront lieu de 9h50 à 10h20 et 14h30-15h00 le lundi, mardi et mercredi. Le café sera disponible à deux endroits : l'Atrium du Centre de recherche et d'innovation Bruneau (IIC) et l'Atrium du bâtiment des arts et de l'administration (A).

Les pauses dîner auront lieu entre 11h50 et 13h30 le lundi, mardi et mercredi. En raison de la rareté des restaurants à proximité de l'Université Memorial, le dîner est fourni tous les jours dans le cadre de votre inscription au Hall R Gushue (voir les billets quotidiens dans votre trousse de participant) ; des boîtes à lunch seront disponibles au Centre Bruneau pour la recherche et l'innovation uniquement pour les délégués qui participent aux réunions des comités à l'heure du dîner. Vous pourrez aussi peut-être vous restaurer dans certains des restaurants du campus au Centre universitaire.

Athletics Facilities • Installations sportives

Memorial University offers state-of-the art athletics facilities at the Field House. Information on our fitness centre is available at

L'Université Memorial offre des installations sportives de pointe au Field House. Vous trouverez plus d'informations sur notre centre de fitness à l'adresse suivante :

https://www.theworksonline.ca/fitness-field-house-and-memberships/field-house/.

The Conference • Le congrès

Language • Langue

An important feature of our meetings is the presentation of the abstracts and of the plenary session visual aids in both official languages. This translation was once again very ably carried out under the supervision of the Bilingualism Committee (chaired by Thierry Chekouo Tekougang) by the translators Catherine Cox, Caroline Gras, Michelle Blaquière and Olivier Tremblay.

At the time that they submitted their abstract, speakers were asked to provide the language in which they intend to give their oral presentation as well as the language of their visual aids. Icons are used to provide this information for each paper. For the oral presentation, we use the icons Ⓔ and Ⓕ, whereas E and F indicate the language of the visual aids. The letter inside identifies the language: E for English and F for French. Please note that the visual aids for the plenary talks will be provided in both languages.

Une caractéristique importante de nos congrès est la présentation des résumés et des supports visuels des sessions plénières dans les deux langues officielles. Cette traduction a été encore une fois très habilement menée sous la supervision du Comité du bilinguisme (présidé par Thierry Chekouo Tekougang) par les traducteurs Catherine Cox, Caroline Gras, Michelle Blaquière et Olivier Tremblay.

Lorsque les conférenciers ont soumis leur résumé, ils ont spécifié la langue dans laquelle ils comptaient faire leur présentaton orale, ainsi que la langue du support visuel. À titre informatif, nous avons inclus cette information à l'aide d'icônes pour chaque présentation. Pour la présentation orale nous avons utilisé les icônes Ⓔ et Ⓕ, tandis que E et F indiquent le support visuel. La lettre à l'intérieur identifie la langue : F pour le français et E pour l'anglais (English). Veuillez noter que le support visuel des conférences plénières sera présenté dans les deux langues.

Bilingualism Committee • Comité du bilinguisme

- Thierry Chekouo Tekougang (Chair • Président), University of Calgary
- Denis Talbot, Université Laval
- Sarah-Anne Savard, Statistics Canada
- Marie-Hélène Descary, Université du Québec à Montréal
- Claude Girard, Statistics Canada
- Anne-Sophie Charest, Université Laval
- Ismaila Baldé, Université de Moncton

Locations of the Conference • Lieux de la conférence

All SSC2024 talks are planned to be in Chemistry - Physics (C), Arts and Administration Building (A), Education Building (ED) and - Bruneau Centre for Research and Innovation (IIC – formerly known as Inco Innovation Centre). (Note that all buildings are connected by a tunnel system which you may wish to use if inclement weather)

Toutes les sessions du congrès SSC2024 se dérouleront dans les bâtiments suivants : Chimie - Physique (C), Arts et Administration (A), Éducation (ED) et Centre de recherche et d'innovation Bruneau (IIC – anciennement Centre d'innovation Inco). (Notez que tous les édifices sont reliés par un système de tunnels que vous pouvez utiliser en cas de mauvais temps).

Poster Sessions and Case Studies • Séances d'affichage et les Études de cas

Contributed posters and Case Study posters will be displayed in Core Science Facility (CSF) Atrium. All the case study posters will be displayed between noon and 4:00 p.m., the authors being with their posters from 13:30 until 15:00. The contributed poster presentations are scheduled for Tuesday, 13:30-15:00.

Les affiches libres et des études de cas seront exposées dans l'atrium de la Core Science Facility (CSF). Les affiches d'études de cas seront exposées entre midi et 16h00, les auteurs étant présents de 13h30 à 15h00. Les présentations des affiches libres sont prévues le mardi, de 13h30 à 15h00.

Presentation of Awards • Remise des prix

Presentation of the predetermined SSC Awards for 2024 (Gold Medal, Honorary Memberships, Distinguished Service, CRM-SSC, Canadian Journal of Statistics Best Paper, Pierre Robillard, Impact of Applied and Collaborative Work, Distinguished Educator, Early Career Educator, Lise Manchester) will be presented at the Banquet on Monday June 2. Awards for student presentations at the SSC 2024 Conference will be presented in a special session on Wednesday June 4 at 9:30am in IIC 2001, immediately following the presentation by the 2023 SSC Gold Medal winner. These awards include all of the Student Presentation Awards (including the competitions put on by SSC Sections) and the Case Studies in Data Analysis Awards. Please join us to honour those young researchers who win these important awards. The New Investigator Presentation Award will be announced in Liaison at a later date.

La remise des prix prédéterminés de la SSC pour 2024 (Médaille d'or, membre honoraire, services insignes, CRM-SSC, meilleur article de La revue canadienne de statistique, Pierre-Robillard, impact du travail appliqué et collaboratif, prix d'excellence en enseignement pour professeur.e ou chercheur.e, prix d'enseignement pour professeur.e ou chercheur.e en début de carrière, Lise-Manchester) sera présentée lors du banquet du lundi 2 juin. Les prix pour les présentations d'étudiantes au congrès 2024 de la SSC seront remis lors d'une session spéciale le mercredi 4 juin à 9h30 dans la salle IIC 2001, immédiatement après la présentation du récipiendaire de la Médaille d'or de la SSC 2023. Ces prix incluent tous les prix pour les présentations de recherche étudiantes (y compris les concours organisés par les Groupes de la SSC), et les prix d'études de cas en analyse de données. Joignez-vous à nous pour rendre hommage aux jeunes chercheurs qui remportent ces prix importants. Le Prix pour la présentation d'un nouveau chercheur sera annoncé dans Liaison à une date ultérieure.

Workshops • Ateliers

Workshops organized by the Sections will be held on Sunday June 2 in rooms in Arts and Administration (A) and Henrietta Harvey (HH – Mathematics). Refreshment breaks will take place in Bruneau Centre for Research and Innovation Atrium between 10:15-10:45 a.m. and 2:15-2:45 p.m. Lunch for Workshops will take place in R Gushue Hall between noon and 1:00 p.m.

Les ateliers organisés par les sections se tiendront le dimanche 2 juin dans les salles Arts et administration (A) et Henrietta Harvey (HH - Mathématiques). Les pauses rafraîchissement auront lieu dans l'atrium du Centre Bruneau pour la recherche et l'innovation entre 10h15 et 10h45 et entre 14h15 et 14h45. Le dîner pour les ateliers aura lieu dans le Hall R Gushue entre midi et 13 heures.

Other Meetings • Autres réunions

NSERC Discovery Grant Information Session: Tuesday June 4, 15:30-17:00 in room A-1049

This workshop will be presented by NSERC Research Grants staff and will cover the Notification of Intent to Apply (NOI) and Full Application process, the Discovery Grant evaluation process principles (criteria and ratings), the Conference Model and tips for preparing a Discovery Grant application. Following the Workshop, there will be an opportunity for participants to ask questions.

LaTeX Class Presentation (Canadian Journal of Statistics): Tuesday June 4 5:00-6:00 p.m. in room SN 2109

This special presentation will be the opportunity to unveil the new class 'cjs-rcs-article', give a tour of its features, and explain how to quickly get started for your next article! The presentation will be given by Vincent Goulet, Université Laval who was developed the new class.

Séance d'information sur les subventions à la découverte du CRSNG : mardi 4 juin, de 15h30 à 17h00 dans la salle A-1049

Cet atelier, présenté par le personnel des subventions à la découverte du CRSNG, couvrira l'Avis d'intention de présenter une demande de subvention à la découverte et le processus de demande détaillée, les principes du processus d'évaluation des subventions à la découverte (critères et cotes), le modèle de conférence et présentera certains conseils pour la préparation d'une demande de subvention à la découverte. À la fin de l'atelier, les participants seront invités à poser leurs questions.

Présentation de la classe LaTeX (La revue canadienne de statistique) : mardi 4 juin 17h - 18h, salle Science SN 2109.

Cette présentation spéciale sera l'occasion de dévoiler la nouvelle classe « cjs-rcs-article », de faire un tour d'horizon de ses fonctionnalités, et d'expliquer comment rapidement démarrer votre prochain article ! La présentation sera donnée par Vincent Goulet, de l'Université Laval qui a développé la nouvelle classe.

Social Events • Activités sociales

Welcome Reception • Réception de bienvenue

Sunday, June 2, 6:00 - 8:00 pm • Dimanche 2 juin, 18h00 - 20h00

The Welcome Reception will be held in Core Science Facility Whale Atrium. All conference attendees are welcome to join us to share a drink and some appetizers in good company. One drink ticket will be given to all registrants for the reception in their registration badge. For those who will be arriving directly from the workshops it is a five minute walk from the workshop meeting rooms.

La réception de bienvenue se tiendra dans l'Atrium des baleines de la Core Science Facility. Tous les participants à la conférence sont invités à se joindre à nous pour partager un verre et quelques amuse-gueules en bonne compagnie. Un ticket boisson sera remis à toutes les personnes inscrites à la réception dans leur badge d'inscription. Pour ceux qui arrivent directement des ateliers, la réception se trouve à cinq minutes à pied des salles de réunion.

Banquet •

Monday, June 3, 6:00pm – 10:30pm • Lunde 3 juin, 18h00 – 22h30 ..... Fort William Ballroom

The banquet will be held at Sheraton Hotel with a cash bar/cocktail at 18:00 and dinner service beginning promptly at 19.00. Due to capacity restrictions, we could not provide banquet ticket to all the registered participants. All conference participants who have selected a meal for the banquet will find one banquet ticket with their selected meal in their registration badge. You must place the banquet ticket on the table in front of you when your table is served. *Please notify your server of any food allergies.*

Le banquet se tiendra à l'hôtel Sheraton, avec un bar/cocktail à 18h00 et un dîner qui débutera à 19h00. En raison des restrictions de capacité, nous n'avons pas pu fournir de billet de banquet à tous les participants inscrits. Les participants à la conférence qui ont choisi un plat pour le banquet trouveront un billet de banquet avec le plat choisi dans leur badge d'inscription. Vous devez placer ce ticket sur la table devant vous lorsque votre table est servie. *Veuillez informer votre serveur de toute allergie alimentaire.*

Shuttle buses will transport conference attendees between MUN to Sheraton Hotel both at the start and end of the banquet. Buses will be available as of 17.30 in front of School of Music (Parking Lot 15); return buses will be available as of 10:30 p.m. in front of the Sheraton Hotel.

Des navettes transporteront les participants à la conférence entre MUN et l'hôtel Sheraton au début et à la fin du banquet. Les autobus seront disponibles à partir de 17h30 devant l'École de musique (stationnement 15) ; les navettes de retour seront disponibles à partir de 22h30 devant l'hôtel Sheraton.

Registration for this year's SSC Annual Conference was overwhelming and we are delighted that so many of you will join us in St. John's. Unfortunately, however, the banquet location was unable to accommodate all registrants. Those of you who registered after February 22 were unable to select a banquet ticket option and have been put on a waiting list (in order of registration).

Les inscriptions au congrès annuel de la SSC de cette année ont été très nombreuses et nous sommes ravis qu'autant d'entre vous se joignent à nous à Saint-Jean. Malheureusement, le lieu du banquet ne peut pas accueillir toutes les personnes inscrites. Ceux d'entre vous qui se sont inscrits après le 22 février n'ont pas pu choisir une option de billet de banquet et ont été placés sur une liste d'attente (par ordre d'inscription).

We have a very short time window in which to accommodate waitlist people due to the fact that this year's banquet is on Monday rather than Tuesday evening. In order to not have empty seats at the Monday banquet when there are people who would like to attend but do not have tickets, we ask anyone who selected a banquet ticket:

Nous disposons d'un délai très court pour répondre aux besoins des personnes inscrites sur la liste d'attente, étant donné que le banquet de cette année aura lieu le lundi et non le mardi soir. Afin de ne pas avoir de sièges vides au banquet alors que certaines personnes qui souhaiteraient y assister n'ont pas de billets, nous demandons à toutes les personnes qui ont choisi un billet pour le banquet :

- **Please pick up your registration package before Monday June 3 at 1:30 p.m. NDT.** After that time, we may re-
- **de bien vouloir passer retirer votre trousse de parti-**

move banquet tickets from packages not yet collected from the registration desk.

- **if you no longer need it, return your banquet ticket to the registration desk by 1:30 p.m. (NDT) on Monday June 3 so that someone on the waitlist can use it.**
- **If you will be arriving late on Monday and still intend to go to the banquet, email es-admin@ssc.ca before 1: 30 p.m. (NDT) on Monday June 3** or your ticket may be distributed to those on the waiting list.

We can then have any available banquet tickets ready for pickup during the Monday afternoon coffee break.

Thank you for your co-operation and understanding.

cipant avant le lundi 3 juin à 13h30 HAT. Après cette heure, il est possible que nous ôtions les billets de banquet des trousses qui n'auront pas été récupérées au bureau des inscriptions.

- **si vous n'en avez plus besoin, remettez votre billet de banquet au bureau des inscriptions avant 13h30 (HAT) le lundi 3 juin afin que quelqu'un sur la liste d'attente puisse l'utiliser.**
- **Si vous arrivez tard le lundi et que vous avez l'intention de participer au banquet, envoyez un courriel à es-admin@ssc.ca avant 13h30 (HAT) le lundi 3 juin,** sans quoi votre billet pourra être distribué à ceux qui sont sur la liste d'attente.

Les billets de banquet ainsi rendus disponibles seront alors prêts à être retirés pendant la pause-café du lundi après-midi.

Nous vous remercions de votre coopération et de votre compréhension.

Student Barbeque • Barbecue des étudiants

Tuesday, June 4 6:00 - 8:00 pm • Mardi 4 mai, 18h00 - 20h00 ..... R Gushue Hall

The student BBQ is free for undergraduate and graduate students; advance registration is required. Those who have registered for the BBQ will receive a ticket with their registration badge.

Le barbecue des étudiants est gratuit pour les étudiants de premier et deuxième cycles ; une inscription préalable est nécessaire. Les personnes inscrites au barbecue recevront un billet avec leur badge d'inscription.

Other Social Events • Autres évènements sociaux

Accredited Members Social Event
Tuesday June 4, 5:00–7:00pm at The Breezeway in the Memorial University Centre (UC on the campus map).

Biostatistics Section and CSEB Social Event
Tuesday June 4, 5:00pm–7:00pm at Quintana's de la Plaza, 57 Rowan Street, St. John's NL A1B 2X2

New Investigators Social Event
Tuesday June 4, 5:30pm–7:30pm at Arribas Bar, 57 Rowan Street, St. John's NL A1B 2X2

Événement social pour les membres accrédités
Mardi 4 juin, de 17h00 à 19h00 au Breezeway dans le Memorial University Centre (UC sur le plan du campus).

Événement social du Groupe de biostatistique et de la SCEB

Mardi 4 juin, de 17h00 à 19h00, au Quintana's de la Plaza, 57, rue Rowan, St John's NL A1B 2X2.

Événement sociale des nouveaux chercheurs
Mardi 4 juin, de 17h30 à 19h30 à Arribas Bar, 57, rue Rowan, St John's NL A1B 2X2.

Program • Programme

Sunday June 2 — dimanche 2 juin

10:00-11:00 — **CSF 1203**

SSC Executive Committee Meeting (tentative)/Réunion du comité éxécutif de la SSC (provisoire)

11:00-17:00 — **CSF 1203**

SSC Board of Directors Meeting/Réunion du conseil d'administration de la SSC

Monday June 3 — lundi 3 juin

12:00-13:30 — **A 2071**

Accreditation Committee Meeting/Réunion du comité d'accréditation

---

12:00-13:30 — **A 1049**

Business and Industrial Statistics Section Executive Committee Meeting/Réunion du comité exécutif du Groupe de statistique industrielle et de gestion

---

12:00-13:30 — **A 2065**

Statistical Education Section Executive Committee Meeting/Réunion du comité éxécutif du Groupe d'éducation en statistique

---

12:00-13:30 — **A 1045**

Biostatistics Section Executive Committee Meeting/Réunion du comité éxécutif du Groupe de biostatistique

Tuesday June 4 — mardi 4 juin

12:00-13:30 — **A 1045**

Fundraising Committee Meeting/Réunion du comité de collecte de fonds

---

12:00-13:30 — **A 1049**

Student Research Presentation Award Committee Meeting/Réunion du comité du prix pour les présentations de recherche étudiantes

---

12:00-13:30 — **A 2065**

Student and Recent Graduate Committee Meeting/Réunion du comité des étudiants et diplômés récents

---

12:00-13:30 — **A 2071**

Research Committee Meeting/Réunion du comité de la recherche

Wednesday June 5 — mercredi 5 juin

12:00-13:30 — **A 1045**

CJS Publishing Committee Meeting/Réunion du comité de publication de la RCS

12:00-13:30 **A 2065**

Pierre Robillard Award Committee Meeting/Réunion du comité du prix Pierre-Robillard

12:00-13:30 **A 1049**

Committee on Membership Meeting/Réunion du comité de recrutement

Social Events • Événements sociaux

Sunday June 2 — dimanche 2 juin

18:00-20:00 — **Core Science Facility Whale Atrium**

Welcoming Reception/Réception de bienvenue

Monday June 3 — lundi 3 juin

18:00-22:30 — **Fort William Ballroom, Sheraton Hotel Newfoundland, 115 Cavendish Square**

SSC 2024 Conference Banquet/Banquet de la conférence SSC 2024

Tuesday June 4 — mardi 4 juin

17:00-19:00 — **Quintanas de la Plaza, 57 Rowan Street**

Biostatistics Section and CSEB Social Event/Événement social du Groupe de biostatistique et de la SCEB

17:00-19:00 — **The Breezeway, University Centre (UC)**

Accredited Members Social Event/Événement social pour les membres accrédités

---

17:30-19:30 — **Arribas Bar, 57 Rowan Street**

New Investigators' Social Event/Événement sociale des nouveaux chercheurs

---

18:00-20:00 — **R. Gushue Hall**

Student BBQ/Barbecue d'étudiants

Workshops • Ateliers

Sunday June 2 — dimanche 2 juin

09:00-16:00 Workshop / Atelier **A 1045**

Actuarial Science Workshop
Atelier du Groupe de science actuarielle

09:00-16:00 **Maciej Augustyniak** (Université de Montréal) **Jean-François Bégin** (Simon Fraser University) **Frédéric Godin** (Concordia University) **Hong Li** (University of Guelph)
Exploring synergies between tools for longevity and financial risk management in retirement / Explorer les synergies entre les outils de gestion de la longévité et des risques financiers à la retraite

---

09:00-16:00 Workshop / Atelier **A 1046**

Business and Industrial Statistics Workshop
Atelier du GSIG

09:00-16:00 **Tim Swartz** (Simon Fraser University)
An Introduction to Sports Analytics / Introduction à l'analyse sportive

---

09:00-16:00 Workshop / Atelier **A 1049**

Accreditation Program Workshop
Atelier sur le Programme d'Accréditation

09:00-16:00 **Robert Platt** (McGill University)
Statistician as Expert Witness: Data and the Legal System / Le statisticien en tant que témoin expert : Les données et le système juridique

---

09:00-16:00 Workshop / Atelier **A 2071**

Statistical Education Workshop
Atelier du Groupe d'éducation en statistique

09:00-16:00 **Mireille Schnitzer** (Université de Montréal) **Denis Talbot** (Université Laval)
Teaching Causal Inference: A Workshop for Educators / Enseigner l'inférence causale : atelier pour les éducateurs

---

09:00-16:00 Workshop / Atelier **A 3017**

Probability Workshop
Atelier du Groupe de probabilité

09:00-16:00 **Bouchra Nasri** (École de Santé Publique de l'Université de Montréal)
Introduction to Infectious Disease Modeling / Introduction à la modélisation des maladies infectieuses

---

09:00-16:00 Workshop / Atelier **HH 3017**

Survey Methods Workshop
Atelier du Groupe des méthodes d'enquête

09:00-16:00 **Anne-Sophie Charest** (Université Laval)
Statistical Data Privacy / Confidentialité des données statistiques Ⓔ E

09:00-16:00 Workshop / Atelier **A 2065**

Data Science and Analytics Workshop
Atelier du Groupe de science des données et analytiques

09:00-16:00 **Tiffany Timbers** (University of British Columbia) **Daniel Chen** (University of British Columbia) **G. Alexi Rodríguez-Arelis** (University of British Columbia) **Katie Burak** (University of British Columbia)
How to create and distribute R packages / Comment créer et distribuer des paquets R Ⓔ E

09:00-16:30 Workshop / Atelier **A 1043**

Biostatistics Workshop
Atelier biostatistique

09:00-16:30 **Babette Brumback** (University of Florida)
Fundamentals of Causal Inference: with R / Principes fondamentaux de l'inférence causale : avec R Ⓔ E

Workshop Descriptions • Descriptifs des ateliers

Actuarial Science Workshop • Atelier du Groupe de science actuarielle

Room/Salle: A 1045

Date: Sunday June 2 / dimanche 2 juin

Time/Heure: 09:00-16:00

Maciej Augustyniak (Université de Montréal) **Jean-François Bégin** (Simon Fraser University) **Frédéric Godin** (Concordia University) **Hong Li** (University of Guelph)

Exploring synergies between tools for longevity and financial risk management in retirement

Explorer les synergies entre les outils de gestion de la longévité et des risques financiers à la retraite

This workshop delves into the crucial intersection of longevity and financial risk management in retirement. As individuals increasingly seek to secure their financial futures in an era of extended life expectancies, this workshop offers a unique opportunity to explore the dynamic strategies and tools available for comprehensive retirement planning. The topics covered include longevity risk, tontines and variable annuities. The use of machine learning and data analytics techniques within these topics will also be discussed.

Cet atelier se penche sur l'intersection cruciale de la gestion de la longévité et des risques financiers à la retraite. Alors que les individus cherchent de plus en plus à assurer leur avenir financier à une époque où l'espérance de vie s'allonge, cet atelier offre une occasion unique d'explorer les stratégies dynamiques et les outils permettant une planification complète de la retraite. Les sujets abordés incluent le risque de longévité, les tontines et les annuités variables. Nous aborderons également l'utilisation de l'apprentissage automatique et les techniques d'analyse de données dans ces domaines.

Business and Industrial Statistics Workshop • Atelier du GSIG

Room/Salle: A 1046

Date: Sunday June 2 / dimanche 2 juin

Time/Heure: 09:00-16:00

Tim Swartz (Simon Fraser University)

An Introduction to Sports Analytics

Introduction à l'analyse sportive

Although Moneyball brought sports analytics to the attention of the general public, it has been the introduction of tracking data which has been a game changer for researchers. With tracking data, we have big data where the location of the ball and each player are recorded at high frequencies (say, 10-25 times per second). Such data permit the evaluation of off-the-ball activities, something that is not possible with box score data and event data. The spatio-temporal aspect of tracking data has been exploited in major sports such as soccer, basketball and hockey.
In this Workshop, attendees will bring in their own computer, and will be given a sample game of tracking data from soccer. Attendees will learn how to manage the data, and carry out some basic calculations. Hopefully, they will gain an appreciation of the opportunities that are available with this new source of data.

Bien que Moneyball ait attiré l'attention du grand public sur l'analyse du sport, c'est l'introduction des données de suivi qui a changé la donne pour les chercheurs. Avec les données de suivi, nous disposons de mégadonnées où l'emplacement du ballon et de chaque joueur est enregistré à haute fréquence (par exemple, 10 à 25 fois par seconde). Ces données permettent d'évaluer les activités en dehors du ballon, ce qui n'est pas possible sur la seule base des données de score et événementielles. L'aspect spatio-temporel des données de suivi a été exploité dans des sports majeurs tels que le soccer, le basket-ball et le hockey.
Dans cet atelier, les participants apporteront leur propre ordinateur et recevront un exemple de jeu de suivi de données de soccer. Ils apprendront à gérer les données et à effectuer quelques calculs de base. Nous espérons qu'ils prendront conscience des possibilités offertes par cette nouvelle source de données.

Accreditation Program Workshop • Atelier sur le Programme d'Accréditation

Room/Salle: A 1049

Date: Sunday June 2 / dimanche 2 juin

Time/Heure: 09:00-16:00

Robert Platt (McGill University)

Statistician as Expert Witness: Data and the Legal System

Le statisticien en tant que témoin expert : Les données et le système juridique

Workshop summary: Expert witnesses play an integral role in the legal system. Experts are people with specialized skill sets whose opinion may help a judge or jury make sense of factual evidence in a case. Testimony from expert witnesses can have a tremendous influence on the final decision of the judge. Statisticians have a natural role as expert witnesses, both in conducting and presenting data analyses, in interpreting and providing opinions on data and studies, and in providing opinions on the statistical methods used by other experts. I will describe the role of the expert in the court system and discuss specific examples and case studies where statisticians have contributed to the legal process. Through working sessions we will role-play as experts and review statisticians' reports in high-profile cases.

Résumé de l'atelier : Les témoins experts jouent un rôle essentiel dans le système juridique. Les experts sont des personnes possédant des compétences spécialisées dont l'opinion peut aider un juge ou un jury à comprendre les preuves factuelles dans une affaire. Les témoignages des experts peuvent exercer une influence considérable sur la décision finale du juge. Les statisticiens jouent un rôle naturel en tant que témoins experts, pouvant à la fois mener et présenter des analyses de données, interpréter et émettre des avis sur les données et les études, et émettre des avis sur les méthodes statistiques utilisées par d'autres experts. Je décrirai le rôle de l'expert dans le système judiciaire et discuterai d'exemples spécifiques et d'études de cas où les statisticiens ont contribué au processus juridique. Pendant des séances de travail, nous jouerons le rôle d'experts et examinerons les rapports des statisticiens dans des affaires très médiatisées.

Statistical Education Workshop • Atelier du Groupe d'éducation en statistique

Room/Salle: A 2071

Date: Sunday June 2 / dimanche 2 juin

Time/Heure: 09:00-16:00

Mireille Schnitzer (Université de Montréal) **Denis Talbot** (Université Laval)

Teaching Causal Inference: A Workshop for Educators

Enseigner l'inférence causale : atelier pour les éducateurs

Causal inference at the interface of statistics, study design, and epistemology, is the science of learning effects from data and background knowledge. Fundamentally important across multiple domains, including epidemiology, economics, political science, psychology, clinical science, engineering, and social sciences, there is increasing interest in the teaching of causal inference theory and statistical methods within many academic units. Different from more traditional statistics, causal inference fundamentally relies on non-statistical assumptions in order to make inferences. Entire systems of notation and graphical conventions have been developed to produce the framework within which statistical analysis can be planned. In addition, a vast literature of statistical (or so called "causal") methods have been developed to address the purely quantitative components of causal inference. Therefore, relaying the basic ideas of causal inference in relatively simple terms may seem like a daunting task.

In this workshop, we will outline and explain the elements of causal inference that we teach and have found to be the most relevant for an advanced undergraduate or graduate-level course, and the exercises that accompany them. We will focus on explaining how these elements are interconnected and give a global view on how causality can be addressed in study planning and analysis. These elements include

• Counterfactual theory, notation, and parameters,
• Identifiability of causal/counterfactual parameters via statistical estimands,
• Directed acyclic graphs and their role in identifiability,
• The interface between study design (e.g. randomized controlled trials, pseudo-experimental studies, observational studies, target trial emulation) and parameter identifiability,
• The targeted learning roadmap,
• Statistical estimation for counterfactual parameters, including marginal structural models (Regression, inverse probability of treatment weighting, G-computation),

L'inférence causale, à l'interface de la statistique, de la conception des études et de l'épistémologie, est la science de l'apprentissage d'effets à partir de données et de connaissances de base. D'une importance fondamentale dans de nombreux domaines, notamment en épidémiologie, économie, science politique, psychologie, science clinique, ingénierie et sciences sociales, l'enseignement de la théorie de l'inférence causale et des méthodes statistiques suscite un intérêt croissant dans de nombreuses unités universitaires.

Contrairement à la statistique plus traditionnelle, l'inférence causale se fonde principalement sur des hypothèses non statistiques afin d'effectuer des déductions. Des systèmes entiers de notation et de conventions graphiques ont été développés pour produire le cadre dans lequel l'analyse statistique peut être planifiée. En outre, une vaste littérature de méthodes statistiques (dites « causales ») a été développée pour traiter les éléments purement quantitatifs de l'inférence causale. Par conséquent, relayer les idées de base de l'inférence causale en termes simples peut sembler ardu.

Dans cet atelier, nous présenterons et expliquerons les éléments de l'inférence causale que nous enseignons et que nous avons trouvés les plus pertinents pour un cours avancé de premier ou de deuxième cycle, ainsi que les exercices qui les accompagnent. Nous nous attacherons à expliquer comment ces éléments sont interconnectés et à donner une vue d'ensemble de la manière dont la causalité peut être abordée dans la planification et l'analyse d'une étude. Ces éléments sont les suivants :

• Théorie, notation et paramètres contrefactuels
• Identifiabilité des paramètres causaux/contrefactuels par le biais d'estimandes statistiques
• Graphes acycliques dirigés et leur rôle dans l'identifiabilité
• Interface entre la conception de l'étude (par exemple, essais contrôlés randomisés, études pseudo-expérimentales, études d'observation, émulation d'essais ciblés) et l'identifiabilité des paramètres
• Feuille de route de l'apprentissage ciblé
• Estimation statistique des paramètres contrefactuels, y compris des modèles structurels marginaux (régression, pondération de la probabilité inverse de traitement, calcul G)
• Cadres semi-paramétriques (y compris estimation ciblée du maximum de vraisemblance) et intégration de l'apprentissage au-

• Semiparametric frameworks (including targeted maximum likelihood estimation) and machine learning integration,
• Alternative identifiability through instrumental variable methods,
• Overview of more advanced topics such as mediation analysis and longitudinal treatments.

We will also include discussion on how to target your course material to your audience and some approaches to evaluation.

Prerequisites of workshop: Interest in causal inference, understanding of basic statistical theory and methods, generalized linear regression.

tomatique
• Identifiabilité alternative par des méthodes de variables instrumentales
• Aperçu de sujets plus avancés tels que l'analyse de la médiation et les traitements longitudinaux.

Nous discuterons également de la manière de cibler votre matériel de cours en fonction de votre public et de certaines approches en matière d'évaluation.

Conditions préalables à l'atelier : Intérêt pour l'inférence causale, compréhension de la théorie et des méthodes statistiques de base, régression linéaire généralisée.

Probability Workshop • Atelier du Groupe de probabilité

Room/Salle: A 3017

Date: Sunday June 2 / dimanche 2 juin

Time/Heure: 09:00-16:00

Bouchra Nasri (École de Santé Publique de l'Université de Montréal)

Introduction to Infectious Disease Modeling

Introduction à la modélisation des maladies infectieuses

An overview of Modelling Infectious disease by Iain Moyles (Math & Stats, York U), Bouchra Nasri (École de Santé Publique, Université de Montréal) and Idriss Sekkak (École de Santé Publique, Université de Montréal)
This 6-hour course aims to give an overview about infectious disease modeling. The workshop will begin with the concepts mathematical modelling and how they can be used to model disease outbreaks. It will explore different model structures including compartmental and stochastic models and how modelling provides a deeper understanding of disease dynamics. Finally, we focus on the selection and use of models as predictive tools and to better understand fundamental epidemiological processes. This course offers a step-by-step introduction to infectious disease modeling, from basic concepts to the appropriate choice of models.

The outline of the course is detailed below:

Part 1: Introduction: Philosophy of modelling, the basic elements of infection and transmission, Definitions, types, and characterization of infectious diseases. Epidemiological models. Notions of differential equations and the differential equation limit of a model.

Part 2: Compartmental Models: Compartment-based epidemiological models. The basic epidemiological model SIR (Susceptible-Infected-Recovered) and related concepts, including the calculation of R0, the equilibrium point, and final size. Analysis of host strategies and heterogeneities (compartments by age, by risk, etc.). Transition to PDE.

Part 3: Stochastic Models: Stochastic epidemiological models. Stochastic compartment models: Motivation and concepts. Mathematical analysis and numerical approximations of stochastic epidemiological models.

Part 4: Model Determination: Selecting an appropriate model for an infectious disease. The different types of

Un aperçu de la Modélisation des Maladies Infectieuses par Iain Moyles (Département de Mathématiques et Statistiques, Université York), Bouchra Nasri (École de Santé Publique, Université de Montréal) et Idriss Sekkak (École de Santé Publique, Université de Montréal).

Ce cours intensif de 6 heures est conçu pour offrir une introduction complète à la modélisation des maladies infectieuses. Nous commencerons par une exploration des principes fondamentaux de la modélisation mathématique et leur application à l'étude des épidémies. Le cours se poursuivra avec une analyse des différentes structures de modèles, y compris les modèles compartimentaux et stochastiques, pour démontrer comment la modélisation peut enrichir notre compréhension de la dynamique des maladies. L'accent sera ensuite mis sur la sélection et l'application de modèles en tant qu'outils prédictifs et pour approfondir notre compréhension des processus épidémiologiques sous-jacents. Ce cours propose une approche progressive de la modélisation des maladies infectieuses, des notions de base à la sélection judicieuse de modèles.

Contenu du cours :

Partie 1 : Introduction : Philosophie de la modélisation, éléments fondamentaux de l'infection et de la transmission, définitions, types et caractéristiques des maladies infectieuses, introduction aux modèles épidémiologiques, et aperçu des équations différentielles et de leur importance dans la modélisation.

Partie 2 : Modèles Compartimentaux : Exploration des modèles épidémiologiques basés sur les compartiments, y compris le modèle SIR (Susceptible-Infecté-Récupéré) de base, le calcul de R0, l'analyse des points d'équilibre et de la taille finale de l'épidémie, et l'examen des stratégies hôtes et des hétérogénéités (par âge, par risque, etc.), avec une transition vers les équations aux dérivées partielles (EDP).

Partie 3 : Modèles Stochastiques : Introduction aux modèles épidémiologiques stochastiques, motivation et concepts clés des modèles compartimentaux stochastiques, analyse mathématique et approximations numériques.

compartment models. Modeling immunity and incubation time. Selection of model structure.

This 6-hour course is based on lectures with exercises based on case studies in infectious disease epidemiology. R codes for basic models and structural modifications.

Partie 4 : Détermination du Modèle : Critères de sélection d'un modèle approprié pour une maladie spécifique, exploration des différents types de modèles compartimentaux, modélisation de l'immunité et du temps d'incubation, et choix de la structure du modèle.

Le cours combine des présentations théoriques à des exercices pratiques basés sur des études de cas en épidémiologie des maladies infectieuses. Des codes R pour les modèles de base et leurs modifications structurelles seront fournis aux participants.

Survey Methods Workshop • Atelier du Groupe des méthodes d'enquête

Room/Salle: HH 3017

Date: Sunday June 2 / dimanche 2 juin

Time/Heure: 09:00-16:00

Anne-Sophie Charest (Université Laval)

Statistical Data Privacy

Confidentialité des données statistiques

In this workshop we will learn about the science of collecting, analyzing and sharing confidential data without disclosing personal information. We will first provide an overview of the different goals and approaches in this vast field of study, making connections with work from the computer science community, often published under different terminology. We will then consider a specific approach known as differential privacy which is the focus of much research and is used in practice by certain statistical agencies and private companies. We will explain the origin of this formal privacy measure, look in details at its mathematical definition and its meaning, and sow how to implement it for simple tasks. The rest of the workshop will focus on the use of synthetic datasets for privacy purposes, looking at how to generate such datasets and evaluate their quality in terms of risk and utility. All the content will be illustrated with R code and some time will be reserved for the participants to experiment with the methods on real datasets.

Outline:

1. Statistical data privacy
2. Differential privacy
3. Creating synthetic datasets
4. Evaluating and using synthetic datasets

Dans cet atelier, nous découvrirons la science de la collecte, de l'analyse et du partage de données confidentielles sans divulguer de renseignements personnels. Nous commencerons par un aperçu des objectifs et approches de ce vaste domaine d'étude, en établissant des liens avec les travaux de la communauté des informaticiens, souvent publiés sous une terminologie différente. Nous exploreronss ensuite une approche spécifique connue sous le nom de protection différentielle de la vie privée, qui fait l'objet de nombreuses recherches et est utilisée en pratique par certains organismes statistiques et entreprises privées. Nous expliquerons l'origine de cette mesure de confidentialité formelle, examinerons en détail sa définition mathématique et sa signification, puis montrerons comment la mettre en œuvre pour des tâches simples. Le reste de l'atelier se concentrera sur l'utilisation d'ensembles de données synthétiques à des fins de protection de la vie privée, en examinant comment générer de tels jeux de données et évaluer leur qualité en termes de risque et d'utilité. Nous illustrerons le contenu par du code R et réserverons une partie du temps pour laisser les participants expérimenter les méthodes sur des ensembles de données réels.

Aperçu :

1. Confidentialité des données statistiques
2. Protection différentielle de la vie privée
3. Création d'ensembles de données synthétiques
4. Évaluation et utilisation d'ensembles de données synthétiques

Data Science and Analytics Workshop • Atelier du Groupe de science des données et analytiques

Room/Salle: A 2065

Date: Sunday June 2 / dimanche 2 juin

Time/Heure: 09:00-16:00

Tiffany Timbers (University of British Columbia) **Daniel Chen** (University of British Columbia) **G. Alexi Rodríguez-Arelis** (University of British Columbia) **Katie Burak** (University of British Columbia)

How to create and distribute R packages

Comment créer et distribuer des paquets R

In this workshop learners will be introduced to what an R package is and when they should invest the time to make one. By the end of the workshop, learners should be able to build their own package that can be easily shared and installed by others. Additional important topics such as code testing, documentation and licenses will also be attended to. The workshop outline is shown below below:

- What is an R package and when should I make one?
- Hands-on building practice building an R package
- Ensuring your code works as expected - and introduction to testing
- Package documentation
- Introduction to continuous integration using GitHub Actions
- Sharing and publishing packages on GitHub and CRAN
- Copyright & Licenses (who owns the code?)

Dans cet atelier, les participants apprendront ce qu'est un paquet R et quand ils devraient prendre le temps d'en créer un. À la fin de l'atelier, ils seront en mesure de créer leur propre paquet qui peut être facilement partagé et installé par d'autres. D'autres sujets importants tels que les tests de code, la documentation et les licences seront également abordés. Le plan de l'atelier est présenté ci-dessous :

- Qu'est-ce qu'un paquet R et quand dois-je en créer un ?
- Entraînement pratique à la construction d'un paquet R
- S'assurer que votre code fonctionne comme prévu - et introduction aux tests
- Documentation du paquet
- Introduction à l'intégration continue à l'aide des actions GitHub
- Partage et publication de paquets sur GitHub et CRAN
- Droits d'auteur et licences (qui est propriétaire du code ?)

Biostatistics Workshop • Atelier biostatistique

Room/Salle: A 1043

Date: Sunday June 2 / dimanche 2 juin

Time/Heure: 09:00-16:30

Babette Brumback (University of Florida)

Fundamentals of Causal Inference: with R

Principes fondamentaux de l'inférence causale : avec R

One of the primary motivations for clinical trials and observational studies of humans is to infer cause and effect. Disentangling causation from confounding is of utmost importance. Fundamentals of Causal Inference: With R explains and relates different methods of confounding adjustment in terms of potential outcomes and graphical models, including standardization, doubly robust estimation, difference-in-differences estimation, front-door estimation, and instrumental variables estimation. These methods are compared in terms of estimating the average effect of treatment on the treated (ATT). The fundamentals of mediation analysis and adjusting for time-dependent confounding are also presented. Several real data examples, simulation studies, and analyses using R motivate and illustrate the methods throughout. The course assumes familiarity with basic statistics and probability, regression, and R. The course will be taught with a blend of lecture, worked examples and hands-on examples in R.

L'une des principales motivations des essais cliniques et des études d'observation chez l'homme est de déduire les causes et les effets. Il est extrêmement important de démêler la causalité des facteurs de confusion. ≪ Principes fondamentaux de l'inférence causale : avec R ≫ explique et met en relation différentes méthodes d'ajustement des facteurs de confusion en termes de résultats potentiels et de modèles graphiques, y compris la standardisation, l'estimation doublement robuste, l'estimation des différences dans les différences, l'estimation frontale (front door) et l'estimation des variables instrumentales. Ces méthodes sont comparées en termes d'estimation de l'effet moyen du traitement sur le traité (ATT). Les principes fondamentaux de l'analyse de la médiation et de l'ajustement pour les facteurs de confusion dépendants du temps sont également présentés. Plusieurs exemples de données réelles, des études de simulation et des analyses utilisant R motivent et illustrent les méthodes tout au long du cours. Le cours suppose une certaine familiarité avec la statistique et la probabilité de base, la régression et R. Le cours sera enseigné à l'aide d'un mélange de cours magistraux, d'exemples travaillés et d'exemples pratiques dans R.

Scientific Program • Programme scientifique

Monday June 3 lundi 3 juin

08:30-09:50 **Invited / Sur invitation** (abstract/résumé 87) **IIC 2001**

SSC Presidential Invited Address
Allocution de l'invité de la présidente de la SSC

Chair/Président: Shirley E. Mills
Organizer/Responsable: Shirley E. Mills

08:30-09:50 **Karen Kafadar** (University of Virginia)
Statistics FOR Data Science: Combining Statistics and Exploratory Data Analysis / Statistique pour la science des données : Combiner la statistique et l'analyse exploratoire des données

---

10:20-11:50 **Invited / Sur invitation** (abstract/résumé 88) **A 1045**

Advances in Spatial and Spatiotemporal Modeling: Uncovering Complex Patterns and Enhancing Inference
Progrès en modélisation spatiale et spatiotemporelle : découverte de modèles complexes et amélioration de l'inférence

Chair/Président: Cindy Feng
Organizer/Responsable: Cindy Feng

10:20-10:42 **Rob Deardon** (University of Calgary) **Yirao Zhang** (University of Calgary) **Lorna Deeth** (University of Guelph)
Composite Spatial Epidemic Models: Computational Efficiency via Clustering / Modèles d'épidémies spatiales composites : efficacité de calcul grâce au regroupement

10:42-11:05 **Mahmoud Torabi** (University of Manitoba) **Charmaine B. Dean** (University of Waterloo) **Georges Bucyibaruta** (Imperial College London)
Innovative Strategies for Influenza Data Examination / Stratégies innovantes pour l'examen des données sur la grippe

11:05-11:27 **Patrick Brown** (Unity Health Toronto) **Jamie Stafford** (University of Toronto)
A Historical Look at Geostatistical Models and Gaussian Markov Random Fields / Un regard historique sur les modèles géostatistiques et les champs aléatoires de Markov gaussiens

11:27-11:50 **Dirk Douwes-Schultz** (McGill)
Markov Switching Zero-Inflated Space-Time Multinomial Models for Comparing Multiple Infectious Diseases / Modèles multinomiaux spatio-temporels à commutation de Markov avec excès de zéros pour la comparaison de maladies infectieuses multiples

---

10:20-11:50 **Invited / Sur invitation** (abstract/résumé 91) **IIC 2001**

Risk Quantification in Actuarial Science
Quantification des risques en science actuarielle

Chair/Président: Silvana Pesenti
Organizer/Responsable: Silvana Pesenti
Sponsor/Commanditaires: Actuarial Science Section/Groupe de science actuarielle

10:20-10:50 **Fangda Liu** (University of Watelroo)
Distributional Uncertainty and Risk Sharing / Incertitude distributionnelle et partage des risques

10:50-11:20 **Thai H. Nguyen** (Université Laval)
Pareto-Optimal Investments and Contracting for Non-linear Payoffs / Investissements d'optimum de Pareto et passation de contrat pour des gains non linéaires

11:20-11:50 **Ilaria Peri** (Birkbeck, University of London) **Akif Ince** (Birkbeck, University of London) **Marlon Moresco** (Federal University of Rio Grande do Sul) **Silvana Pesenti** (University of Toronto)
Elictable Risk Functionals with Quasi-convex Score / Fonctionnelles de risque élicitables avec score quasi-convexe

---

10:20-11:50 **Invited / Sur invitation** (abstract/résumé 93) **A 1043**

Data Science and Analytics Section Keynote Lecture
Présentation principale du Groupe de science des données et analytique

Chair/Président: Nathaniel Tyler Stevens
Organizer/Responsable: Nathaniel Tyler Stevens
Sponsor/Commanditaires: Data Science and Analytics Section/Groupe de science des données et analytique

10:20-11:50 **Xiaoli Meng** (Harvard University)
Privacy, Data Privacy, and Differential Privacy / Vie privée, confidentialité des données et confidentialité différentielle

---

10:20-11:50 **Invited / Sur invitation** (abstract/résumé 94) **C 2045**

Harnessing Statistical and Computational Models for Neuroscience
Exploiter les modèles statistiques et informatiques pour la neuroscience

Chair/Président: Reza Ramezan
Organizer/Responsable: Reza Ramezan
Sponsor/Commanditaires: Biostatistics Section/Groupe de biostatistique

10:20-10:50 **Reza Ramezan** (University of Waterloo) **Farouk Nathoo** (University of Victoria) **Cedric Beualac** (Université du Québec à Montréal) **Michelle Miranda** (University of Victoria) **Jiguo Cao** (Simon Fraser University) **Liangliang Wang** (Simon Fraser University) **Mirsa Beg** (Simon Fraser University) **Yin Song** (University of Victoria) **Leno Rocha** (University of Victoria) **Sidi Wu** (Simon Fraser University) **Erin Gibson** (Simon Fraser University)
Neural Network Feature Extraction and Bayesian Spatial Modeling for Imaging Genetics / Extraction de caractéristiques d'un réseau neuronal et modélisation spatiale bayésienne pour l'imagerie génétique

10:50-11:20 **Lloyd T. Elliott** (Simon Fraser University)
Mediation Analysis shows effects for the LIFO Network in Brain Imaging Genetics / L'analyse de médiation montre des effets pour le réseau LIFO dans la génétique par imagerie cérébrale

11:20-11:50 **Meixi Chen** (University of Waterloo) **Martin Lysy** (University of Waterloo) **Reza Ramezan** (University of Waterloo)
Insights into Brain Dynamics: A Scalable Spike-Train Model for Neuronal Interactions / Perspectives sur les dynamiques cérébrales : un modèle extensible de trains d'impulsions nerveuses pour les interactions neuronales

---

10:20-11:50 **Invited / Sur invitation** (abstract/résumé 96) **ED 2018A**

Inference for Autoregressive and Markov Models; A Memorial Session for Wolfgang Wefelmeyer
Inférence des modèles autorégressifs et de Markov ; session commémorative en l'honneur de Wolfgang Wefelmeyer

Chair/Président: Thomas Salisbury
Organizer/Responsable: Priscilla E. Greenwood
Sponsor/Commanditaires: Probability Section/Groupe de probabilité

10:20-10:50 **Priscilla E. Greenwood** (The University of British Columbia)
Work of Wolfgang Wefelmeyer: Asymptotic Efficiency, Inference for Stochastic Processes, Nonparametric and Semiparametric Estimation / Travail de Wolfgang Wefelmeyer : efficacité asymptotique, inférence pour les processus stochastiques, estimation semi-paramétrique et non paramétrique

10:50-11:20 **Ursula U. Müller** (Texas A&M University)
Estimation for Markov Chains with Periodically Missing Observations / Estimation pour les chaînes de Markov avec observations périodiquement manquantes

11:20-11:50 **Anton Schick** (Binghamton University) **Wolfgang Wefelmeyer** (Universitaet zu Koeln)
Efficient Density Estimation in an AR(1) Model / Estimation efficace de la densité dans un modèle AR(1)

10:20-11:50 **Invited / Sur invitation** (abstract/résumé 98) **ED 2018B**

Networking for Large Programs of Research (Panel)
Le réseautage dans les grands programmes de recherche (Table ronde)

Chair/Président: Thérèse A. Stukel
Organizer/Responsable: Thérèse A. Stukel

10:20-11:50 **Lisa M. Lix** (University of Manitoba) **Robyn Tamblyn** (McGill University) **Robert W. Platt** (McGill University) **Shelley Bull** (Lunenfeld-Tanenbaum Research Institute, Sinai Health, and University of Toronto)
Networking for Large Programs of Research / Le réseautage dans les grands programmes de recherche

---

10:20-11:50 **Contributed / Communications libres** (abstract/résumé 99) **A 1049**

Sampling, Multi-level and Clustered Data
Données d'échantillonnage, multi-niveaux et groupées

Chair/Président: Gyanendra Pokharel

10:20-10:35 **Abigail McGrory** (University of Toronto) **Anna Heath** (The Hospital for Sick Children)
Enhancing the Efficiency of Adaptive Platform Trials Through the Exploration of Alternative Treatment Ranking Methods / Augmentation de l'efficacité des essais de plateforme adaptatifs grâce à l'exploration de méthodes de classement de traitement de deuxième ligne

10:35-10:50 **Zelalem Firisa Negeri** (University of Waterloo) **Narayanaswamy Balakrishnan** (McMaster University)
Nonparametric statistical methods for diagnostic test meta-analyses / Méthodes statistiques non paramétriques pour les méta-analyses de tests diagnostiques

10:50-11:05 **Chen Chen** (University of Toronto) **Aya A. Mitani** (University of Toronto)
A Joint Model of Hierarchical Data with Multivariate Skewed-t Distribution and Informative Cluster Size / Modèle conjoint de données hiérarchiques avec distribution multivariée asymétrique-t et taille de grappe informative

11:05-11:20 **Cody B. Halden** (University of Ottawa) **Jemila Hamid** (University of Ottawa)
Performance of Iteratively Reweighted Growth Curve Model / Performance du modèle de courbe de croissance pondéré de manière itérative

11:20-11:35 **Mamadou Yauck** (Université du Québec à Montréal (UQAM))
A statistical Test for Detecting Homophily and Preferential Recruitment in Link-Tracing Sampling Surveys / Test statistique pour détecter l'homophilie et le recrutement préférentiel dans les enquêtes par sondage à dépistage de liens

11:35-11:50 **Sean Xinyang Feng** (University of Toronto) **Aya A. Mitani** (University of Toronto)
Multivariate Joint Modeling for Clustered Data with Application to Periodontal Disease / Modélisation conjointe multivariée pour des données regroupées appliquée à la parodontopathie

---

10:20-11:50 **Contributed / Communications libres** (abstract/résumé 103) **A 2071**

Biostatistics Student Research Session #1
Session de recherche étudiante en biostatistique #1

Chair/Président: Mohammad Kaviul Anam Khan

10:20-10:35 **Mohammad Reza Fahimi** (University of Toronto Dalla Lana School of Public Health) **Aya A. Mitani** (University of Toronto) **Osvaldo Espin-Garcia** (Western University)
Optimal Sampling Fractions in Two-Phase Designs with Ordinal Outcomes / Fractions d'échantillonnage optimales dans les plans à deux phases avec réponses ordinales

10:35-10:50 **Emily Somerset** (University of Toronto)
Estimating and Forecasting Disease Trends from Wastewater Surveillance Data / Estimation et prévision des tendances d'une maladie à partir des données de surveillance des eaux usées

10:50-11:05 **Luke Hagar** (University of Waterloo) **Nathaniel T. Stevens** (University of Waterloo)
Quantile Estimation for Sampling Distributions of Posterior Probabilities / Estimation des quantiles pour les distributions d'échantillonnage des probabilités a posteriori

11:05-11:20 **Zijin Liu** (University of Toronto) **Zhihui (Amy) Liu** (University Health Network) **Olli Saarela** (University of Toronto)
A Bayesian Joint Model for Mediation Analysis With Matrix-Valued Mediators / Modèle conjoint bayésien pour l'analyse de la médiation avec des médiateurs à valeur matricielle

11:20-11:35 **Shijie Min** (University of Toronto)
A Copula-infused Graph Neural Network for Cell Type Classification in Single-cell RNA Sequencing Data / Réseau neuronal graphique basé sur des copules pour la classification des types de cellules dans des données de séquençage de l'ARN unicellulaire

11:35-11:50 **Amanda Qiu** (University of Victoria) **Hong Wang** (SANOFI) **Luc Essermeant** (SANOFI) **Weiliang Qiu** (SANOFI) **Xuekui Zhang** (University of Victoria)
Novel Evaluation of Cell-Type Deconvolution Algorithms in Bulk RNAseq Data Analyses / Nouvelle évaluation d'algorithmes de déconvolution de type de cellule dans les analyses de grand nombre de données de séquençage de l'ARN

10:20-11:50 **Invited / Sur invitation** (abstract/résumé 107) **A 1046**

Advancements in Sports Analytics
Progrès en analyse du sport

Chair/Président: Tianyu Guan
Organizer/Responsable: Tianyu Guan
Sponsor/Commanditaires: Data Science and Analytics Section/Groupe de science des données et analytique

10:20-10:50 **Shirley E. Mills** (Carleton University)
Evaluating Player and Team Performance / Évaluer un joueur et la performance d'équipe

10:50-11:20 **Paramjit S. Gill** (University of British Columbia Okanagan)
Tennis Analytics Based on Point-by-Point Data / Analyse du tennis basée sur des données point par point

11:20-11:50 **Tim B. Swartz** (Simon Fraser University)
Causal Inference Problems in Soccer using Tracking Data / Problèmes d'inférence causale dans le soccer et données de suivi

---

10:20-11:50 **Contributed / Communications libres** (abstract/résumé 109) **A 2065**

Biostatistics Student Research Session #2
Session de recherche étudiante en biostatistique #2

Chair/Président: Ana Carolina da Cruz

10:20-10:35 **Malcolm Risk** (University of Michigan) **Xu Shi** (University of Michigan) **Lili Zhao** (University of Michigan)
Distributed Kaplan-Meier Curves via the Influence Function / Courbes de Kaplan-Meier distribuées avec la fonction d'influence

10:35-10:50 **Shiyao Ying** (University of Toronto) **Yun-Hee Choi** (Western University, London, Ontario) **Laurent Briollais** (Lunenfeld-Tanenbaum Research Institute, Mount Sinai Hospital, Toronto, Ontario)
Correlated Frailty Models with Kinship for Analysis of Time-to-Event Outcomes within Families / Modèles de fragilité corrélés avec la parenté pour l'analyse des résultats temporels au sein des familles

10:50-11:05 **Amin Abed** (University of Manitoba) **Mahmoud Torabi** (University of Manitoba) **Zeinab Mashreghi** (University of Winnipeg)
Modeling of Infectious Disease with Reinfection: Tuberculosis Transmission in Manitoba / Modélisation des maladies infectieuses avec réinfection : transmission de la tuberculose au Manitoba

11:05-11:20 **Mei Dong** (University of Toronto) **Linbo Wang** (University of Toronto) **Wei Xu** (University of Toronto)
Robust Estimator for Average Treatment Effect with Continuous Instrumental Variables / Estimateur robuste de l'effet moyen du traitement avec des variables instrumentales continues

11:20-11:35 **Fatema Tuj Johara** (University of Toronto Dalla Lana School of Public Health) **Eleanor M. Pullenayegum** (Hospital for Sick Children)
Methods of Quantifying Within Person Variability for Longitudinal Data With Irregular Observation / Méthodes de quantification de la variabilité intra-personnelle pour données longitudinales avec observations irrégulières

11:35-11:50 **Wensha Zhang** (Dalhousie University) **Toby J. Kenney** (Dalhousie University) **Lam Ho** (Dalhousie University)
Detection of Evolutionary Shifts in Variance under an Ornstein–Uhlenbeck Model / Détection des changements évolutifs de la variance dans le cadre d'un modèle d'Ornstein-Uhlenbeck

---

10:20-11:50 **Contributed / Communications libres** (abstract/résumé 113) **C 2033**

New Approaches for Statistical Modelling and Design of Experiments
Nouvelles approches pour la modélisation statistique et les plans d'expériences

Chair/Président: Armin Hatefi

10:20-10:35 **Golshid Aflaki** (HEC Montréal) **Jean-François Plante** (HEC Montreal) **Juliana Schulz** (HEC Montreal)
A New Bivariate Zero-Inflated Poisson Model / Un nouveau modèle Poisson bivarié à excès de zéros

10:35-10:50 **Pranath Pussella** (Brock University) **Tianyu Guan** (Brock University) **Robert Nguyen** (University of New South Wales)
Simulation for Cricket: A Machine Learning Approach / Simulation pour le cricket : une approche d'apprentissage automatique

10:50-11:05 **David Awosoga** (University of Waterloo)
Investigating Player Contribution in Volleyball Using Bayesian Spatiotemporal Data Analysis / Enquête sur la contribution des joueurs au volley-ball à l'aide d'une analyse de données spatio-temporelle bayésienne

11:05-11:20 **Yuying Huang** (University of Waterloo) **Samuel Wong** (University of Waterloo)
Sequential Design Strategy for Mean Response Surface Modeling of Expensive Stochastic Simulation with Heterogeneous Noise via Bayesian Framework / Stratégie de conception séquentielle pour la modélisation de surface de réponse moyenne de simulation stochastique coûteuse avec bruit hétérogène dans le cadre bayésien

---

10:20-11:50 **Contributed / Communications libres** (abstract/résumé 116) **C 4036**

Strategies in Teaching Statistics and Evaluating Statistical Knowledge
Stratégies d'enseignement de la statistique et l'évaluation des connaissances statistiques

Chair/Président: Mark Reesor

10:20-10:35 **Scott Andrew Robison** (University of Calgary)
Interesting Courses, Need to be Interesting / Les cours intéressants doivent susciter l'intérêt

10:35-10:50 **Tharshanna Nadarajah** (McGill University)
Enhance Student Learning by Reviewing Previous Learning Regularly / Amélioration de l'apprentissage des étudiants par la révision régulière des cours précédents

10:50-11:05 **Danika M. Lipman** (University of Calgary) **Scott Andrew Robison** (University of Calgary)
Creating a Positive Class Environment Through Assessment / Créer un environnement de classe positif au moyen de l'évaluation

11:05-11:20 **Lengyi Spectrum Han** (The University of British Columbia)
Using WipeBooks to Increase Student Engagement / L'emploi de WipeBooks pour augmenter l'engagement étudiant

11:20-11:35 **Shahriar Shams** (University of Toronto Scarborough) **Sotirios Damouras** (University of Toronto Scarborough)
Implementing Computer-Based Assessments in Statistics Courses. / Implémentation d'évaluations informatisées dans les cours de statistique

11:35-11:50 **Chelsea Uggenti** (University of Waterloo) **Divya Lala** (University of Waterloo)
Reflecting on the First Year of our Teaching Assistant Program / Réflexion sur la première année de notre programme d'assistants d'enseignement

---

10:20-11:50 **Contributed / Communications libres** (abstract/résumé 120) **C 3053**

Big Data Analysis
Analyse des données volumineuses

Chair/Président: Joel A. Dubin

10:20-10:35 **Kyu Min Shim** (University of Waterloo)
Variance Reduction with Model-Based Counterfactual Estimation / Réduction de la variance par estimation contrefactuelle basée sur un modèle

10:35-10:50 **Trang Bui** (University of Waterloo) **Stefan Steiner** (University of Waterloo) **Nathaniel T. Stevens** (University of Waterloo)
Analysis of Experiments on Networks with Binary Outcomes / Analyse d'expériences sur les réseaux avec réponses binaires

10:50-11:05 **Nathan Phelps** (University of Western Ontario)
Challenges when Calibrating a Random Forest Fit to Undersampled Data / Problèmes de calibrage pour une forêt aléatoire ajustée à des données sous-échantillonnées

11:05-11:20 **Yutong Lu** (University of Toronto) **Yan Yi Li** (University of Toronto)
Knowledge Fusion of Large Language Models for Molecular Property Prediction / Fusion des connaissances des grands modèles de langage pour la prédiction des propriétés moléculaires

11:20-11:35 **Bahram Moeinianfar** (University of Manitoba) **Mohammad Jafari Jozani** (University of Manitoba)
Elite-Driven Support Vector Machine (EDSVM): Building Classifiers with Insights from a Collective of Support Vectors / Machine vectorielle de support dirigée par les élites (EDSVM) : construction de classifieurs avec des informations provenant d'une collection de vecteurs de support

---

10:20-11:50 **Invited / Sur invitation** (abstract/résumé 123) **C 3033**

Funding Opportunities for Statistics and Data Science Research
Opportunités de financement pour la recherche en statistiques et en science des données

Chair/Président: Saman Muthukumarana
Organizer/Responsable: Saman Muthukumarana
Sponsor/Commanditaires: Research Committee/Comité de la recherche

10:20-11:50 **Donald Estep** (Simon Fraser University/CANSSI) **Pascal Marchand** (Natural Sciences and Engineering Research Council of Canada) **Ilana Gombos** (Canadian Institutes of Health Research) **Katarina Dedovic** (Canadian Institutes of Health Research) **Heidi Crummel** (MITACS)
Funding Opportunities for Statistics and Data Science Research / Possibilités de financement pour la recherche en statistique et science des données Ⓔ E

13:30-15:00 **Invited / Sur invitation** (abstract/résumé 124) **A 1043**

Fairness and Discrimination in Insurance
Équité et discrimination en assurance

Chair/Président: Marie-Pier Côté
Organizer/Responsable: Marie-Pier Côté
Sponsor/Commanditaires: Actuarial Science Section/Groupe de science actuarielle

13:30-14:00 **Carlos Andres Araiza Iturria** (University of Waterloo) **Mary Hardy** (University of Waterloo) **Paul Marriott** (University of Waterloo)
Discrimination in Insurance Pricing / La discrimination dans la tarification de l'assurance Ⓔ E

14:00-14:30 **Chengguo Weng** (University of Waterloo)
Optimal Prediction under Several Fairness Criteria / Prévision optimale selon plusieurs critères d'équité Ⓔ E

14:30-15:00 **Marie-Pier Côté** (Université Laval) **Olivier Côté** (Université Laval) **Arthur Charpentier** (Université du Québec à Montréal)
A Fair Price to Pay: Exploiting Causal Graphs for Fairness in Insurance / Un juste prix à payer : exploiter les graphes causaux pour l'équité en assurance Ⓔ E

13:30-15:00 **Invited / Sur invitation** (abstract/résumé 126) **A 1045**

Survey Methods Section Presidential Invited Address
Allocution de l'invité du président du Groupe des méthodes d'enquête

Chair/Président: Éric Gagnon
Organizer/Responsable: Éric Gagnon
Sponsor/Commanditaires: Survey Methods Section/Groupe des méthodes d'enquête

13:30-15:00 **David Haziza** (University of Ottawa) **Mehdi Dagdoug** (McGill University) **Camelia Goga** (Université de Bourgogne Franche Comté)
Statistical Inference in the Presence of Imputed Survey Data through Regression Trees and Random Forests / Inférence statistique en présence de données d'enquête imputées au moyen d'arbres de régression et de forêts aléatoires Ⓔ E

13:30-15:00 **Invited / Sur invitation** (abstract/résumé 127) **A 1046**

New Advances in Statistics and Data Science
Nouvelles avancées en statistique et science des données

Chair/Président: Dehan Kong
Organizer/Responsable: Dehan Kong
Sponsor/Commanditaires: ICSA-Canada Chapter/Chapitre canadien de l'ICSA

13:30-14:00 **Annie Qu** (University of California, Irvine) **Hansen Ye** (UC Irvine) **Wenzhuo Zhou** (UC Irvine) **Ruoqing Zhu** (UIUC)
Stage-Aware Learning for Dynamic Treatments / Apprentissage conscient des étapes pour les traitements dynamiques Ⓔ E

14:00-14:30 **Peter X. Song** (University of Michigan)
Supervised Homogeneity Pursuit via Mixed Integer Optimization / Recherche d'homogénéité supervisée par optimisation en nombres entiers mixtes Ⓔ E

14:30-15:00 **Peijun Sang** (University of Waterloo) **Yao Luo** (University of Toronto)
Penalized Sieve Estimation of Structural Models / Estimation pénalisée par tamis des modèles structurels

---

13:30-15:00 **Invited / Sur invitation** (abstract/résumé 129) **C 2045**

Reproducibility in Machine Learning and Statistics
Reproductibilité en apprentissage automatique et statistique

Chair/Président: Tiffany A. Timbers
Organizer/Responsable: Tiffany A. Timbers
Sponsor/Commanditaires: Data Science and Analytics Section/Groupe de science des données et analytique

13:30-14:00 **Rohan Alexander** (University of Toronto)
Reproducibility and Code - Using LLMs to Translate Replication Packages to Enhance Credibility / Reproductibilité et code — utilisation des LLMs pour traduire les progiciels de réplication pour hausser la crédibilité

14:00-14:30 **Callandra Moore** (The Hospital for Sick Children)
Reproducibility in Clinical NLP / Reproductibilité dans le traitement automatique des langues en clinique

---

13:30-15:00 **Invited / Sur invitation** (abstract/résumé 131) **A 2071**

Generative Artificial Intelligence and What's Next for the Teaching and Learning of Statistics (Panel)
Intelligence artificielle générative et l'avenir de l'enseignement et de l'apprentissage de la statistique (Table ronde)

Chair/Président: Alison L. Gibbs
Organizer/Responsable: Alison L. Gibbs
Sponsor/Commanditaires: Statistical Education Section/Groupe d'éducation en statistique

13:30-15:00 **David Riegert** (Trent University) **Nathan Taback** (University of Toronto)
Generative Artificial Intelligence and What's Next for the Teaching and Learning of Statistics / Intelligence artificielle générative et l'avenir de l'enseignement et de l'apprentissage de la statistique

---

13:30-15:00 **Invited / Sur invitation** (abstract/résumé 132) **A 2065**

Modern Approaches for Clustering Data
Approches modernes pour le regroupement des données

Chair/Président: Sanjeena Dang
Organizer/Responsable: Brian Franczak
Sponsor/Commanditaires: Business and Industrial Statistics Section/Groupe de statistique industrielle et de gestion

13:30-13:52 **Mateen Shaikh** (Thompson Rivers University)
Applicants of Sequences of Surrogate Functions in Statistcal Learning / Applications de séquences de fonctions de remplacement à l'apprentissage statistique

13:52-14:15 **John R.J. Thompson** (The University of British Columbia) **Jesse Ghashti** (The University of British Columbia)
Kernel Metric Learning for Mixed-type Distance Shrinkage and Variable Selection / Apprentissage de mesure de noyaux pour le rétrécissement de distance de type mixte et la sélection de variables

14:15-14:37 **Paul David McNicholas** (McMaster University) **Mackenzie Neal** (McMaster University)
Flexible Variable Selection for Clustering / Sélection de variables flexibles pour le regroupement

14:37-15:00 **Brian Franczak** (MacEwan University)
Outlier Detection using an Asymmetric Laplace Distribution / Détection des valeurs aberrantes à l'aide d'une distribution de Laplace asymétrique

---

13:30-15:00 **Invited / Sur invitation** (abstract/résumé 134) **IIC 2001**

Distinguished Educator Award Address
Allocution du recipiendaire du Prix d'excellence en enseignement

Chair/Président: Wesley S. Burr
Organizer/Responsable: Wesley S. Burr

13:30-15:00 **Jim B. Stallard** (University of Calgary)
How to Clear a Room / Comment vider la salle

13:30-15:00 **Invited / Sur invitation** (abstract/résumé 135) **A 1049**

Advances in Methodology for Clinical Trials
Progrès en méthodologie des essais cliniques

Chair/Président: Wendy Lou
Organizer/Responsable: Wendy Lou

13:30-14:00 **Motomi Mori** (St. Jude Children's Research Hospital)
Clinical Trial Designs to Evaluate Gene and Cell Therapies in Rare Diseases / Conception d'essais cliniques pour évaluer les thérapies géniques et cellulaires dans les maladies rares

14:00-14:30 **Bingshu Chen** (Queen's University) **Wenyu Jiang** (Queen's University) **Parisa Gavanji** (Queen's University)
Penalized Likelihood Ratio Test for a Biomarker Threshold Effect in Clinical Trials Based on Generalized Linear Models / Test de rapport de vraisemblance pénalisé pour l'effet de seuil d'un biomarqueur dans des essais cliniques basés sur des modèles linéaires généralisés

14:30-15:00 **Aaron Hudson** (Fred Hutchinson Cancer Center) **Oliver Dukes** (Ghent University) **Mats Stensrud** (École Polytechnique Fédérale de Lausanne (EPFL)) **Riccardo Brioschi** (École Polytechnique Fédérale de Lausanne (EPFL))
A Nonparametric Test and Estimand for Qualitative Interactions / Test non paramétrique et estimation des interactions qualitatives

---

13:30-15:00 **Invited / Sur invitation** (abstract/résumé 137) **ED 2018A**

Advanced Statistical Inference for Emerging Infectious Disease Epidemics
Inférence statistique avancée pour les épidémies de maladies infectieuses émergentes

Chair/Président: Lam Ho
Organizer/Responsable: Lam Ho
Sponsor/Commanditaires: Biostatistics Section/Groupe de biostatistique

13:30-14:00 **Cindy Feng** (Dalhousie University)
Spatial Generalized Additive Location Scale Models for Modeling Infectious Disease Risk / Modèles de position-échelle spatiaux additifs généralisés pour la modélisation du risque de maladies infectieuses

14:00-14:30 **Justin James Ian Slater** (University of Guelph)
Overdispersed or Underreported? Inference for Infectious Disease Models with Underreported Case Counts / Surdispersion ou sous-déclaration ? Inférence pour des modèles de maladies infectieuses avec comptages de cas sous-déclarés

14:30-15:00 **Jason Xu** (Duke University)
Data-Augmented MCMC for Stochastic Epidemic Models / MCMC avec données augmentées pour des modèles épidémiques stochastiques

13:30-15:00 **Contributed / Communications libres** (abstract/résumé 139) **C 2033**

Actuarial Science 1
Science actuarielle 1

Chair/Président: Johanna G. Nešlehová

13:30-13:45 **Jean-François Bégin** (Simon Fraser University) **Barbara Sanders** (Simon Fraser University)
Benefit Volatility-targeting Strategies in Lifetime Pension Pools / Stratégies de ciblage de la volatilité de prestations dans les rentes viagères à paiements variables

13:45-14:00 **Kyran Cupido** (Kyran Cupido) **Petar Jevtic** (Arizona State University) **Luca Regis** (University of Torino) **Kenneth Zhou** (Arizona State University)
Spatial Natural Hedging – A General Framework with Application to the Mortality of U.S. States / Couverture naturelle spatiale — un cadre général appliqué à la mortalité aux États-Unis

14:00-14:15 **Barbara Sanders** (Simon Fraser University) **Jean-François Bégin** (Simon Fraser University) **Nikhil Kapoor** (Simon Fraser University)
A New Approximation of Annuity Prices for Age–Period–Cohort Models / Nouvelle approximation des coûts des rentes pour des modèles âge-période-cohorte

14:15-14:30 **Dante Mata Lopez** (Université du Québec à Montréal (UQAM))
On an optimal dividend problem with a concave bound on the dividend rate / Problème de dividende optimal avec borne concave du taux de dividende

14:30-14:45 **Hélène Cossette** (Université Laval) **Etienne Marceau** (Université Laval) **Benjamin Côté** (Université Laval)
Risk Models Defined on a Family of Tree-Based Markov Random Fields with Poisson Marginals / Modèles de risque définis sur une famille de champs aléatoires de Markov arborescents avec des distributions marginales de Poisson

14:45-15:00 **Etienne Marceau** (Université Laval)
Risk Sharing and Risk Allocation: Generating Function Approach / Partage et répartition des risques : méthode de génération de fonctions

13:30-15:00 **Contributed / Communications libres** (abstract/résumé 142) **C 4036**

Recent Developments in Inference and Computation
Développements récents en inférence et calcul

Chair/Président: Yunhong Lyu

13:30-13:45 **Yi Meng Chang** (University of Toronto Dalla Lana School of Public Health) **Petros Pechlivanoglou** (The Hospital for Sick Children) **Olli Saarela** (University of Toronto) **Eleanor M. Pullenayegum** (The Hospital for Sick Children)
Issues and Bias in Phase-of-Care Costing When Estimating Attributable Healthcare Costs of a Disease / Problèmes et biais relatifs au calcul des coûts par phase de soin lors de l'estimation des coûts de soins de santé imputables à une maladie

13:45-14:00 **Tessa Reimer** (University of Manitoba) **Alexandre Leblanc** (University of Manitoba) **Saman Muthukumarana** (University of Manitoba)
Bayesian Analysis of Batting Outcomes from Major League Baseball Using a Nested Dirichlet Prior Distribution / Analyse bayésienne des succès au bâton de la Ligue majeure de baseball à l'aide d'une distribution a priori de Dirichlet imbriquée

14:00-14:15 **Martin Lysy** (University of Waterloo)
PFJAX: Differentiable Particle Filtering in Python / PFJAX : Filtrage de particule différentiable dans Python

14:15-14:30 **Jeffrey Negrea** (University of Waterloo) **Jun Yang** (University of Copenhagen) **Haoyue Feng** (Boston University) **Daniel Roy** (University of Toronto) **Jonathan Huggins** (Boston University)
Statistical Inference with Stochastic Gradient Algorithms / Inférence statistique avec algorithmes de gradient stochastique Ⓔ E

14:30-14:45 **Lulu Zhang** (University of New Brunswick) **Renjun Ma** (University of New Brunswick) **Guohua Yan** (University of New Brunswick) **Xifen Huang** (Yunnan Normal University)
A New Logistic Model with Subject-Specific and Serially Correlated Time-Specific Distribution-Free Random Effects on the Unit Interval for Intensive Longitudinal Binary Data / Nouveau modèle logistique avec effets aléatoires de distribution libre à sujet spécifique et à temps spécifiques sériellement corrélés sur l'intervalle unité pour des données binaires longitudinales à forte intensité Ⓔ E

14:45-15:00 **Dingding Hu** (University of Waterloo)
ROC Curve Analysis under Non-ignorable Verification Bias / Analyse de la courbe ROC en présence d'un biais de vérification non négligeable Ⓔ E

13:30-15:00 **Contributed / Communications libres** (abstract/résumé 146) **C 3053**

Recent Developments in Survey Methods 1
Développements récents en méthodes d'enquête 1

Chair/Président: Wei Liu

13:30-13:45 **Zeinab Mashreghi** (University of Winnipeg)
Bootstrap Resampling Methods for Survey Data in R / Méthodes de rééchantillonnage bootstrap pour des données d'enquête avec R Ⓔ E

13:45-14:00 **Gradon Nicholls** (University of Waterloo)
Model-Assisted Double-Coding of Open-Ended Survey Questions with Large Language Models / Double codage assisté par modèle de questions d'enquête ouvertes à l'aide de grands modèles de langage Ⓔ E

14:00-14:15 **Michael John Ilagan** (McGill University) **Carl F. Falk** (McGill University)
A Mixture Model for p-values to Detect Bots in Likert-type Questionnaire Data / Un modèle de mélange pour les valeurs p afin de repérer les robots dans des données de questionnaire de type Likert Ⓔ E

14:15-14:30 **Atefeh Kheirollahi** (Memorial University of Newfoundland) **Nan Zheng** (Memorial University of Newfoundland) **Yildiz Yilmaz** (Memorial University of Newfoundland)
Accounting for Age Measurement Errors in Fish Growth Model Estimation using Length-stratified Age Sampling Data / Prise en compte des erreurs de mesure de l'âge dans l'estimation du modèle de croissance de poissons à l'aide de données d'échantillonnage par âge stratifiées selon la longueur Ⓔ E

13:30-15:00 **Contributed / Communications libres** (abstract/résumé 149) **ED 2018B**

Teaching Statistics With a Data-Centric Perspective
Enseigner la statistique dans une perspective centrée sur les données

Chair/Président: Harsha Harsha Perera

13:30-13:45 **Michael Wallace** (University of Waterloo)
Real-world Data Analysis in an Introductory Statistics Course: Assignments using Data from the Stanford Open Policing Project / Analyse de données réelles dans un cours d'introduction aux statistiques : travaux réalisés à l'aide des données du Stanford Open Policing Project Ⓔ E

13:45-14:00 **Katherine Daignault** (University of Toronto)
In-Depth Exploration of Linear Regression Concepts through Self-Paced LearnR Modules / Explorations des concepts de régression linéaire à travers les modules LearnR Ⓔ E

14:00-14:15 **Wanhua Su** (MacEwan University)
Applying Statistical Learning Methods to Complex Survey Data / Application des méthodes d'apprentissage statistique aux données d'enquêtes complexes Ⓔ E

14:15-14:30 **Bethany J.G. White** (University of Toronto) **Jastaranpreet Singh** (University of Toronto)
Going Hybrid: Transforming an Introductory Statistics Course for Life Sciences / Enseignement hybrique : transformer un cours d'introduction à la statistique pour les sciences de la vie

14:30-14:45 **Samantha-Jo Caetano** (University of Toronto) **Emily Somerset** (University of Toronto) **Andrea Portt** (University of Toronto)
Flexible Deadlines for Written Assessments / Échéances flexibles pour les évaluations écrites

14:45-15:00 **Sohee Kang** (University of Toronto Scarborough)
Giving Students Choice: A Flexible Weight for Final Project / Donner le choix aux étudiants : une pondération flexible pour les projets finaux

---

13:30-15:00 **Poster / Poster** (abstract/résumé 153) **CSF Whale Atrium**

Case Study 1: Examining Graduate Student Perspectives on Quality of Supervision, Program and University Experiences

Étude de cas 1 : Examen des perspectives des étudiants diplômés sur la qualité de la supervision, du programme et de l'expérience universitaire

Chair/Président: Chel Hee Lee
Organizer/Responsable: Chel Hee Lee
Sponsor/Commanditaires: Award for Case Studies in Data Analysis Committee/Comité du prix pour les études de cas en l'analyse de données

13:30-15:00 **Amanda Qiu** (University of Victoria) **Jingtong Hu** (University of Victoria) **Jindi Huang** (University of Victoria) **Mingyang Chen** (University of Victoria)
University of Victoria 1 / Université de Victoria 1

13:30-15:00 **Hunter Pozzebon** (University of Toronto Dalla Lana School of Public Health) **Harieswar Sundaram** (University of Toronto) **Xueer Bi** (University of Toronto) **XiaoXuan Han** (University of Toronto)
University of Toronto 1 / Université de Toronto 1

13:30-15:00 **Aadesh Warren Nunkoo** (University of Prince Edward Island) **Paul Alexander Seward** (University of Prince Edward Island) **Mobasherah Falak** (University of Prince Edward Island) **Maleeha Haris** (University of Prince Edward Island)
University of Prince Edward Island / Université de l'Île-du-Prince-Édouard

13:30-15:00 **Winner Pathak** (University of Manitoba) **Avanthi Moragammana Gedara** (University of Manitoba) **Thimani Dananjana Ranathungage** (University of Manitoba) **Jervis Gallanosa** (University of Manitoba)
University of Manitoba 1 / Université du Manitoba 1

13:30-15:00 **Helen Bian** (McGill University) **Rubiya Akter** (McGill University) **Qicheng Zhao** (McGill University)
McGill University 1 / Université McGill 1

13:30-15:00 **Sara Haroon** (Carleton University) **Christiana Koebel** (Carleton University) **Yuliya Nesterova** (Carleton University)
Carleton University / Université Carleton

13:30-15:00 **Jingwen Ji** (University of Toronto) **Ruiyang Wang** (University of Toronto) **Ruochen Zhao** (University of Toronto) **Yanyue Zhang** (University of Toronto)
University of Toronto 2 / Université de Toronto 2

13:30-15:00 **Larry Dong** (University of Toronto Dalla Lana School of Public Health) **George Stefan** (University of Toronto) **Fatema Tuj Johara** (University of Toronto Dalla Lana School of Public Health)
University of Toronto 3 / Université de Toronto 3

13:30-15:00 **Poster / Poster** (abstract/résumé 155) **CSF Whale Atrium**

Case Study 2: Predicting Length of ICU Stay in People with Acute Traumatic Spinal Cord Injury
Étude de cas 2 : Prévision de la durée du séjour en USI des personnes souffrant d'une lésion traumatique aiguë de la moelle épinière

Chair/Président: Chel Hee Lee
Organizer/Responsable: Chel Hee Lee
Sponsor/Commanditaires: Award for Case Studies in Data Analysis Committee/Comité du prix pour les études de cas en l'analyse de données

13:30-15:00 **Balage Don Harshani Hiranthika De Silva** (University of Manitoba) **Samuel Morrissette** (University of Manitoba) **Ashani N. Wickramasinghe** (University of Manitoba) **Nayanthi Karunanayake** (University of Manitoba)
University of Manitoba 2 / Université du Manitoba 2

13:30-15:00 **Siqi Cheng** (McGill University) **Sebastian Garneau** (McGill University) **Yu Gu** (McGill University)
McGill University 2 / Université McGill 2

13:30-15:00 **Alysha Cooper** (University of Guelph) **Patrick McMillan** (University of Guelph) **Madeline Ward** (University of Calgary)
University of Guelph / University of Calgary / Université de Guelph / Université de Calgary

13:30-15:00 **Nasim Feizinazhadgheshlaghi** (University of Manitoba) **Elham Afzali** (University of Manitoba) **Funmilola Mary Taiwo** (University of Manitoba) **Bahram Moeinianfar** (University of Manitoba)
University of Manitoba 3 / Université du Manitoba 3

13:30-15:00 **Linke Li** (University of Toronto Dalla Lana School of Public Health) **Jasper Zhongyuan Zhang** (University of Toronto) **Ziqian Zhuang** (University of Toronto) **Mei Dong** (University of Toronto)
University of Toronto 4 / Université de Toronto 4

13:30-15:00 **Myron Moskalyk** (University of Toronto Dalla Lana School of Public Health) **Zhaoyu Ding** (University of Toronto) **Yan Yi Li** (University of Toronto) **Jinyu Luo** (University of Toronto)
University of Toronto 5 / Université de Toronto 5

13:30-15:00 **Amin Abed** (University of Manitoba) **Md. Hasan** (University of Manitoba) **Narges Amiri** (University of Manitoba) **Justin Dyck** (University of Manitoba)
University of Manitoba 4 / Université du Manitoba 4

13:30-15:00 **Nam-Anh Tran** (McGill University) **Kent Lu** (McGill University) **Mingchi Xu** (McGill University)
McGill University 3 / Université McGill 3

13:30-15:00 **Priyonto Saha** (University of Toronto Dalla Lana School of Public Health) **Xueying Han** (University of Toronto) **Youxue Ren** (University of Toronto) **Yucheng Jiang** (University of Toronto)
University of Toronto 6 / Université de Toronto 6

13:30-15:00 **Xiao Yan** (University of Toronto Dalla Lana School of Public Health) **Yixiao Chen** (University of Toronto)
University of Toronto 7 / Université de Toronto 7

13:30-15:00 **Hao He** (University of Ottawa) **Xiao Liang** (University of Ottawa) **Yuewen Pan** (University of Ottawa) **Chang Qu** (University of Ottawa)
University of Ottawa / Université d'Ottawa

13:30-15:00 **Yacine Marouf** (University of Toronto) **Yutong Lu** (University of Toronto) **Hongyan Chen** (University of Toronto) **Haiqi Yang** (University of Toronto)
University of Toronto 8 / Université de Toronto 8

13:30-15:00 **Chen Chen** (University of Toronto) **Hongyu Chen** (University of Toronto) **Kiara Wu** (University of Toronto) **Zunaira Mehmood** (University of Toronto)
University of Toronto 9 / Université de Toronto 9

13:30-15:00 **Abdulaziz Sherif** (University of Toronto Dalla Lana School of Public Health) **Feifan Xiang** (University of Toronto) **Rachel Yeung** (University of Toronto) **Yankai Feng** (University of Toronto)
University of Toronto 10 / Université de Toronto 10

13:30-15:00 **Brynn O'Connell** (MacEwan University) **Alex Lyndon** (MacEwan University) **Adrian Neumann** (MacEwan University) **Stuart Dovey** (MacEwan University)
MacEwan University / Université MacEwan

13:30-15:00 **Jiachen Pan** (University of Western Ontario) **Chengqian Xian** (University of Western Ontario) **Jingyu Tu** (University of Western Ontario) **Xianglong Fu** (University of Western Ontario)
University of Western Ontario 1 / Université de Western Ontario

13:30-15:00 **Ellen Song** (Western University) **Yiming Hu** (University of Western Ontario) **Yini Cheng** (University of Western Ontario) **Hanrui Dou** (University of Western Ontario)
University of Western Ontario 2 / Université de Western Ontario 2

13:30-15:00 **Alexandra Mossman** (University of Waterloo) **Bryn Candles** (University of Waterloo) **Megan French** (University of Waterloo)
University of Waterloo / Université de Waterloo

15:30-17:00 **Invited / Sur invitation** (abstract/résumé 158) **A 1043**

Machine Learning in Actuarial Science and Finance
Apprentissage automatique en science actuarielle et finance

Chair/Président: Chengguo Weng
Organizer/Responsable: Chengguo Weng
Sponsor/Commanditaires: Actuarial Science Section/Groupe de science actuarielle

15:30-16:00 **Frédéric Godin** (Concordia University) **Andrei Neagu** (Concordia University) **Leila Kosseim** (Concordia University) **Clarence Simard** (Université du Québec à Montréal)
Deep Hedging under Imperfect Liquidity / Stratégie de couverture profonde en présence de liquidité imparfaite

16:00-16:30 **Shu Li** (Western University)
Use of Prediction Bias in Active Learning for Variable Annuity Portfolio Valuation / Utilisation d'un biais de prédiction dans l'apprentissage actif pour l'évaluation de portefeuille de rente variable

16:30-17:00 **Himchan Jeong** (Simon Fraser University) **Hashan Peiris** (Simon Fraser University) **Jae-Kwang Kim** (Iowa State University) **Hangsuck Lee** (Sungkyunkwan University)
Integration of Traditional and Telematics Data for Efficient Insurance Claims Prediction / Intégration de données traditionnelles et télématiques pour la prédiction efficace de réclamations d'assurance

---

15:30-17:00 **Invited / Sur invitation** (abstract/résumé 160) **ED 2018A**

Presenting your technical work to non-statistical audiences
Présenter son travail technique à des publics non statisticiens

Chair/Président: Eleanor M. Pullenayegum
Organizer/Responsable: Eleanor M. Pullenayegum
Sponsor/Commanditaires: Biostatistics Section/Groupe de biostatistique

15:30-16:00 **Aasthaa Bansal** (University of Washington)
A Framework for Personalizing the Timing of Surveillance Testing / Cadre de personnalisation d'un calendrier de tests de surveillance

16:00-16:30 **Aya A. Mitani** (University of Toronto) **Sean Xinyang Feng** (University of Toronto) **Elizabeth Kaye** (Boston University)
Modelling Time-Varying Risk Factors of Tooth Loss: Results From Joint Model Compared With Extended Cox Regression Model / Modélisation des facteurs de risque de perte de dents variant dans le temps : résultats du modèle conjoint comparés à ceux du modèle de régression de Cox étendu

---

15:30-17:00 **Invited / Sur invitation** (abstract/résumé 162) **C 2045**

Embedding Equity, Diversity, and Inclusion in Statistical Research and Practice (Panel)
Intégrer l'équité, la diversité et l'inclusion dans la recherche et la pratique statistiques (Table ronde)

Chair/Président: Michael Wallace
Organizer/Responsable: Michael Wallace
Sponsor/Commanditaires: Committee on Equity, Diversity and Inclusion/Comité pour l'équité, la diversité et l'inclusion

15:30-17:00 **Josée Dupuis** (McGill University) **Tolulope Sajobi** (University of Calgary) **Bei Jiang** (University of Alberta)
Embedding Equity, Diversity, and Inclusion in Statistical Research and Practice / Intégrer l'équité, la diversité et l'inclusion dans la recherche et la pratique statistiques Ⓔ E

15:30-17:00 **Invited / Sur invitation** (abstract/résumé 163) **A 1049**

Recent Advances in Random Walks
Avancées récentes en marches aléatoires

Chair/Président: Jean Vaillancourt
Organizer/Responsable: Jean Vaillancourt
Sponsor/Commanditaires: Probability Section/Groupe de probabilité

15:30-16:00 **Deli Li** (Lakehead University) **Andrew Rosalsky** (University of Florida, USA)
Some Limit Theorems for Negative Quadrant Dependent Random Variables / Des théorèmes limites pour des variables aléatoires négativement dépendantes d'un quadrant Ⓔ E

16:00-16:30 **Hélène Guérin** (Université du Québec à Montréal)
Elephant Random Walk, Polya Urns and Asymptotic Behavior / La marche aléatoire de l'éléphant, les urnes de Polya et leur comportement asymptotique Ⓔ E

16:30-17:00 **Lucile Laulin** (Université Paris Nanterre) **Alice Contat** (Université Sorbonne Paris Nord)
Scaling Limit for Amnesic Step-Reinforced Random Walks / Limite d'échelle pour les marches aléatoires renforcées amnésiques Ⓔ E

15:30-17:00 **Invited / Sur invitation** (abstract/résumé 165) **A 2071**

Past, Present, and Future of Statistical Education
Le passé, le présent et l'avenir de l'enseignement de la statistique

Chair/Président: Yildiz Yilmaz
Organizer/Responsable: Yildiz Yilmaz
Sponsor/Commanditaires: Statistical Education Section/Groupe d'éducation en statistique

15:30-16:00 **Richard J. Cook** (University of Waterloo)
Reflections on Some Experiential Graduate Training Programs in Biostatistics / Réflexions sur certains programmes universitaires de formation expérimentale en biostatistique Ⓔ E

16:00-16:30 **Alison L. Gibbs** (University of Toronto)
An Undergraduate Program in Statistics: Past Influences, Current Curriculum, and Future Questions / Programme de statistique de premier cycle universitaire : influences du passé, programme actuel et questions sur l'avenir Ⓔ E

16:30-17:00 **Donald Estep** (Simon Fraser University/CANSSI) **Donald Estep** (Canadian Statistical Sciences Institute and Simon Fraser University)
The Vision for a CANSSI Research Training Library: Providing Canadian Students with Cutting Edge Statistical Knowledge and Skills / Objectif d'une bibliothèque de formation à la recherche de l'INCASS : fournir aux étudiants canadiens des connaissances et des compétences statistiques de pointe Ⓔ E

15:30-17:00 **Invited / Sur invitation** (abstract/résumé 167) **A 2065**

Innovative Strategies in High-Dimensional Data Analysis with Applications to Business and Industry
Stratégies innovantes en analyse de données à haute dimension avec les applications aux entreprises et à l'industrie

Chair/Président: Armin Hatefi
Organizer/Responsable: S. Ejaz Ahmed
Sponsor/Commanditaires: Business and Industrial Statistics Section/Groupe de statistique industrielle et de gestion

15:30-16:00 **Anand N Vidyashankar** (George Mason University) **Crissa Marshburn** (McKesson Corporation)
Assessing Privacy and Security Risk and Mitigation Strategies / Évaluation des risques pour la vie privée et la sécurité et stratégies d'atténuation Ⓔ E

16:00-16:30 **Yi Li** (University of Michigan)
Penalized Deep Partially Linear Cox Models / Modèles de Cox partiellement linéaires profonds pénalisés Ⓔ E

16:30-17:00 **S. Ejaz Ahmed** (Brock University)
Post-Shrinkage Strategies in Semiparametric Models for High-Dimensional Data Application / Stratégies post-rétrécissement dans les modèles semi-paramétriques pour l'application à des données de forte dimension Ⓔ E

15:30-17:00 **Invited / Sur invitation** (abstract/résumé 169) **IIC 2001**

Isobel Loutit Invited Address
Allocution Isobel-Loutit

Chair/Président: Farouk Nathoo
Organizer/Responsable: Farouk Nathoo
Sponsor/Commanditaires: Business and Industrial Statistics Section/Groupe de statistique industrielle et de gestion

15:30-17:00 **Johanna G. Nešlehová** (McGill University)
Can Bayes Spaces Help to Detect Anomalous Risk Element Concentrations in Agricultural Soils? / Les espaces de Bayes peuvent-ils aider à détecter des concentrations anormalement élevées d'éléments dangereux dans les sols agricoles ? Ⓔ E F

15:30-17:00 **Invited / Sur invitation** (abstract/résumé 170) **A 1045**

Advances in Modern Data Analysis Techniques
Progrès en techniques modernes d'analyse des données

Chair/Président: Peijun Sang
Organizer/Responsable: Peijun Sang

15:30-16:00 **Linglong Kong** (University of Alberta) **Bei Jiang** (University of Alberta)
Gaussian Differential Privacy on Riemannian Manifolds / Confidentialité différentielle gaussienne appliquée à des variétés riemanniennes Ⓔ E

16:00-16:30 **Dehan Kong** (University of Toronto)
LLOT: application of Laplacian Linear Optimal Transport in Spatial Transcriptome Reconstruction / Application du transport optimal linéaire laplacien (LLOT) à la reconstruction spatiale du transcriptome Ⓔ E

16:30-17:00 **Gregory Rice** (University of Waterloo) **Sebastian Kuhnert** (University of California at Davis) **Alexander Aue** (University of California at Davis) **Jeremy VanderDoes** (University of Waterloo)
An operator-level functional GARCH model / Un modèle GARCH fonctionnel au niveau de l'opérateur Ⓔ E

15:30-17:00 **Invited / Sur invitation** (abstract/résumé 172) **A 1046**

Ensemble Learning for Developing Predictive Models
Apprentissage d'ensemble pour le développement de modèles prédictifs

Chair/Président: Rob Deardon
Organizer/Responsable: Jabed H Tomal
Sponsor/Commanditaires: Data Science and Analytics Section/Groupe de science des données et analytique

15:30-16:00 **Thomas M. Loughin** (Simon Fraser University) **Jiahao Tian** (Simon Fraser University) **Hugh Chipman** (Acadia University)
MLCBART: Multilabel Classification with Bayesian Additive Regression Trees / MLCBART : Classification multi-étiquettes avec arbres de régression additifs bayésiens Ⓔ E

16:00-16:30 **Geoff Pleiss** (The University of British Columbia)
Ensembles in the Age of Overparameterization: Promises and Pathologies / Les méthodes d'ensemble à l'ère du surparamétrage : promesses et pathologies Ⓔ E

16:30-17:00 **Jabed H Tomal** (Thompson Rivers University)
Ensemble Learning Based on Subsets of Variables / Apprentissage d'ensemble basé sur des sous-ensembles de variables Ⓔ E

15:30-17:00 **Contributed / Communications libres** (abstract/résumé 174) **C 4036**

Epidemiology and Health
Épidémiologie et santé

Chair/Président: Zelalem Firisa Negeri

15:30-15:45 **Jianan Peng** (Acadia University)
Simultaneous Identifications of the Minimum Effective Dose in Each of Several Groups by the DR Method / Identifications simultanées de la dose minimale efficace dans chacun de plusieurs groupes par la méthode DR Ⓔ E

15:45-16:00 **Christopher Gravel** (University of Ottawa) **Muhammad Mullah** (University of Ottawa)
Evaluation of Phenomenological Models for the Short-term Forecasting of Daily COVID-19 Case Incidence in Canada in the Presence of Multiple Waves / Évaluation de modèles phénoménologiques pour la prévision à court terme de l'incidence des cas de la COVID-19 au Canada en présence de vagues multiples Ⓔ E

16:00-16:15 **Wei Liu** (York University) **Dongwei Wei** (York University)
Semiparametric Nonlinear Mixed-Effects Models with Covariate Measurement Errors and Change Points, with Application to AIDS Studies / Modèles semi-paramétriques à effets mixtes non linéaires avec erreurs de mesure de la covariable et points de rupture, appliqués à des études sur le sida Ⓔ E

16:15-16:30 **Samuel Perreault** (University of Toronto) **Gracia Y. Dong** (University of Toronto) **Alex Stringer** (University of Waterloo) **Hwashin Shin** (Health Canada) **Patrick Brown** (University of Toronto)
Case-Crossover Designs and Overdispersion with Application in Air Pollution Epidemiology / Analyses cas-croisé et surdispersion avec application en épidémiologie de la pollution atmosphérique Ⓔ E

15:30-17:00 **Contributed / Communications libres** (abstract/résumé 177) **C 3053**

Recent Developments in Survey Methods 2
Développements récents en méthodes d'enquête 2

Chair/Président: Daniel J. McDonald

15:30-15:45 **Martin St-Pierre** (Statistics Canada) **Samer Farfour** (Statistics Canada)
The Sampling and Weighting Methodology of the 2021 Census Undercoverage Study / La méthodologie d'échantillonnage et de pondération de l'Étude sur le sous-dénombrement du recensement de 2021 Ⓔ E F

15:45-16:00 **Anne Mather** (Statistics Canada) **Cilanne Boulet** (Statistics Canada) **Golshid Chatrchi** (Statistics Canada) **Andrew Brennan** (Statistics Canada)
Subsampling for Non-response Follow-up in Social Surveys: A Case Study / Sous-échantillonnage pour le suivi de la non-réponse dans les enquêtes sociales : Une étude de cas Ⓔ E

16:00-16:15 **Ariane Boivin** (Université Laval) **Anne-Sophie Charest** (Université Laval)
Between Quality and Confidentiality: Generation of Robust Synthetic Data / Entre qualité et confidentialité : génération de données synthétiques robustes ⒺⒻ E

16:15-16:30 **Alexander Imbrogno** (Statistics Canada)
Using Non-Binary Gender to Calibrate Survey Weights for the Canadian Long-Form Census Sample / Utilisation du genre non-binaire pour le calage de l'échantillon du questionnaire détaillé du Recensement Canadien Ⓔ E

16:30-16:45 **Oluwagbohunmi Adetunji Awosoga** (University of Lethbridge) **Nse Odunaiya** (University of Ibadan, College of Medicine, Nigeria) **Adesola Odole** (University of Ibadan, College of Medicine, Nigeria) **Olufemi Oyewole** (Olabisi Onabanjo University, Teaching Hospital, Nigeria) **Mercy Adegoke** (University of Ibadan, College of Medicine, Nigeria) **Chiedozie Alumona** (University of Lethbridge) **Ogochukwu Onyeso** (University of Lethbridge) **Abiodun Adeoye** (University of Ibadan, College of Medicine, Nigeria) **Happiness Aweto** (University of Lagos, Lagos University Teaching Hospital, Nigeria)
Cardiovascular Disease Perception and Risk among Community-Dwelling Adults in Southwest Nigeria / Perception et risque de maladie cardiovasculaire chez les adultes vivant en communauté dans le sud-ouest du Nigeria Ⓔ E

16:45-17:00 **Mohammed Sanda** (Social Security and National Insurance Trust)
Revitalizing Social Security Systems through Innovative Survey Methods: A Case Study of SSNIT's Transformational Journey / Étude de cas sur le parcours transformationnel du SSNIT : modernisation des systèmes de sécurité sociale par des méthodes d'enquête novatrices Ⓔ E

15:30-17:00 **Contributed / Communications libres** (abstract/résumé 181) **ED 2018B**

Regression Analysis
Analyse de régression

Chair/Président: Yanglei Song

15:30-15:45 **Qing Wang** (Ca' Foscari University of Venice) **Roberto Casarin** (Ca' Foscari University of Venice) **Radu Craiu** (University of Toronto)
Markov Switching Tensor Regression / Régression tensorielle à commutation de Markov Ⓔ E

15:45-16:00 **Hedayat Fathi** (Université Laval) **Marzia A. Cremona** (Université Laval) **Federico Severino** (Université Laval)
Selection of Functional Predictors and Smooth Coefficient Estimation for Scalar-on-function Regression Models / Sélection de prédicteurs fonctionnels et estimation lisse des coefficients pour les modèles de régression scalaire-sur-fonction Ⓔ E

16:00-16:15 **Jonathan Jalbert** (Polytechnique Montreal) **Auguste Paoli** (Polytechnique Montréal)
Goodness-of-Fit Tests for Multivariate Scaling Models of IDF Curves / Tests d'adéquation pour les modèles de scaling des courbes IDF Ⓔ E

16:15-16:30 **Ziang Zhang** (University of Toronto) **Patrick Brown** (University of Toronto) **Jamie Stafford** (University of Toronto)
Unveiling Quasi-Periodic Patterns with Seasonal Gaussian Processes / Dévoilement de tendances quasi périodiques à l'aide de processus saisonniers gaussiens Ⓔ E

16:30-16:45 **Marcus Hlady** (University of Manitoba) **Yuliya V. Martsynyuk** (University of Manitoba)
Estimators of Reliability Ratio, Their Asymptotic Properties and Use for Inference in Linear Structural Errors-in-Variables Models / Estimateurs du ratio de fiabilité, propriétés asymptotiques et utilisation pour l'inférence dans les modèles structurels linéaires avec erreur sur les variables

---

15:30-17:00 **Contributed / Communications libres** (abstract/résumé 184) **C 2033**

Causal Inference
Inférence causale

Chair/Président: William Ruth

15:30-15:45 **Glen McGee** (University of Waterloo) **Lan Wen** (University of Waterloo)
Estimating Average Causal Effects with Incomplete Exposure and Confounders / Estimation des effets causaux moyens en cas d'exposition incomplète et de facteurs de confusion

15:45-16:00 **Yuliang Shi** (University of Waterloo) **Yeying Zhu** (University of Waterloo) **Joel A. Dubin** (University of Waterloo)
Causal Inference on Missing Exposure via Robust Estimation / Inférence causale sur l'exposition manquante au moyen d'une estimation robuste

16:00-16:15 **Xiaoya Wang** (University of Waterloo) **Richard J. Cook** (University of Waterloo) **Yeying Zhu** (University of Waterloo)
Two-stage Regression for Causal Inference Involving Semi-continuous Exposures and Two-dimensional Propensity Scores / Régression en deux étapes pour l'inférence causale impliquant des expositions semi-continues et des scores de propension bidimensionnels

16:15-16:30 **Henan Xu** (University of Waterloo) **Yeying Zhu** (University of Waterloo)
Functional Mediation Analysis with Zero-inflated Count Data / Analyse fonctionnelle de la médiation à partir de données de dénombrement avec excès de zéros

16:30-16:45 **Sumeet Kalia** (University of Manitoba) **Olli Saarela** (University of Toronto) **Michelle Greiver** (University of Toronto) **Frank Sullivan** (University of St. Andrews)
Continuous-time Causal Inference With Marked Point Process Weights: An Example on Sodium-Glucose Co-Transporters 2 Inhibitor Medications and Urinary Tract Infection / Inférence causale en temps continu avec des poids de processus ponctuels marqués : exemple des médicaments inhibiteurs du cotransporteur sodium-glucose de type 2 contre l'infection urinaire

---

15:30-17:00 **Contributed / Communications libres** (abstract/résumé 187) **C 3033**

Biostatistics Student Research Session #3
Session de recherche étudiante en biostatistique #3

Chair/Président: Qihuang Zhang

15:30-15:45 **Yu Shi** (University of Toronto Dalla Lana School of Public Health)
Unsupervised Deep Domain Adaptation for Predicting Patient-Specific Cancer Dependency Maps / Adaptation profonde non supervisée par domaine pour prédire les cartes de dépendance du cancer propres aux patients

15:45-16:00 **Muditha L. Bodawatte Gedara** (University of Manitoba) **Lisa M. Lix** (Department of Community Health Sciences, University of Manitoba) **Ridwan Sanusi** (Smart Mobility and Logistics, King Fahd University of Petroleum & Minerals) **Tolulope Sajobi** (Department of Community Health Sciences, University of Calgary)
Comparison of Cross-validation Methods for Tree-based Item-Focused Models to Detect Differential Item Functioning in Patient-reported Outcome Measures / Comparaison de méthodes de validation croisée pour des modèles d'arbres axés sur les items (IFT) pour la détection du fonctionnement différentiel d'un item (DIF) dans les mesures de résultats rapportés par les patients

16:00-16:15 **Ziqian Zhuang** (University of Toronto Dalla Lana School of Public Health) **Wei Xu** (University Health Network, Biostatistics)
Joint Modeling of Complex Multivariate Adverse Events in Clinical Trial Data / Modélisation conjointe des événements indésirables multivariés complexes dans les données d'essais cliniques

16:15-16:30 **Aoqi Xie** (University of Toronto) **Peijin Wang** (School of Medicine, Duke University) **Aya A. Mitani** (Dalla Lana School of Public Health, University of Toronto) **Madison Aitken** (Department of Psychology, York University; Centre for Addiction and Mental Health) **Wendy Lou** (Dalla Lana School of Public Health, University of Toronto) **Clement Ma** (Dalla Lana School of Public Health, University of Toronto; Centre for Addiction and Mental Health)
A novel Sequential Multiple Assignment Randomized Trial (SMART) design with Internal Pilot Study and Unblinded Sample Size Re-estimation / Un nouveau concept d'essai randomisé séquentiel à évaluation multiple (SMART) avec une étude pilote interne et une réestimation de la taille d'échantillon non aveugle

16:30-16:45 **Qirui (Dylan) Hou** (University of Toronto) **Amy Liu** (Princess Margaret Cancer Centre, University Health Network) **Peter Szatmari** (Centre for Addiction and Mental Health) **Clement Ma** (Centre for Addiction and Mental Health)
A Novel Multi-Arm, Two-stage Basket Design / Nouveau modèle de panier à deux étapes et à bras multiples

16:45-17:00 **Sirikkathuge Ishanka Randini Fernando** (McMaster University)
Time-aligned Latent Dirichlet Allocation for Longitudinal Microbiome Data / Répartition de Dirichlet latente alignée dans le temps pour les données longitudinales de microbiome

Tuesday June 4 — mardi 4 juin

08:30-09:50 **Invited / Sur invitation** (abstract/résumé 191) **IIC 2001**

CRM-SSC Prize in Statistics Invited Address
Allocution du recipiendaire du Prix CRM-SSC en statistique

Chair/Président: Erica E. M. Moodie
Organizer/Responsable: Erica E. M. Moodie

08:30-09:50 **Alexandre Bouchard-Côté** (University of British Columbia)
Computational Lebesgue integration / Intégration computationelle de Lebesgue

---

10:20-11:50 **Invited / Sur invitation** (abstract/résumé 192) **A 1045**

Statistical Modelling and Computational Intelligence for Complex Data in Medical Research
Modélisation statistique et intelligence informatique des données complexes en recherche médicale

Chair/Président: You Liang
Organizer/Responsable: You Liang

10:20-10:50 **Longhai Li** (University of Saskatchewan)
Z-Residual Diagnostic Tool for Assessing Covariate Functional Form in Proportional Hazards Models with Shared Frailty / Outil diagnostique Z-résiduel pour l'évaluation de la forme fonctionnelle des covariables dans les modèles de risques proportionnels avec fragilité partagée

10:50-11:20 **Li Xing** (University of Saskatchewan)
Concurrent Prediction of Multiple Survival Outcomes with a Refined Stacking Algorithm / Prédiction concurrente de plusieurs résultats de survie avec un algorithme d'empilage raffiné

10:20-11:50 **Invited / Sur invitation** (abstract/résumé 194) **C 2033**

Innovative Design and Analysis of Clinical Trials
Conception et analyse innovantes des essais cliniques

Chair/Président: Yanqing Yi
Organizer/Responsable: Yanqing Yi

10:20-10:50 **Grace Y. Yi** (University of Western Ontario) **Yasin Khadem Charvadeh** (University of Western Ontario)
Accommodating Misclassification Effects on Optimizing Dynamic Treatment Regimes with Q-Learning / Accommodation des effets de l'erreur de classification pour l'optimisation des régimes de traitement dynamique avec Q-learning

10:50-11:20 **Xikui Wang** (University of Manitoba)
Bayesian Adaptive Design of Phase I Clinical Trials / Conception adaptative bayésienne d'essais cliniques de phase I

10:20-11:50 **Invited / Sur invitation** (abstract/résumé 195) **SN 2109**

Mortality Forecasting and Longevity Risk Management
Prévision de la mortalité et gestion du risque de longévité

Chair/Président: Yingli Qin
Organizer/Responsable: Yingli Qin
Sponsor/Commanditaires: Actuarial Science Section/Groupe de science actuarielle

10:20-10:50 **Liqun Diao** (University of Waterloo) **Yechao Meng** (University of Prince Edward Island) **Chengguo Weng** (University of Waterloo)
Mortality Prediction via Age-Specific Band Selection / Prévision de la mortalité par sélection de bandes spécifiques à l'âge

10:50-11:20 **Hong Li** (University of Guleph) **David Landriault** (University of Waterloo) **Bin Li** (University of Waterloo) **Yuanyuan Zhang** (University of Waterloo)
Risk Aversion and Longevity Risk Transfers: Reinsurance vs. Capital Market Solutions / Aversion pour le risque et transfert du risque de longévité : réassurance et solutions du marché des capitaux

11:20-11:50 **Yechao Meng** (University of Prince Edward Island)
Mortality Prediction: a Parameter Transfer Approach / Prédiction de mortalité : une approche de transfert de paramètre

---

10:20-11:50 **Invited / Sur invitation** (abstract/résumé 197) **A 1046**

Teaching Introductory Statistics to Non-specialists
Enseigner l'introduction à la statistique à des non-spécialistes

Chair/Président: Léo Belzile
Organizer/Responsable: Léo Belzile
Sponsor/Commanditaires: Statistical Education Section/Groupe d'éducation en statistique

10:20-10:50 **Tiffany A. Timbers** (The University of British Columbia)
Reflections on Scaling an Introduction to Data Science / Réflexions sur la généralisation d'une introduction à la science des données

10:50-11:20 **Nathalie Moon** (University of Toronto)
Principles and Practices in Teaching STA130: Introduction to Statistical Reasoning and Data Science at the University of Toronto - A Collaborative Partnership Approach / Principes et pratiques de l'enseignement de STA130 : Introduction au raisonnement statistique et à la science des données à l'Université de Toronto - Un partenariat collaboratif

11:20-11:50 **Carolyn Augusta** (University of Saskatchewan)
One Size Does Not Fit All / Tous n'entrent pas dans le même moule

---

10:20-11:50 **Invited / Sur invitation** (abstract/résumé 199) **ED 2018A**

Addressing Practical Challenges in Longitudinal Causal Inference
Relever les défis pratiques de l'inférence causale longitudinale

Chair/Président: Mireille Schnitzer
Organizer/Responsable: Arthur Chatton, Mireille Schnitzer

10:20-10:50 **Mohammad Ehsanul Karim** (The University of British Columbia) **Lucy Mosquera** (University of British Columbia) **Md Belal Hossain** (University of British Columbia)
Properties of inverse probability of adherence weighted estimator of the per-protocol effect for sustained treatment strategies under different data-generating mechanisms and adherence patterns / Propriétés de l'estimateur de pondération par probabilité inverse d'adhésion de l'effet per-protocole pour les stratégies de traitement soutenu dans le cadre de différents mécanismes de génération de données et de modèles d'adhésion

10:50-11:20 **Arthur Chatton** (Université de Montréal) **Robert W. Platt** (McGill University) **Michael Schomaker** (Ludwig-Maximilans-Universität München, Germany) **Miguel-Angel Luque-Fernandez** (University of Granada, Spain) **Mireille Schnitzer** (Université de Montréal)
A Diagnostic Tool for Sequential Positivity Violations in Longitudinal Causal Inference / Un outil de diagnostic pour les violations de positivité séquentielle dans l'inférence causale longitudinale

11:20-11:50 **Eleanor M. Pullenayegum** (Hospital for Sick Children)
Causal Inference With Longitudinal Data Subject to Irregular Assessment Times / Inférence causale avec des données longitudinales soumises à des temps d'évaluation irréguliers

---

10:20-11:50 **Contributed / Communications libres** (abstract/résumé 201) **C 3033**

Quantitative Finance and Financial Econometrics
Finance quantitative et économétrie financière

Chair/Président: Po Yang

10:20-10:35 **Mark Reesor** (Wilfrid Laurier Universtiy) **Mark Drmac Walid Mnif Arie Zeldenrijk**
Incorporating Climate Risk into Portfolio Credit Risk Models via Distortion / Intégrer le risque climatique dans des modèles de risque de portefeuille de crédit par distorsion Ⓔ E

10:35-10:50 **Manal Teto** (University of Ottawa)
Dynamic Programming Approach to Price a Panel of American Options / Approche de programmation dynamique pour fixer le prix d'un panel d'options américaines Ⓔ E

10:50-11:05 **Esam Mahdi** (Carleton University)
New Mixed Portmanteau Tests for Time Series Models / Nouveaux tests portmanteau mixtes pour les modèles de séries temporelles Ⓔ E

11:05-11:20 **Haixu Wang** (University of Calgary) **Jiguo Cao** (Simon Fraser University)
Nonlinear Prediction of Functional Time Series / Prédiction non linéaire de séries temporelles fonctionnelles Ⓔ E

11:20-11:35 **Maciej Augustyniak** (Université de Montréal) **Alexandru Badescu** (University of Calgary) **Jean-François Bégin** (Simon Fraser University) **Sarath Kumar Jayaraman** (University of Calgary)
On the Relation Between Discrete and Continuous-Time Affine Option Pricing Models / À propos de la relation entre les modèles d'évaluation des options affines à temps continu et ceux à temps discret Ⓔ E

10:20-11:50 **Contributed / Communications libres** (abstract/résumé 204) **C 4036**

Clustering and Machine Learning
Regroupement et apprentissage automatique

Chair/Président: Brian Franczak

10:20-10:35 **Mina Aminghafari** (University of Calgary) **Saeid Hoseinipour Hoseinipour** (Amirkabir University of Technology) **Adel Mohammadpour** (Amirkabir University of Technology) **Mohamed Nadif** (Université Paris Cité)
Exponential Family and Latent Block Model for Co-clustering / Famille exponentielle et modèle de blocs latents pour le Co-clustering ⒺⒻ E

10:35-10:50 **Julie Carreau** (Polytechnique Montreal)
A Spatially Adaptive Multi-resolution Generative Algorithm: Application to Simulating Flood Wave Propagation / Algorithme génératif multi-résolutions spatialement adaptatif : application à la simulation de la propagation des ondes de crue Ⓔ E

10:50-11:05 **Samuel Morrissette** (University of Manitoba) **Saman Muthukumarana** (University of Manitoba) **Maxime Turgeon** (University of Manitoba)
Parsimonious Dirichlet Process Mixture Models for Clustering with Dissimilarities / Modèles de mélange de processus de Dirichlet parcimonieux pour le regroupement avec dissimilarités Ⓔ E

11:05-11:20 **Yi-Shu Lin** (CHES, Sickkids) **Linke Li** (The Hospital for Sick Children) **Anna Heath** (The Hospital for Sick Children) **James O'Mahony** (University College Dublin)
Using Machine Learning Algorithms to Identify Relevant Strategies for Simulation in Cost-Effectiveness Analysis of Screening / Utilisation d'algorithmes d'apprentissage machine (ML) pour identifier des stratégies adéquates de simulation dans l'analyse de la rentabilité (CEA) du dépistage Ⓔ E

11:20-11:35 **Devan G. Becker** (Wilfrid Laurier University)
Defining SARS-CoV-2 Lineages with Temporally Consistent Mutation Clusters in Wastewater Samples / Définition des lignées SARS-CoV-2 avec des grappes de mutations cohérentes dans le temps dans les échantillons d'eaux usées Ⓔ E

11:35-11:50 **Elif Fidan Acar** (University of Manitoba) **Martin Lysy** (University of Waterloo)
Automated Statistical Methods for High-Throughput Phenotyping Experiments / Méthodes statistiques automatisées pour expériences de phénotypage à haut débit Ⓔ E

10:20-11:50 **Invited / Sur invitation** (abstract/résumé 208) **A 1043**

Advanced Development on Time Series and Their Applications
Développements avancés en séries temporelles et applications

Chair/Président: Bruno N. Rémillard
Organizer/Responsable: Bruno N. Rémillard
Sponsor/Commanditaires: Probability Section/Groupe de probabilité

10:20-10:50 **Bouchra Nasri** (Université de Montréal)
Tests of Serial Dependence for Multivariate Time Series with Arbitrary Distributions / Tests de dépendance sérielle pour des séries chronologiques multivariées de distribution arbitraire Ⓔ E

10:50-11:20 **Mohamedou Ould Haye** (Carleton University) **Anne Philippe** (Nantes University)
Inference for Discrete Randomized Linear Processes / Inférence pour processus linéaires aléatoires discrets Ⓔ E

11:20-11:50 **Masoud M. Nasari** (Bank of Canada) **Mohamedou Ould Haye** (Carleton University)
A New Inferential Framework / Un nouveau cadre d'inférence Ⓔ E

10:20-11:50 **Invited / Sur invitation** (abstract/résumé 210) **A 1049**

Statistical Modeling of Complex Medical Research Data
Modélisation statistique de données complexes de recherche médicale

Chair/Président: Neil A. Spencer
Organizer/Responsable: Tessema Astatkie

10:20-10:50 **Demissie Alemayehu** (Columbia University)
An Enriched Approach to Combining High-dimensional Genomic and Low-dimensional Phenotypic Data / Une approche enrichie pour combiner des données génomiques en grande dimension et des données phénotypiques en petite dimension Ⓔ E

10:50-11:20 **Birol Emir** (Columbia Univ)
A Flexible Alternative to Standard Modeling Techniques for Extrapolated Mean Survival Times Needed for Cost-Effectiveness Analyses / Solution de rechange polyvalente aux techniques de modélisation standard pour l'extrapolation des durées moyennes de survie nécessaires aux analyses coût-efficacité Ⓔ E

11:20-11:50 **Javier Cabrera** (Rutgers University)
Differential Projection Pursuit Methods and Their Applications to Differential Experiments / Méthodes de poursuite de la projection différentielle et ses applications aux expériences différentielles Ⓔ E

10:20-11:50 **Invited / Sur invitation** (abstract/résumé 212) **C 2045**

Recent Advances in Epidemiology and Ecology
Progrès récents en épidémiologie et écologie

Chair/Président: Rob Deardon
Organizer/Responsable: Rob Deardon
Sponsor/Commanditaires: Biostatistics Section/Groupe de biostatistique

10:20-10:42 **Jee Yeon (Joanne) Kim** (The Ohio State University) **Andrew B. Lawson** (Medical University of South Carolina)
A Novel Bayesian Spatio-temporal Surveillance Metric to Predict Emerging Infectious Disease High-risk Clusters / Nouvelle mesure bayésienne de surveillance spatio-temporelle pour prédire les nouveaux foyers de contagion à risque élevé

10:42-11:05 **Madeline Ward** (University of Calgary) **Rob Deardon** (University of Calgary) **Lorna Deeth** (University of Guelph) **Caitlin Ward** (University of Minnesota)
Accounting for Behavioural Changes in Epidemic Models / Prendre en compte les changements comportementaux dans les modèles épidémiques

11:05-11:27 **Joanna Elizabeth Mills Flemming** (Dalhousie University)
Exploring Encounter Processes: From Cell-Cell Interactions to Interspecies Dynamics / Exploration des processus de rencontre : des interactions entre cellules aux dynamiques inter-espèces

11:27-11:50 **Laura L.E. Cowen** (University of Victoria)
From Ecology to Epidemiology / De l'écologie à l'épidémiologie

---

10:20-11:50 **Contributed / Communications libres** (abstract/résumé 214) **C 3053**

Survival and Reliability Analysis
Analyse de survie et de fiabilité

Chair/Président: Tolu Sajobi

10:20-10:35 **Yixuan Li** (McGill) **Ariane Marelli** (McGill Adult Unit for Congenital Heart Disease (MAUDE Unit), McGill University of Health Centre) **Yi Yang** (McGill) **Yue Li** (McGill)
MixEHR-SurG: a Joint Proportional Hazard and Guided Topic Model for Inferring Mortality-Associated Topics from Electronic Health Records / MixEHR-SurG : un modèle conjoint à risques proportionnels et à sujets prédéfinis pour inférer des sujets liés à la mortalité à partir de dossiers médicaux électroniques

10:35-10:50 **Laura Bumbulis** (University of Waterloo) **Richard J. Cook** (University of Waterloo)
Testing Process Reliability under a Limit of Detection: Issues of Robustness and Efficiency / Test de fiabilité de processus selon une limite de détection : problèmes de robustesse et d'efficacité

10:50-11:05 **Xianwei Li** (University of Waterloo) **Richard J. Cook** (University of Waterloo) **Liqun Diao** (University of Waterloo)
Prediction for Illness-death Processes under Intermittent Observation / Prévision pour un processus maladie-décès en cas d'observation intermittente

11:05-11:20 **Wenling Zhang** (University of Waterloo) **Cecilia A. Cotton** (University of Waterloo) **Lan Wen** (University of Waterloo)
Targeted Maximum Likelihood and Other Robust Estimators for Recurrent Causal Events / Estimation ciblée du maximum de vraisemblance et autres estimateurs robustes pour événements causaux récurrents

11:20-11:35 **Connie Stewart** (University of New Brunswick) **Tyler Rideout** (University of New Brunswick Saint John) **Matthew Stephenson** (Quantics)
Prey Selection for Fatty Acid Signature Analysis Using the Akaike Information Criterion / Sélection de proie pour l'analyse des signatures des acides gras en utilisant le critère d'information d'Akaike

---

10:20-11:50 **Invited / Sur invitation** (abstract/résumé 217) **A 2071**

Recent Advances By New Investigators Across Canada
Progrès récents réalisés par les nouveaux chercheurs au Canada

Chair/Président: Kevin McGregor
Organizer/Responsable: Kevin McGregor
Sponsor/Commanditaires: Committee on New Investigators/Comité des nouveaux chercheurs

10:20-10:50 **Alexander Shestopaloff** (Memorial University of Newfoundland) **Mihai Cucuringu** (University of Oxford) **Yichi Zhang** (University of Oxford) **Stefan Zohren** (University of Oxford)
Robust Detection of Lead-Lag Relationships in Lagged Multi-Factor Models / Détection robuste de relations lead-lag dans les modèles multifacteurs décalés Ⓔ E

10:50-11:20 **James H. McVittie** (University of Regina)
Survival Analysis Methodologies for Wildlife Studies / Méthodologie d'analyse de survie dans les études sur la faune Ⓔ E

10:20-11:50 **Invited / Sur invitation** (abstract/résumé 219) **A 2065**

On Some Multivariate Distributions and Recent Advances in Robust Inference
Des distributions multivariées et avancées récentes en inférence robuste

Chair/Président: Mai Ghannam
Organizer/Responsable: Mai Ghannam

10:20-10:50 **Sévérien Nkurunziza** (University of Windsor) **Mai Ghannam** (University of Ottawa)
Some Recent Identities in Tensor Elliptically Contoured Distributions and Their Applications / Quelques identités récentes pour des distributions elliptiques tensorielles et leurs applications Ⓕ E

10:50-11:20 **Serge B. Provost** (The University of Western Ontario)
Identities Stemming from Matrix-variate Density Functions / Identités découlant de fonctions de densité des variables matricielles Ⓕ E

11:20-11:50 **Éric P. Marchand** (Université de Sherbrooke)
The search for efficient predictive densities for multivariate data / La recherche de densités prédictives efficientes pour des données multivariées Ⓕ E

10:20-11:50 **Contributed / Communications libres** (abstract/résumé 221) **ED 2018B**

Bayesian Methods
Méthodes bayésiennes

Chair/Président: Joseph Beyene

10:20-10:35 **Thierry Chekouo** (University of Minnesota) **Samuel Babatunde Samuel Babatunde** (University of Calgary) **Samuel Babatunde** (University of Calgary)
A Bayesian Variable Selection for Semicontinuous Response data: Application to cardiovascular disease / Un modèle de sélection bayésien pour des données de réponses semi-continues : application aux maladies cardiovasculaires Ⓔ E

10:35-10:50 **Larry Dong** (University of Toronto Dalla Lana School of Public Health) **Eleanor M. Pullenayegum** (The Hospital for Sick Children) **Olli Saarela** (University of Toronto)
On Bayesian Joint Modelling with Irregularly Observed Data to Estimate Optimal Treatment Regimes / Au sujet des modèles bayésiens conjoints avec des données irrégulières pour l'estimation des régimes de traitement optimaux Ⓔ E

10:50-11:05 **Yushu Zou** (University of Toronto Dalla Lana School of Public Health) **Aya A. Mitani** (Dalla Lana School of Public Health, University of Toronto) **Olli Saarela** (Dalla Lana School of Public Health, University of Toronto) **Kuan Liu** (Dalla Lana School of Public Health, University of Toronto; Institute of Health Policy, Management, and Evaluation, University of Toronto)
A Bayesian Sensitivity Analysis Approach for Unmeasured Confounding in Longitudinal Data / Une approche d'analyse de sensibilité bayésienne pour les variables confondantes non-mesurées pour des données longitudinales Ⓔ E

11:05-11:20 **Wen Teng** (The Hospital for Sick Children) **Niall Ferguson** (University Health Network) **Ewan Goligher** (University Health Network) **Anna Heath** (The Hospital for Sick Children)
Bayesian Joint Modeling for Longitudinal Magnitude Data with Informative Dropout: an Application to Critical Care Data / Modélisation bayésienne conjointe pour données longitudinales d'amplitude avec abandon informatif : une application aux données de soins intensifs Ⓔ E

11:20-11:35 **Michelle F. Miranda** (University of Victoria)
A CANDECOMP/PARAFAC basis for fast Bayesian Estimation of Multi-Subject fMRI / Une base CANDECOMP/PARAFAC pour une estimation bayésienne rapide d'une IRMf à multisujet

11:35-11:50 **Lara Maleyeff** (McGill University) **Shirin Golchi** (McGill University) **Erica Moodie** (McGill University)
An Adaptive Enrichment Design using Bayesian Model Averaging for the Identification of Tailoring Variables / Plan d'enrichissement adaptatif utilisant la moyenne des modèles bayésiens pour l'identification des variables d'adaptation

---

13:30-15:00 **Invited / Sur invitation** (abstract/résumé 225) **ED 2018A**

New Insights and Developments in Mixture Models and Their Applications
Nouvelles perspectives et développements en modèles de mélange et applications

Chair/Président: Zeny Feng
Organizer/Responsable: Zeny Feng

13:30-14:00 **Jiahua Chen** (The University of British Columbia)
Moment Estimator and the Optimal Minimax Convergence Rate / Estimateur de moment et taux de convergence optimal minimax

14:00-14:30 **Sanjeena Dang** (Carleton University) **Andrea Payne** (Carleton University) **Anjali Silva** (University of Toronto) **Steven Rothstein** (University of Guelph) **Paul David McNicholas** (McMaster University)
A Parsimonious Family of Mixtures of Multivariate Poisson Log-Normal Factor Analyzers for Clustering Count Data / Une famille parcimonieuse de mélanges d'analyseurs de facteur log-normal de Poisson multivariés pour le regroupement de données de dénombrement

14:30-15:00 **Pengfei Li** (University of Waterloo) **Tao Yu** (National University of Singapore) **Jing Qin** (National Institutes of Health)
Maximum Binomial Likelihood for Multivariate Mixture Data / Vraisemblance binomiale maximale pour données de mélanges multivariées

---

13:30-15:00 **Invited / Sur invitation** (abstract/résumé 227) **A 1045**

Machine Learning Strategies for Health Science Data
Stratégies d'apprentissage automatique des données des sciences de la santé

Chair/Président: Liqun Diao
Organizer/Responsable: Liqun Diao

13:30-14:00 **Joel A. Dubin** (University of Waterloo) **Minzee Kim** (University of Waterloo) **Tatiana Krikella** (University of Waterloo)
Advances in Similarity-based Predictive Modeling Methods / Progrès dans les méthodes de modélisation prédictive par similarités

14:00-14:30 **Ameer Dharamshi** (University of Washington) **Anna Neufeld** (Fred Hutchinson Cancer Center) **Keshav Motwani** (University of Washington) **Lucy L. Gao** (University of British Columbia) **Daniela Witten** (University of Washington) **Jacob Bien** (University of Southern California)
Data Thinning with Applications in the Health Sciences / Affinage des données avec application en sciences de la santé

14:30-15:00 **Jon Steingrimsson** (Brown University)
Generalizability of Study Results / Généralisabilité des résultats d'études

---

13:30-15:00 **Invited / Sur invitation** (abstract/résumé 229) **A 1049**

Statistical Learning and Decision Making in Biostatistics
Apprentissage statistique et prise de décision en biostatistique

Chair/Président: Xikui Wang
Organizer/Responsable: Xikui Wang

13:30-14:00 **Yanqing Yi** (Memorial University of Newfoundland)
Stochastic Modeling and Optimal Adaptive Design of Clinical Trials / Modélisation stochastique et conception adaptative optimale d'essais cliniques

14:00-14:30 **Wenqing He** (University of Western Ontario) **Grace Y. Yi** (University of Western Ontario) **Raymond Carroll** (Taxas A & M University)
Feature Screening with Large Scale and High Dimensional Survival Data / Sélection de caractéristiques pour données de survie à grande échelle et en grande dimension

14:30-15:00 **You Liang** (Toronto Metropolitan University) **Aleksandar Popovic** (Toronto Metropolitan University) **Na Yu** (Toronto Metropolitan University) **Xun Zhou** (St. Michael's Hospital) **Keanu Uchida** (St. Michael's Hospital) **Tomasz Tkaczyk** (Rice University) **Neeru Gupta** (St. Michael's Hospital; University of Toronto) **Yeni Yucel** (St. Michael's Hospital; University of Toronto)
A Graph-based Semantic Segmentation Algorithm for Hyperspectral Fluorescence Microscopy Imaging Data / Algorithme de segmentation sémantique basé sur les graphes pour données d'imagerie de microscopie à fluorescence hyperspectrale

13:30-15:00 **Invited / Sur invitation** (abstract/résumé 231) **A 2071**

New Development in Functional Data Analysis
Nouveaux développements en analyse des données fonctionnelles

Chair/Président: Jiguo Cao
Organizer/Responsable: Jiguo Cao

13:30-14:00 **Edward Gunning** (University of Pennsylvania) **Giles Hooker** (University of Pennsylvania)
A New Perspective on Principal Differential Analysis / Nouvelle perspective de l'analyse différentielle principale

14:00-14:30 **Luo Xiao** (North Carolina State University) **Ruonan Li** (North Carolina State University)
Latent Factor Model for Multivariate Functional Data / Modèle à facteurs latents pour les données fonctionnelles multivariées

14:30-15:00 **Tianyu Guan** (Brock University) **Shifan Jia** (Simon Fraser University) **Haolun Shi** (Simon Fraser University)
Semiparametric Function-on-function Regression Models / Modèles de régression fonction-sur-fonction semi-paramétriques

13:30-15:00 **Invited / Sur invitation** (abstract/résumé 233) **A 2065**

Recent Advances in Capital Structure Models and Contingent Capital
Avancées récentes en modèles de structure du capital et capital contingent

Chair/Président: Mark Reesor
Organizer/Responsable: Mark Reesor

13:30-14:00 **Francois Michel Boire** (University of Ottawa) **Mark Reesor** (Wilfrid Laurier University) **Hatem Ben-Ameur** (HEC Montréal) **Pascal François** (HEC Montréal) **Lars Stentoft** (University of Western Ontario)
A Dynamic Structural Model for Contingent Convertible Debt / Un modèle dynamique structurel pour des obligations convertibles à coupon

14:00-14:30 **Di Meng** (Wilfrid Laurier University) **Adam Metzler** (Wilfrid Laurier University) **Mark Reesor** (Wilfrid Laurier University)
Capital Structural Models and Contingent Convertible Securities / Modèles structurels de capitaux et titres convertibles contingents Ⓔ E

14:30-15:00 **Joe Campolieti** (Wilfrid Laurier University)
Last Hitting Times, Excursions and Meanderings of Solvable Diffusions / Derniers temps de passage, excursions et méandres des diffusions solubles Ⓔ E

13:30-15:00 **Invited / Sur invitation** (abstract/résumé 235) **SN 2109**

Bridging the Gap: Navigating Collaborative Research with Non-Statisticians (Panel)
Combler le fossé : la recherche collaborative avec des non-statisticiens (table ronde)

Chair/Président: Reza Ramezan
Organizer/Responsable: Reza Ramezan
Sponsor/Commanditaires: Business and Industrial Statistics Section/Groupe de statistique industrielle et de gestion

13:30-15:00 **Daniel J. McDonald** (University of British Columbia) **Tolulope Sajobi** (University of Calgary) **Andrew Irwin** (Dalhousie University) **Martin Lysy** (University of Waterloo) **Mireille Schnitzer** (Université de Montréal)
Bridging the Gap: Navigating Collaborative Research with Non-Statisticians / Combler le fossé : Naviguer la recherche collaborative avec des non-statisticien-ne-s Ⓔ E

13:30-15:00 **Invited / Sur invitation** (abstract/résumé 236) **C 2033**

Probability Models in Finance
Modèles de probabilité en finance

Chair/Président: Marcos Escobar-Anel
Organizer/Responsable: Marcos Escobar-Anel

13:30-14:00 **Marcos Escobar-Anel** (Western University)
Portfolio Optimization in Affine GARCH models / Optimisation de portefeuille dans les modèles affines GARCH Ⓔ E

14:00-14:30 **Bruno N. Rémillard** (HEC Montréal) **Jean Vaillancourt** (HEC Montréal) **Pierre Laroche** (Banque Nationale du Canada)
Parrondo's Paradox and Financial Applications / Le paradoxe de Parrondo et applications financières Ⓔ E

14:30-15:00 **Anatoliy V. Swishchuk** (University of Calgary)
Applications of Geometric Compound Hawkes Process in Finance / Applications du processus de Hawkes composé géométrique en finance Ⓔ E

13:30-15:00 **Invited / Sur invitation** (abstract/résumé 238) **C 2045**

New Approaches to Genetic and Genomic Problems by Young Canadian Researchers
Nouvelles approches des problèmes génétiques et génomiques par de jeunes chercheurs canadiens

Chair/Président: Lei Sun
Organizer/Responsable: Lei Sun

13:30-14:00 **Qihuang Zhang** (McGill University) **Qicheng Zhao** (McGill University)
Bayesian Model for Disease-Specific Gene Detection in High-Dimensional Spatially Resolved Transcriptomics / Modèle bayésien pour la détection de gènes spécifiques à la maladie dans la transcriptomique à haute résolution spatiale Ⓔ E

14:00-14:30 **Lin Zhang** (Simon Fraser University, Burnaby) **Lei Sun** (University of Toronto) **Andrew Paterson** (The Hospital for Sick Children)
Allele-frequency Estimation and Ancestry Informative Marker Identification via Retrospective Regression / Estimation de la fréquence allélique et identification de marqueur informatif sur l'ascendance à l'aide d'une régression rétrospective Ⓔ E

14:30-15:00 **Yongjin P. Park** (The University of British Columbia) **Sishir Subedi** (University of British Columbia) **Tomokazu Sumida** (Yale University)
Probabilistic Topic Modelling to Eavesdropping Cell-Cell Communication Patterns in Spatial Gene Expression Data / Modélisation de sujet probabiliste afin de reconnaître les tendances de communication entre cellules dans des données d'expression génique spatiale Ⓔ E

13:30-15:00 **Invited / Sur invitation** (abstract/résumé 240) **A 1046**

20th Anniversary of SSC Accreditation: Core Principles
Le 20è anniversaire de l'accréditation par la SSC : principes fondamentaux

Chair/Président: Hugh Chipman
Organizer/Responsable: Judy-Anne W. Chapman
Sponsor/Commanditaires: Accreditation Committee/Comité d'accréditation

13:30-14:00 **Peter D.M. Macdonald** (McMaster University)
University Course Requirements for Accreditation / Exigences en matière de cours universitaires pour l'accréditation Ⓔ E

14:00-14:30 **Tony Panzarella** (University of Toronto)
The Statistical Society of Canada's Code of Ethical Statistical Practice: An Effective Road Map to Promoting High Professional Standards / Le Code de déontologie statistique de la Société statistique du Canada : Une feuille de route efficace pour promouvoir des normes professionnelles élevées Ⓔ E

14:30-15:00 **Milena Kurtinecz** (Bayer Pharmaceuticals)
Accessible Variety of Professional Development / Formation professionnelle diversifiée et accessible Ⓔ E

13:30-15:00 **Invited / Sur invitation** (abstract/résumé 242) **A 1043**

2023 SSC Impact Award
Prix Impact de la SSC de 2023

Chair/Président: Jemila Seid Hamid
Organizer/Responsable: Jemila Seid Hamid

13:30-15:00 **Pierre R. L. Dutilleul** (McGill University)
"Spatial, temporal and multidimensional Statistics: Estimation, testing, and applications in the environmental sciences" – An overview with novelties / « Statistique spatiale, temporelle et multidimensionnelle : estimation, test et applications aux sciences de l'environnement » - une vue d'ensemble avec des nouveautés ⒺⒻ E F

13:30-15:00 **Contributed / Communications libres** (abstract/résumé 243) **C 4036**

Genetic Epidemiology and Statistics
Épidémiologie génétique et statistique

Chair/Président: Jinko Graham

13:30-13:45 **Hong Gu** (Dalhousie University) **Chaoyue Liu** (Dalhousie University) **Toby J. Kenney** (Dalhousie University) **Robert Beiko** (Dalhousie University) **Zesheng Jia** (Dalhousie University)
The Community Coevolution Model and Machine Learning Approach for Phylogenetic Comparative Analysis / Modèle de coévolution communautaire et approche d'apprentissage automatique pour l'analyse comparative phylogénétique Ⓔ E

13:45-14:00 **Jiaqi Bi** (University of Western Ontario) **Osvaldo Espin-Garcia** (University of Western Ontario) **Yun-Hee Choi** (University of Western Ontario)
Correlated Shared Frailty Model Incorporating Ascertainment Correction with Missing Covariates in Family-Based Studies / Modèle de fragilité partagée corrélée intégrant la correction de l'incertitude avec des covariables manquantes dans les études familiales

14:00-14:15 **Chenyang Li** (Western University) **Osvaldo Espin-Garcia** (Western University)
Optimizing Linear Polygenic Risk Score Combinations for Two-Phase Re-sequencing Study Design / Optimiser les combinaisons de scores de risques polygéniques linéaires pour une étude de reséquençage à deux phases

14:15-14:30 **Patrick McMillan** (University of Guelph) **Zeny Feng** (University of Guelph) **Lewis Lukens** (University of Guelph)
Improving Crop Variety Recommendations for Farmers: An Integrated Approach using Machine Learning and Genetics / Recommandations pour améliorer la variété des types de cultures agricoles : une approche intégrée utilisant l'apprentissage machine et la génétique

14:30-14:45 **Yuan Sun** (Lunenfeld-Tanenbaum Research Institute, Sinai Health, Toronto, Canada) **Laurent Briollais** (Lunenfeld-Tanenbaum Research Institute, Sinai Health, Toronto, Canada; Dalla Lana School of Public Health, University of Toronto, Toronto, Canada) **Xuming He** (Department of Statistics and Data Science, Washington University in St. Louis, St. Louis, USA)
A Two-Stage Model for Genome-Wide Association Study / Modèle à deux étapes pour études d'association à l'échelle du génome

14:45-15:00 **Brady Ryan** (University of Michigan) **Michael Boehnke** (University of Michigan) **Ryan Welch** (University of Michigan) **Christian Fuchsberger** (Eurac Research)
Using External Reference Panel and Single-Variant Summary Statistics for Rare-Variant Aggregation Tests / Utilisation d'un panel de référence externe et de statistiques récapitulatives à variante unique pour les tests d'agrégation de variants rares

---

13:30-15:00 **Contributed / Communications libres** (abstract/résumé 247) **C 3053**

Infectious Disease
Maladies infectieuses

Chair/Président: Lisa M. Lix

13:30-13:45 **Jeffrey W. Peitsch** (University of Calgary)
Directionally Dependent Individual Level Models for Infectious Disease / Modèles au niveau individuel avec dépendance directionnelle pour une maladie infectieuse

13:45-14:00 **Rado Malalatiana Ramasy** (Université de Montréal) **William Ruth** (Université de Montréal)
Multilevel Mediation Analysis : Deciphering the Impact of Information Sources on Adherence to Restrictive Measures during the COVID-19 Pandemic. / Analyse de médiation multiniveau : Décrypter l'impact des sources d'information sur l'adhésion aux mesures restrictives durant la pandémie de COVID-19

14:00-14:15 **Gyanendra Pokharel** (The University of Winnipeg)
Predictive Probability-based Gaussian Process Emulators for Infectious Disease Models / Émulateurs de processus gaussiens basés sur des probabilités prédictives pour les modèles de maladies infectieuses

14:15-14:30 **Cong Jiang** (University of Montreal) **Mireille Schnitzer** (University of Montreal) **Denis Talbot** (Laval University)
COVID-19 Vaccine Effectiveness Estimation Under the Test-Negative Design / Estimation de l'efficacité du vaccin contre la COVID-19 dans le cadre d'un devis test-négatif

14:30-14:45 **Jiaping (Olivia) Liu** (University of British Columbia) **Zhenglun Cai** (University of British Columbia) **Paul Gustafson** (University of British Columbia) **Daniel J. McDonald** (University of British Columbia)
RtEstim: Effective Reproduction Number Estimation With Trend Filtering / RtEstim : estimation du nombre effectif de reproduction avec filtrage des tendances

13:30-15:00 **Contributed / Communications libres** (abstract/résumé 250) **ED 2018B**

Probability Models
Modèles de probabilité

Chair/Président: Aya A. Mitani

13:30-13:45 **François A. Marshall** (Self employed) **Lenin Arango-Castillo** (Bank of Mexico)
Coherent Nationwide Variation in U.S. Air Pollution: A Novel Switching Model / Variation cohérente à l'échelle nationale de la qualité de l'air aux États-Unis : Un nouveau modèle à transfère

13:45-14:00 **Yunhong Lyu** (University of Montreal) **Bouchra Nasri** (University of Montreal) **Bruno N. Rémillard** (HEC Montréal)
Sequential Change-point Detecting with Generalized Ornstein–Uhlenbeck Processes / Détection séquentielle de points de changement à l'aide de processus d'Ornstein-Uhlenbeck généralisés

14:00-14:15 **Adam B. Kashlak** (University of Alberta)
Asymptotic Invariance in Randomization Tests / Invariance asymptotique dans les tests de randomisation

14:15-14:30 **Klaus Peter Herrmann** (Université de Sherbrooke) **Johanna G. Nešlehová** (McGill University) **Marius Hofert** (The University of Hong Kong)
Transformations of Stable Tail Dependence Functions / Transformations des fonctions de dépendance de queue stable

13:30-15:00 **Contributed / Communications libres** (abstract/résumé 252) **C 3033**

Actuarial Science 2
Science actuarielle 2

Chair/Président: Himchan Jeong

13:30-13:45 **Kathleen E. Miao** (University of Toronto) **Silvana Pesenti** (University of Toronto)
Robustifying Elicitable Functionals under Kullback-Leibler Misspecification / Amélioration de la robustesse des fonctions élicitables à l'aide de la divergence de Kullback-Leibler pour quantifier les erreurs de spécification

13:45-14:00 **Sébastien Jessup** (Concordia University) **Mélina Mailhot** (Concordia University) **Mathieu Pigeon** (Université du Québec à Montréal)
Robust Extreme Thresholds Through Generalised Bayesian Model Averaging / Seuils de valeurs extrêmes robustes en utilisant l'aggrégation de modèles généralisés bayésiens

14:00-14:15 **Xiyue Han** (University of Waterloo) **Alexander Schied** (University of Waterloo)
Statistical Aspects of Rough Stochastic Volatility / Aspects statistiques de la volatilité stochastique rugueuse

14:15-14:30 **Wei Liang** (University of Waterloo) **Changbao Wu** (University of Waterloo)
Model-Assisted Uplift Evaluation / Évaluation du levier assistée par un modèle

14:30-14:45 **Emma Kroell** (University of Toronto) **Silvana Pesenti** (University of Toronto) **Sebastian Jaimungal** (University of Toronto)
Optimal Reinsurance in a Monotone Mean-Variance Framework / Réassurance optimale dans le cadre de moyenne-variance monotone

14:45-15:00 **Taehan Bae** (University of Regina) **Tatjana Miljkovic** (Miami University - Oxford)
The Size-biased Lognormal Mixture with the Entropy Regularized Algorithm / Mélange lognormal biaisé par la taille avec l'algorithme régularisé par l'entropie

13:30-15:00 **Poster / Poster** (abstract/résumé 256) **CSF Whale Atrium**

Poster Presentations
Présentations par affichage

Organizer/Responsable: Tessema Astatkie

13:30-15:00 **Vihotogbé Edouard Houssou** (Polytechnique Montreal)
Probabilistic Spatial Interpolation of Meteorological Data : Exploitation of Spatial Patterns Provided by Regional Climate Models / Interpolation spatiale probabiliste de données météorologiques : exploitation des motifs spatiaux fournis par les modèles de climat régionaux

13:30-15:00 **Parham Pishrobat** (The University of British Columbia) **William Welch** (University of British Columbia) **Stefan Schrunner** (Norwegian University of Life Sciences)
Introducing Dynamic Kernel Regression for Enhancing Hydrological Inference / Présentation d'une régression dynamique à noyaux pour améliorer l'inférence hydrologique

13:30-15:00 **Andrew Putman** (Ontario Tech University) **Shilpa Dogra** (Ontario Tech University)
Initial Validity and Reliability Testing of the SGBA-5: A Measurement Tool for Facilitating Sex- And Gender-Based Analyses in Health Sciences Research / Tests initiaux de validité et de fiabilité du SGBA-5 : Un outil de mesure pour faciliter les analyses fondées sur le sexe et le genre dans la recherche en sciences de la santé

13:30-15:00 **Jiali Wang** (University of Manitoba) **Xikui Wang** (University of Manitoba)
Actuarial Study and Statistical Analysis of Wildfire Insurance Claims / Étude actuarielle et analyse statistique des réclamations d'assurance liées aux feux de forêt

13:30-15:00 **Roberto Primo Curti** (Thompson Rivers University) **Md. Erfanul Hoque** (Thompson Rivers University) **Sean Hellingman** (Thompson Rivers University)
Exploring the Complexity of Collectible Asset Valuation and Forecasting - Insights from Magic: The Gathering / Exploration de la complexité de l'évaluation et de la prévision des actifs de collection – Intuition grâce au jeu Magic : The Gathering

13:30-15:00 **Negar Kalanpour** (Memorial University of Newfoundland) **Armin Hatefi** (Memorial University of Newfoundland)
Shrinkage Estimators for Proportional Hazards Mixture Cure Models / Estimateurs avec rétrécissement pour des modèles de mélange pour guérison à risques proportionnel

15:30-17:00 **Invited / Sur invitation** (abstract/résumé 260) **A 1046**

Modelling Rainfall Extremes
Modélisation des précipitations extrêmes

Chair/Président: Michaël Lalancette
Organizer/Responsable: Léo Belzile

15:30-16:00 **Debbie J. Dupuis** (HEC Montréal) **Luca Trapin** (University of Bologna)
Mixed-frequency Extreme Value Regression: Estimating the Effect of Mesoscale Convective Systems on Extreme Rainfall Intensity / Régression à fréquences mixtes pour valeurs extrêmes : estimation de l'effet des complexes convectifs de méso-échelle sur l'intensité des précipitations extrêmes

16:00-16:30 **Mélina Mailhot** (Concordia University) **Mathieu Pigeon** (Université du Québec à Montréal) **Sébastien Jessup** (Concordia University)
Combination Methods on Extreme and Skewed Data / Méthodes de combinaison de données extrêmes et asymétriques

16:30-17:00 **Léo Belzile** (HEC Montréal) **Rishikesh Yadav** (HEC Montréal)
Can Climate Model Output Adequately Represent Extreme Rainfall? / Est-ce que les précipitations extrêmes de modèles climatiques sont fidèles à la réalité ?

15:30-17:00 **Invited / Sur invitation** (abstract/résumé 262) **C 2045**

The Bayesian Edge: Novel Applications of Bayesian Methods to Clinical Research, Indirect Treatment Comparison, and Public Health.
L'avantage bayésien : nouvelles applications des méthodes bayésiennes à la recherche clinique, à la comparaison indirecte des traitements et à la santé publique

Chair/Président: Audrey Béliveau
Organizer/Responsable: Aaron Springford

15:30-16:00 **Linke Li** (University of Toronto Dalla Lana School of Public Health)
Efficient Computation Methods for Expected Value of Sample Information in Bayesian Clinical Trial Designs / Méthodes de calcul efficaces de la valeur attendue de l'information d'échantillonnage (EVSI) dans des essais cliniques bayésiens

16:00-16:30 **Yiran Wang** (University of Waterloo) **Martin Lysy** (University of Waterloo) **Audrey Béliveau** (University of Waterloo)
Plant-Capture Methods for Estimating Population Size from Uncertain Plant Captures / Méthodes de capture de plantes pour estimer la taille de population provenant de captures incertaines de plantes

16:30-17:00 **Emma K Mackay** (Inka Health)
Bayesian borrowing approaches to address the challenges of evaluating efficacy/effectiveness in rare indications: applications to basket trials and pediatric studies / Approches d'emprunt bayésiennes pour relever les défis de l'évaluation de l'efficacité dans les indications rares : applications aux essais panier et études pédiatriques

15:30-17:00 **Invited / Sur invitation** (abstract/résumé 264) **ED 2018A**

Machine Learning in Causal Inference: Modern Health Research Paradigms
Apprentissage automatique en inférence causale : paradigmes modernes de la recherche en santé

Chair/Président: Mohammad Ehsanul Karim
Organizer/Responsable: Mohammad Ehsanul Karim

15:30-16:00 **Lan Wen** (University of Waterloo)
Estimating the Average Causal Effects of Dietary Substitution Strategies / Estimation des effets causaux moyens des stratégies de substitution alimentaire

16:00-16:30 **Robert W. Platt** (McGill University) **Rubiya Akter** (McGill University) **Enrico Ripamonti** (University of Milan-Bicocca)
Lookback Periods in Observational Epidemiology: Statistical Considerations / Périodes rétrospectives en épidémiologie par observation : considérations statistiques

16:30-17:00 **Mireille Schnitzer** (Université de Montréal) **Cong Jiang** (Université de Montréal) **Miceline Mésidor** (INRS-Institut Armand-Frappier) **Yan Liu** (Université de Montréal) **Edgar Ortiz Brizuela** (McGill University) **Mabel Carabali** (McGill University) **Denis Talbot** (Université Laval)
Methods for the test-negative design: application and analysis of vaccine effectiveness during the pandemic and new approaches / Méthodes pour le devis test négatif : l'application et l'analyse de l'efficacité de la vaccination pendant la pandémie et nouvelles approches

15:30-17:00 **Invited / Sur invitation** (abstract/résumé 266) **A 1045**

Recent Advances in Sequential Methods
Progrès récents en méthodes séquentielles

Chair/Président: Yanglei Song
Organizer/Responsable: Yanglei Song

15:30-16:00 **Yajun Mei** (University)
Active Learning in Sequential Analysis and Change-Point Detection / Apprentissage actif en analyse séquentielle et détection de changement de régime

16:00-16:30 **Jay Bartroff** (University of Texas at Austin)
Group Sequential Testing of a Treatment Effect Using a Surrogate Marker / Test séquentiel de groupe d'un effet de traitement à l'aide d'un marqueur de substitution

16:30-17:00 **Georgios Fellouris** (University of Illinois, Urbana-Champaign) **Yiming Xing** (University of Illinois, Urbana-Champaign)
Centralized and Asynchronous Sequential Multiple Testing / Essais séquentiels multiples centralisés et asynchrones

---

15:30-17:00 **Invited / Sur invitation** (abstract/résumé 268) **A 1043**

Inference Methods in Stochastic Processes with Change-points: Recent Advances
Méthodes d'inférence dans les processus stochastiques avec les points de changement : avancées récentes

Chair/Président: Sévérien Nkurunziza
Organizer/Responsable: Sévérien Nkurunziza
Sponsor/Commanditaires: Probability Section/Groupe de probabilité

15:30-16:00 **Zhou Zhou** (University of Toronto) **Weichi Wu** (Tsinghua University) **David Veitch** (University of Toronto)
Asynchronous Jump Testing and Estimation in High Dimensions Under Complex Temporal Dynamics / Test de saut asynchrone et estimation en grande dimension selon des dynamiques temporelles complexes

16:00-16:30 **Mai Ghannam** (University of Ottawa) **Sévérien Nkurunziza** (University of Windsor)
Estimation and Inference in a Tensor Regression Model with Change-Points / Estimation et inférence dans le modèle de régression tensoriel avec points de rupture

16:30-17:00 **Rogemar S. Mamon** (The University of Western Ontario) **Fuqi Chen** (Health Canada)
Determination of Multiple Change Points in a Multi Dimensional Mean-reverting Process / Détermination de points de changement multiples dans un processus de retour à la moyenne multidimensionnel

---

15:30-17:00 **Invited / Sur invitation** (abstract/résumé 270) **SN 2109**

Publishing for Early Career Researchers (Panel)
Publication pour les chercheurs en début de carrière (Table Ronde)

Chair/Président: James H. McVittie
Organizer/Responsable: Johanna G. Nešlehová
Sponsor/Commanditaires: Canadian Journal of Statistics/Revue canadienne de statistique

15:30-17:00 **Hugh Chipman** (Acadia University) **Josée Dupuis** (McGill University) **Richard A. Lockhart** (Simon Fraser University) **Grace Y. Yi** (University of Western Ontario)
Panel on Publishing for Early Career Researchers / Table ronde sur la publication pour les chercheurs en début de carrière

---

15:30-17:00 **Contributed / Communications libres** (abstract/résumé 271) **A 2071**

Distributions: Theory and Asymptotics
Distributions : théorie et asymptotique

Chair/Président: Zeinab Mashreghi

15:30-15:45 **Nahid Sadr** (Université de Sherbrooke) **Marius Hofert** (The University of Hong Kong) **Klaus Peter Herrmann** (Université de Sherbrooke)
Index-mixed Copulas: A New Class of Multivariate Copulas / Les copules mixtes d'indices : une nouvelle classe de copules multivariées Ⓔ E

15:45-16:00 **Nikola Surjanovic** (University of British Columbia) **Saifuddin Syed** (University of Oxford) **Alexandre Bouchard-Côté** (University of British Columbia) **Trevor Campbell** (University of British Columbia)
Exploration-agnostic Geometric Ergodicity of Parallel Tempering / Ergodicité géométrique à exploration-agnostique de l'atténuation parallèle Ⓔ E

16:00-16:15 **Samuel Valiquette** (Université de Sherbrooke) **Éric P. Marchand** (Université de Sherbrooke) **Gwladys Toulemonde** (Université de Montpellier) **Frédéric Mortier** (CIRAD) **Jean Peyhardi** (Université de Montpellier)
Multivariate Discrete Tree Pólya Splitting Distributions / Modèle multivarié discret Tree Pólya splitting Ⓔ E

16:15-16:30 **Evan Reynolds** (Carleton University) **Song Cai** (Carleton University)
Application of Lasso Methods to Parameter Estimation in Density Ratio Models / Application des méthodes Lasso à l'estimation des paramètres dans les modèles de rapport de densité Ⓔ E

16:30-16:45 **Christine Allard** (Université de Sherbrooke) **Éric P. Marchand** (Université de Sherbrooke)
Bayesian and Minimax Estimators of Loss / Estimateurs bayésiens et minimax de perte Ⓔ E

16:45-17:00 **Ian Waudby-Smith** (Carnegie Mellon University) **Martin Larsson** (Carnegie Mellon University) **Aaditya Ramdas** (Carnegie Mellon University)
Distribution-Uniform Strong Laws of Large Numbers / Uniformité de la loi forte des grands nombres sur des familles de distributions Ⓔ E

15:30-17:00 **Contributed / Communications libres** (abstract/résumé 275) **A 2065**

Developments in Statistical Theory and Bayesian methods
Développements en théorie statistique et méthodes bayésiennes

Chair/Président: Khurram Nadeem

15:30-15:45 **Dayi Li** (University of Toronto)
Bayesian Optimization Sequential Surrogate (BOSS) Algorithm: Fast Bayesian Inference for a Broad Class of Bayesian Hierarchical Models / Algorithme de substitution séquentiel d'optimisation bayésienne : inférence bayésienne rapide pour une vaste catégorie de modèles hiérarchiques bayésiens Ⓔ E

15:45-16:00 **Hui Shen** (McGill University) **Eric Kolaczyk** (McGill University)
Consistent Identification of Top-K Nodes in Noisy Networks / Identification consistante de liens « top-k » dans des réseaux bruités Ⓔ E

16:00-16:15 **Shenita Pramij** (Memorial University of Newfoundland) **Candemir Cigsar** (Memorial University of Newfoundland) **Yildiz Yilmaz** (Memorial University of Newfoundland)
Mediation Analysis for Recurrent Event Data / Analyse de médiation pour les données d'événements récurrents Ⓔ E

16:15-16:30 **Kevin Granville** (University of Windsor) **Douglas Woolford** (University of Western Ontario) **Charmaine B. Dean** (University of Waterloo)
Investigating changes in the timing of Ontario's wildland fire season: a spatial perspective / Étude des changements dans le temps de la saison des feux de forêt en Ontario : une perspective spatiale Ⓔ E

16:30-16:45 **Camila P. E. de Souza** (University of Western Ontario) **Pedro H. T. O. Souza** (Universidade Federal do Paraná) **Ronaldo Dias** (Universidade de Campinas)
Bayesian Variable Selection for Function-on-Scalar Regression Models: a Comparative Analysis / Sélection de variables bayésiennes pour des modèles de régression fonction-sur-scalaire : une analyse comparative Ⓔ E

16:45-17:00 **W. John Braun** (The University of British Columbia)
Monte Carlo Integration of a First Order Differential Equation / Intégration Monte Carlo d'une équation différentielle du premier ordre

---

15:30-17:00 **Contributed / Communications libres** (abstract/résumé 279) **C 4036**

Longitudinal and Time-series Data
Données longitudinales et séries chronologiques

Chair/Président: Qingrun Zhang

15:30-15:45 **Rose Garrett** (University of Toronto) **Eleanor M. Pullenayegum** (The Hospital for Sick Children)
Parametric modelling of irregular longitudinal data: a simulation study / Modélisation paramétrique de données longitudinales irrégulières : une étude en simulations

15:45-16:00 **George Stefan** (University of Toronto / The Hospital for Sick Children) **Eleanor M. Pullenayegum** (University of Toronto / The Hospital for Sick Children)
Methods for Irregularly Measured Longitudinal Data Subject to Informative Dropout / Méthodes pour des données longitudinales irrégulièrement mesurées assujetties à l'abandon informatif

16:00-16:15 **Marc Angelo Parsons** (McGill University) **Andrea Benedetti** (McGill University) **Russell Steele** (McGill University)
Comparing Fractional Polynomial and Spline Meta-Regression Models to Estimate Longitudinal Trajectories in the Presence of Heterogeneity in the Number and Timing of Assessments Between Studies / Une comparaison des modèles de meta-régression employant des bases polynomiales fractionnaires et splines pour l'estimation des trajectoires longitudinales dans la présence de l'hétérogénéité dans le calendrier de l'évaluation des mesures entre les études

16:15-16:30 **Kecheng Li** (University of Waterloo) **Richard J. Cook** (University of Waterloo)
Design and Sequential Analysis of Transfusion Trials / Conception et analyse séquentielle d'essais transfusionnels

16:30-16:45 **Hensley Hubert Mariathas** (Memorial University of Newfoundland) **Shabnam Asghari** (Memorial University of Newfoundland) **Oliver Hurley** (Memorial University of Newfoundland)
An Application of Interrupted Time Series Modeling using Autoregressive Integrated Moving Average for Evaluation of Quality Improvement Intervention / Application d'une modélisation de série chronologique interrompue à l'aide d'une moyenne mobile autorégressive intégrée (ARIMA) pour l'évaluation d'une intervention d'amélioration de la qualité

16:45-17:00 **Mathilde Dicaire-Cartier** (Université de Montréal) **Janie Coulombe** (Université de Montréal)
Estimating the Causal Effect of a Cumulative Exposure on a Continuous Outcome in Studies Prone to Confounding and Irregular Visits / Estimation de l'effet causal d'une exposition cumulative sur une réponse continue dans les études enclines à la confusion et aux visites irrégulières

---

15:30-17:00 **Contributed / Communications libres** (abstract/résumé 283) **C 3053**

New Developments in Statistical Theory and Analysis
Nouveaux développements en théorie et analyse statistiques

Chair/Président: Himchan Jeong

15:30-15:45 **Toby J. Kenney** (Dalhousie University) **Yun Cai Hong Gu** (Dalhousie University)
New Methods and Applications of Deconvolution / Nouvelles méthodes et applications de déconvolution

15:45-16:00 **Jervis Gallanosa** (University of Manitoba) **Yuliya V. Martsynyuk** (University of Manitoba)
On More Powerful Nonparametric Tests for Change in the Mean with Better Controlled Type I Errors / Tests non paramétriques plus puissants pour un changement de la moyenne avec erreurs de type 1 mieux contrôlées

16:00-16:15 **Ethan Lawler** (Dalhousie University) **Joanna Elizabeth Mills Flemming** (Dalhousie University) **Chris Field** (Dalhousie University)
Automatic Outlier Detection and Robust Filtering for Multivariate, Irregular, and Heteroscedastic State-Space Models / Détection automatique des valeurs aberrantes et filtrage robuste pour modèles espace-état multivariés, irréguliers et hétéroscédastiques

16:15-16:30 **Armin Hatefi** (Memorial University of Newfoundland) **Moein Yoosefi** (Memorial University of Newfoundland)
Shrinkage Methods for Contaminated Mixture Models with Matrix-valued Data / Méthodes de rétrécissement pour modèles de mélange contaminés avec des données matricielles

16:30-16:45 **Jia Wei He** **Ayesha Ali** (University of Guelph)
Proximal Projection for Doubly Sparse Regularized Models / Projection proximale pour les modèles régularisés à double parcimonie

16:45-17:00 **Kai Yang** (McGill University) **Masoud Asgharian** (McGill University) **Celia Greenwood** (McGill University)
Tsallis Entropy-Based Method for Sparse Statistical Machine Learning on Correlated Data and a Proximal Conjugate Gradient Algorithm for Nonconvex Nonsmooth Objective Function / Méthode basée sur l'entropie de Tsallis pour l'apprentissage machine statistique éparse de données corrélées et algorithme du gradient conjugué proximal pour une fonction objective non lisse et non convexe

---

15:30-17:00 **Contributed / Communications libres** (abstract/résumé 287) **ED 2018B**

Advances in Experimental Design and Inference
Progrès en conception et inférence expérimentales

Chair/Président: Hedayat Fathi

15:30-15:45 **Skye Paphora Griffith** (Queen's University) **Wesley Burr** (Trent University) **Glen Takahara** (Queen's University)
Boundary Correction and Smoothing Methods for the Spectrograms of Uniformly Modulated Processes / Méthodes de correction aux frontières et méthodes de lissage pour les spectrogrammes de processus uniformément modulés

15:45-16:00 **Bowei Ding** (University of Calgary) **Jingjing Wu** (Department of Mathematics and Statistics, University of Calgary) **Rohana J. Karunamuni** (Department of Mathematical and Statistical Sciences, University of Alberta)
Minimum Profile Hellinger Distance Estimation of Covariate Models / Estimation du profil de distance de Hellinger minimal pour modèles à covariables

16:00-16:15 **Jingyue Huang** (University of Waterloo) **Changbao Wu** (University of Waterloo) **Leilei Zeng** (University of Waterloo)
Empirical likelihood approaches to estimating quantile treatment effects / Approches de vraisemblance empirique pour l'estimation des effets de traitement par quantile

16:15-16:30 **Saba Saghatchi** (University of Calgary) **Xuewen Lu** (University of Calgary) **Jingjing Wu** (University of Calgary)
Variable Selection for Generalized Odds Rate Non-Mixture Cure Models with Current Status Data / Sélection de variable pour les modèles généralisés de non-mélange avec taux de guérison avec des données d'état actuel.

16:30-16:45 **Hao He** (University of Ottawa) **Hao He** (University of Ottawa) **David Haziza** (University of Ottawa) **Song Cai** (Carleton University)
Empirical Likelihood for Density Ratio Model with Missing Data / Vraisemblance empirique d'un modèle de ratio de densité avec données manquantes

16:45-17:00 **Louis Arsenault-Mahjoubi** (Simon Fraser University) **Jean-François Bégin** (Simon Fraser University)
A generalized Computational Method for Nonlinear Non-Gaussian Filtering in Finance / Une méthode computationnelle généralisée pour le filtrage nonlinéaire et non gaussien en finance

---

15:30-17:00 **Contributed / Communications libres** (abstract/résumé 291) **C 2033**

Causal Inference for Complex Data
Inférence causale pour les données complexes

Chair/Président: Anand N Vidyashankar

15:30-15:45 **Ana Carolina da Cruz** (University of Western Ontario) **Camila P. E. de Souza** (University of Western Ontario)
Variational Bayes for Basis Function Selection for Functional Data Representation with Correlated Errors / Un algorithme variationnel bayésien pour la sélection des fonctions de base pour la représentation de données fonctionnelles avec des erreurs corrélées

15:45-16:00 **Mélanie Raymond** (Université du Québec à Montréal)
Building Ancestral Recombination Graphs with Reinforcement Learning / Construire des Généalogies de population en utilisant l'apprentissage par renforcement

16:00-16:15 **Ashani N. Wickramasinghe** (University of Manitoba) **Saman Muthukumarana** (University of Manitoba) **Matt Schaubroeck** (ioAirFlow)
Hotspot Analysis in Buildings using Moran's I Statistic / Analyse des points chauds dans les immeubles avec la statistique I de Moran

16:15-16:30 **Jasper Zhongyuan Zhang** (University of Toronto) **Rafal Kustra** (University of Toronto) **Davide Chicco** (University of Toronto and Università di Milano-Bicocca)
Identifying Clinically Relevant Clusters within Cognitive State Research among a Large Adult Population / Identification de groupes cliniquement pertinents dans la recherche sur l'état cognitif au sein d'une large population adulte

16:30-16:45 **Alex Stringer** (University of Waterloo) **Jeffrey Negrea** (University of Waterloo)
Testing Variance Components the Easy Way / Tester les composantes de la variance en toute simplicité

16:45-17:00 **Kelly Ramsay** (York University) **Shojaeddin Chenouri** (University of Waterloo)
Changepoint Detection in the Variability of Multivariate and Functional Data / Détection de points de changement dans la variabilité des données fonctionnelles et multivariées

15:30-17:00 **Contributed / Communications libres** (abstract/résumé 295) **C 3033**

Methods for High-Dimensional and Large Data
Méthodes pour données de grande dimension et de grande taille

Chair/Président: Tharshanna Nadarajah

15:30-15:45 **Jesse Ghashti** (University of British Columbia - Okanagan) **Jeffrey Andrews** (University of British Columbia) **John R.J. Thompson** (University of British Columbia)
A Bootstrap Augmented k-means Algorithm for Fuzzy Partitions / Algorithme à K moyennes augmenté pour les partitionnements de données diffus

15:45-16:00 **Ladan Tazik** (University of British Columbia) **W. John Braun** (University of British Columbia)
Local Polynomial Lp Norm Regression / Régression polynomiale locale de la norme Lp

16:00-16:15 **Mohammad Kaviul Anam Khan** (University of Toronto) **Rafal Kustra** (Dalla Lana School of Public Health, University of Toronto) **Olli Saarela** (Dalla Lana School of Public Health, University of Toronto)
Conditional Permutation based on Generalized Variable Importance Metric and it's Relation to Causal Inference / Permutation conditionnelle basée sur l'importance de variable généralisée et son lien à l'inférence causale

16:15-16:30 **Thimani Dananjana Ranathungage** (University of Manitoba) **Sulalitha Bowala** (University of Manitoba) **Md. Erfanul Hoque** (Thompson Rivers University) **Aerambamoorthy Thavaneswaran** (University of Manitoba) **Ruppa Thulasiram** (University of Manitoba)
Application of a Novel Fuzzy Pattern Mining Algorithm for Sequence Data / Application d'un nouvel algorithme d'exploration de modèles flous pour des données de séquences

16:30-16:45 **Sarah Organ** (Dalhousie University) **Hong Gu** (Dalhousie University) **Toby J. Kenney** (Dalhousie University)
Vertex Cover Matroid Variable Selection for Controlling the False Discovery Rate and Improving Power With Correlated Predictors / Sélection de variables matroïdes de couverture par sommets pour contrôler le taux de fausses découvertes et améliorer la puissance avec des prédicteurs corrélés

16:45-17:00 **Tia Der** (The University of British Columbia) **John R.J. Thompson** (The University of British Columbia)
Iterative Mean-Shift Clustering for Change-Point Regression Estimation / Regroupement itératif par déplacement de la moyenne pour l'estimation de la régression des points de changement

---

15:30-17:00 **Invited / Sur invitation** (abstract/résumé 299) **A 1049**

NSERC Discovery Grants Information Session
Séance d'information sur les subventions à la découverte de CRSNG

Chair/Président: Saman Muthukumarana
Organizer/Responsable: Saman Muthukumarana
Sponsor/Commanditaires: Research Committee/Comité de la recherche

15:30-17:00 **Adele Ngi-Song** (NSERC)
NSERC Discovery Grants Information Session / Séance d'information sur les subventions à la découverte de CRSNG

17:00-18:00 **Invited / Sur invitation** (abstract/résumé 300) **SN 2109**

LaTeX class presentation (Canadian Journal of Statistics)
Présentation de la classe LaTeX (La revue canadienne de statistique)

Chair/Président: Johanna G. Nešlehová
Organizer/Responsable: Johanna G. Nešlehová
Sponsor/Commanditaires: Canadian Journal of Statistics/La revue canadienne de statistique

17:00-18:00 **Vincent Goulet** (Université Laval)
Introducing the new class for authors of The Canadian Journal of Statistics: cjs-rcs-article / Introduction de la nouvelle classe pour les auteurs de La revue canadienne de statistique : cjs-rcs-article

Wednesday June 5 — mercredi 5 juin

08:30-09:30 — **Invited / Sur invitation** (abstract/résumé 301) — **IIC 2001**

SSC 2023 Gold Medal Address
Allocution du récipiendaire de la Médaille d'or de la SSC 2023

Chair/Président: Grace Y. Yi
Organizer/Responsable: Grace Y. Yi

08:30-09:30 **Charmaine B. Dean** (University of Waterloo)
Optimizing research impact through interdisciplinary and collaborative research / Optimiser l'impact de la recherche par la recherche interdisciplinaire et collaborative

---

09:30-09:50 — **Invited / Sur invitation** (abstract/résumé 302) — **IIC 2001**

Presenting of Student Research Presentation and Case Study Awards
Remise des prix pour les présentations de recherche étudiantes et d'études de cas

Chair/Président: Shirley E. Mills
Organizer/Responsable: Shirley E. Mills

09:30-09:50 **Pingzhao Hu** (Western University)
Student Research Presentation Awards / Prix pour les présentations de recherche étudiantes

09:30-09:50 **Chel Hee Lee** (Alberta Health Services)
Case Studies in Data Analysis Awards / Prix d'études de cas en analyse de données

---

10:20-11:50 — **Invited / Sur invitation** (abstract/résumé 303) — **C 2045**

Statistics in Biosciences (SIBS): Real World Challenges and Recent Methodological Developments
Statistique en biosciences : défis du monde réel et développements méthodologiques récents

Chair/Président: Joan X. Hu
Organizer/Responsable: Joan X. Hu

10:20-10:42 **Naisyin Wang** (University of Michigan)
Utilizing Synthetic Components to Balance Privacy Protection and Data Utility / Utilisation de composants synthétiques pour équilibrer la protection de la vie privée et l'utilité des données

10:42-11:05 **Jie Chen** (Augusta University)
Linking Genomic Features to the Survival Time of GBM Cancer Patients / Lier les caractéristiques génomiques au temps de survie de patients atteints de cancer GBM

11:05-11:27 **Subharup Guha** (University of Florida) **Yi Li** (University of Michigan)
Causal Meta-Analysis by Integrating Multiple Observational Studies with Multivariate Outcomes / Méta-analyse causale par l'intégration de plusieurs études d'observation avec des résultats multivariés

11:27-11:50 **Rui Wang** (Harvard Pilgrim Health Care Institute) **Chia-Rui Chang** (Harvard University) **Yue Song** (Harvard University) **Fan Li** (Duke University)
Covariate Adjustment in Randomized Clinical Trials with Missing Covariate and Outcome Data / Ajustement de covariables dans des essais cliniques randomisés avec covariables et données de résultats manquantes

---

10:20-11:50 — **Invited / Sur invitation** (abstract/résumé 306) — **A 1045**

Novel Spatiotemporal Models for Complex Data in Fisheries and Ecosystem Studies
Nouveaux modèles spatio-temporels pour les données complexes des études sur les pêcheries et écosystèmes

Chair/Président: Noel Cadigan
Organizer/Responsable: Noel Cadigan, Nan Zheng, Asokan Mulayath Variyath

10:20-10:50 **Raphael Robert McDonald** (Dalhousie University) **David Keith** (Fisheries and Oceans Canada) **Jessica Sameoto** (Fisheries and Oceans Canada) **Joanna Elizabeth Mills Flemming** (Dalhousie University)
Improving Spatio-Temporal Stock Assessment Models Through the Inclusion of Habitat Features and Drop-Camera Surveys / Améliorer les modèles spatio-temporels d'évaluation des stocks en incluant les caractéristiques de l'habitat et les sondages par caméras lestées Ⓔ E

10:50-11:20 **James Thorson** (Alaska Fisheries Science Center)
Including Ecological Mechanism in Spatio-temporal Analysis: Habitat Preferences and Structural Multivariate Spatio-temporal Models / Intégration de mécanismes écologiques dans l'analyse spatio-temporelle : préférences en matière d'habitat et modèles structurels spatio-temporels multivariés Ⓔ E

11:20-11:50 **Nan Zheng** (Memorial University of Newfoundland) **Noel Cadigan** (Fisheries and Marine Institute of Memorial University of Newfoundland)
Enhancing Fisheries Stock Assessment: Spatiotemporal Modeling of Zero-Inflated Nonnegative Continuous Data using the Tweedie Distribution / Modélisation spatio-temporelle de données continues non négatives et avec excès de zéros à l'aide de la distribution Tweedie pour améliorer l'évaluation des stocks de pêche Ⓔ E

10:20-11:50 **Invited / Sur invitation** (abstract/résumé 308) **A 1049**

Modelling Dependence of Multivariate Extremes
Modélisation de la dépendance des extrêmes multivariés

Chair/Président: Léo Belzile
Organizer/Responsable: Debbie J. Dupuis

10:20-10:50 **Natalia Nolde** (University of British Columbia) **Vincenzo Coia** (BGC Engineering) **Harry Joe** (University of British Columbia)
Copula-Based Conditional Tail Indices / Indices de queue conditionnels basés sur des copules Ⓔ E

10:50-11:20 **Stanislav Volgushev** (University of Toronto) **Michaël Lalancette** (Université du Québec à Montréal) **Alexander Ryabchenko** (University of Toronto) **Sebastian Engelke** (University of Geneva)
Learning Hüsler-Reiss Graphical Models Under Connectedness Constraints / Apprentissage de modèles graphiques de Hüsler-Reiss sous contraintes de connexité Ⓔ E

11:20-11:50 **Michaël Lalancette** (Université du Québec à Montréal)
On Pairwise Interaction Multivariate Pareto Models and Score Matching / Sur les modèles de Pareto multivariés à interaction de deuxième ordre et le "Score Matching" Ⓔ E

10:20-11:50 **Invited / Sur invitation** (abstract/résumé 310) **A 2071**

Multi-state Modeling for the Analysis of Lifetime Data
Modélisation multi-états pour l'analyse des données de durée de vie

Chair/Président: Yildiz Yilmaz
Organizer/Responsable: Yildiz Yilmaz

10:20-10:50 **Yingwei (Paul) Peng** (Queen's University)
An Additive Hazards Frailty Model with Semi-varying Coefficients / Un modèle de fragilité à risques additifs avec des coefficients semi-variables Ⓔ E

10:50-11:20 **Leilei Zeng** (University of Waterloo) **Yidan Shi** (University of Pennsylvania)
A Mixture Hidden Markov Model for Multiple Types of Disease / Modèle de Markov caché à mélange pour plusieurs types de maladies Ⓔ E

11:20-11:50 **Candemir Cigsar** (Memorial University of Newfoundland) **Leila Torabi** (Memorial University of Newfoundland) **Zhaozhi Fan** (Memorial University of Newfoundland)
Quantile Regression for Sequentially Observed Bivariate Survival Data / Régression quantile (QR) pour des données de survie bivariées séquentiellement observées Ⓔ E

10:20-11:50 **Invited / Sur invitation** (abstract/résumé 312) **A 2065**

Statistical Methods for Animal Studies
Méthodes statistiques des études animales

Chair/Président: Elif Fidan Acar
Organizer/Responsable: Elif Fidan Acar

10:20-10:50 **Saman Muthukumarana** (University of Manitoba)
Modelling Fish Movements and Determining Vulnerability to Fishing Effort in Lake Winnipeg Using Bayesian State-space Models / Modélisation des déplacements de poissons et détermination de la vulnérabilité de l'effort de pêche au lac Winnipeg au moyen de modèles spatiotemporels bayésiens

10:50-11:20 **Théo Michelot** (Dalhousie University)
Multiscale Models of Animal Movement With Irreversible Dynamics / Modèles de déplacements d'animaux avec dynamiques irréversibles

11:20-11:50 **Alysha Cooper** (University of Guelph) **Ayesha Ali** (University of Guelph) **Zeny Feng** (University of Guelph)
Sparse Regression Modeling for Compositional Data: Regularized Dirichlet-Multinomial Regression via Dominating Hyperplane Regularization / Modélisation de la régression parcimonieuse pour les données compositionnelles : régression multinomiale de Dirichlet régularisée par régularisation de l'hyperplan principal

10:20-11:50 **Invited / Sur invitation** (abstract/résumé 314) **ED 2018B**

Recent Developments in Statistical Network Analysis
Développements récents en analyse statistique des réseaux

Chair/Président: Owen G Ward
Organizer/Responsable: Owen G Ward

10:20-10:50 **Neil A. Spencer** (University of Connecticut)
Robust Bayesian Model Selection for Network Data / Sélection d'un modèle bayésien robuste pour les données de réseaux

10:50-11:20 **Jie Jian** (University of Waterloo)
Restricted Tweedie Stochastic Block Models / Modèles de blocs stochastiques Tweedie restreints

11:20-11:50 **Peter W. MacDonald** (McGill University) **Eric Kolaczyk** (McGill University)
Summaries of Markov Models for Evolving Networks - Statistical Properties / Résumés de modèles de Markov pour réseaux évolutifs – propriétés statistiques

10:20-11:50 **Invited / Sur invitation** (abstract/résumé 316) **A 1046**

New Advancements in Formal Privacy Methods and Synthetic Data Generation
Nouvelles avancées en méthodes formelles de protection de la vie privée et génération de données synthétiques

Chair/Président: Bei Jiang
Organizer/Responsable: Bei Jiang, Éric Gagnon
Sponsor/Commanditaires: Survey Methods Section/Groupe des méthodes d'enquête

10:20-10:50 **Weijie Su** (University of Pennsylvania)
Enhancing Privacy Guarantees for the Census Data via Gaussian Differential Privacy / Amélioration des garanties de confidentialité pour les données de recensement grâce à la confidentialité différentielle gaussienne

10:50-11:20 **Jingchen (Monika) Hu** (Vassar College) **Terrance Savitsky** (U.S. Bureau of Labor Statistics) **Matthew Williams** (RTI International)
Mechanisms for Global Differential Privacy under Bayesian Data Synthesis / Mécanismes de confidentialité différentielle globale dans le cadre d'une synthèse bayésienne des données Ⓔ E

11:20-11:50 **Héloïse Gauvin** (Statistique Canada)
Creating a Synthetic version of a Longitudinal and Structured file: challenges and lessons learned / Créer une version synthétique d'un fichier longitudinal et structuré : les défis et les leçons apprises Ⓕ F

10:20-11:50 **Invited / Sur invitation** (abstract/résumé 318) **A 1043**

Spotlight on CANSSI postdocs
Vitrine des étudiants postdoctoraux de l'INCASS

Chair/Président: Andrea Benedetti
Organizer/Responsable: Andrea Benedetti
Sponsor/Commanditaires: CANSSI/INCASS

10:20-10:50 **William Ruth** (University of Montreal) **Rado Malalatiana Ramasy** (University of Montreal) **Rowin Alfaro** (University of Montreal) **Ariel Mundo** (University of Montreal) **Bouchra Nasri** (University of Montreal)
Statistical Considerations in Causal Mediation Analysis / Considérations statistiques dans l'analyse de la médiation causale Ⓔ E

10:50-11:20 **Chi-Kuang Yeh** (University of Waterloo) **Peijun Sang** (University of Waterloo) **Qihuang Zhang** (McGill University) **Archer Yi Yang** (McGill University) **Celia Greenwood** (McGill University)
Multivariate Spatial Functional Principal Component Analysis / Analyse en composantes principales multivariée pour données fonctionnelles spatiales Ⓔ E

10:20-11:50 **Invited / Sur invitation** (abstract/résumé 320) **IIC 2001**

CJS Award Address
Allocution du recipiendaire du Prix de la RCS

Chair/Président: Andrei Volodin
Organizer/Responsable: Andrei Volodin
Sponsor/Commanditaires: The Canadian Journal of Statistics Award Committee/Comité du prix de La revue canadienne de statistique

10:20-11:50 **Cong Jiang** (University of Montreal) **Michael Wallace** (University of Waterloo) **Mary Thompson** (University of Waterloo)
Dynamic Treatment Regimes and Interference / Régimes de traitement dynamiques et interférences Ⓔ E

10:20-11:50 **Invited / Sur invitation** (abstract/résumé 321) **ED 2018A**

Advances in Statistical Models for Single Cell RNA-seq Data
Progrès en modèles statistiques pour les données d'ARN-seq de cellules uniques

Chair/Président: Qihuang Zhang
Organizer/Responsable: Qihuang Zhang

10:20-10:50 **Qingrun Zhang** (University of Calgary) **Sandesh Acharya** (University of Calgary) **Jiami Guo** (University of Calgary)
Stabilized Marker Gene and Pathway Identification in Single-Cell RNA-seq Data / Gène marqueur stabilisé et identification de voie dans des données de séquençage de l'ARN à cellule unique Ⓔ E

10:50-11:20 **Pingzhao Hu** (Western University)
Spatial Transcriptomic Profile Prediction from Histology Images using Novel Contrastive Learning / Prédiction du profil transcriptomique spatial à partir d'images histologiques à l'aide d'un nouvel apprentissage contrastif

11:20-11:50 **Xuekui Zhang** (University of Victoria) **Li Xing** (University of Saskatchewan)
Does increasing sample size inflate false positive rates? / L'augmentation de la taille de l'échantillon gonfle-t-elle les taux de faux positifs ?

10:20-11:50 **Contributed / Communications libres** (abstract/résumé 323) **C 4036**

Ecology and Evolution
Écologie et évolution

Chair/Président: Hina Shaheen

10:20-10:35 **Jacqueline A. May** (University of Waterloo) **Zeny Feng** (University of Guelph) **Sarah J. Adamowicz** (University of Guelph)
Approaches for Handling Missing Values and Their Impacts on Statistical Inferences: A Molecular Rate Case Study / Approches de traitement des valeurs manquantes et de leur impact sur les inférences statistiques : étude de cas sur les taux moléculaires

10:35-10:50 **Yuan Bian** (University of Western Ontario) **Grace Y. Yi** (University of Western Ontario) **Wenqing He** (University of Western Ontario)
Boosting Learning in the Presence of Incomplete Data / Apprentissage par boosting en présence de données incomplètes

10:50-11:05 **Gracia Y. Dong** (University of Toronto/University of Victoria) **Jennifer McNichol** (Simon Fraser University) **Laura L.E. Cowen** (University of Victoria)
Population Size Estimation in a Two-Sample Study using Capture-Recapture Techniques / Estimation de la taille de la population dans une étude à deux échantillons à l'aide de techniques de capture-recapture

11:05-11:20 **Jonathan Babyn** (Dalhousie University)
Evaluating the Feasibility of Juvenile Only CKMR on Grey Seals / Évaluation de la faisabilité de l'application de la méthode de capture-marquage-recapture génétique sur des juvéniles seulement pour les phoques gris

11:20-11:35 **Hoang Nguyen** (Fisheries and Marine Institute of Memorial University of Newfoundland)
Accounting for Movement in Spatial Surplus Production Models: A Case Study on 3LN Redfish / Prise en compte du déplacement dans les modèles spatiaux de production excédentaire : étude de cas du sébaste 3LN

10:20-11:50 **Contributed / Communications libres** (abstract/résumé 326) **C 3053**

Environmental Stress Modelling and Prediction
Modélisation et prévision des stress environnementaux

Chair/Président: Kevin Granville

10:20-10:35 **Orla A. Murphy** (Dalhousie University) **Jonathan Jalbert** (Polytechnique Montréal)
Predicting Extreme Rainfall in Nova Scotia Using a Spatial Bayesian Hierarchical Model / Prévision des précipitations extrêmes en Nouvelle-Écosse à l'aide d'un modèle hiérarchique bayésien spatial

10:35-10:50 **Henrik Stryhn** (University of Prince Edward Island)
Design and Analysis for Ranking of Machine-Rated Applications to a Professional Program / Conception et analyse pour le classement d'applications à un programme de formation professionnelle évaluées par ordinateur

10:50-11:05 **Syeda Fateha Akter** (Memorial University of Newfoundland)
Parameter Estimation of Poisson Autoregressive Moving Average Model / Estimation de paramètres d'un modèle de moyennes mobiles autorégressif de Poisson

11:05-11:20 **Archer Gong Zhang** (University of Toronto) **Nancy Reid** (University of Toronto) **Qiang Sun** (University of Toronto)
A Semiparametric Approach to Data-Integrated Causal Inference / Approche semi-paramétrique pour une inférence causale avec données intégrées

11:20-11:35 **Zixuan Yang** (The University of Western Ontario) **Douglas Woolford** (Statistical & Actuarial Sciences, University of Western Ontario)
Predict the Wildfire Occurrence in Ontario, Canada: An Errors in Variable Modelling Approach / Approche de modélisation des erreurs dans les variables pour prévoir la fréquence des feux de forêt en Ontario, au Canada

11:35-11:50 **Matthias Schonlau** (Department of Statistics and Actuarial Science, University of Waterloo)
Hammock Plots: Visualizing Categorical and Numerical Variables / Graphiques de hamac : visualisation des variables catégorielles et numériques

---

13:30-15:00 **Invited / Sur invitation** (abstract/résumé 330) **A 1046**

Recent Advances in Event History Analysis
Recent Advances in Event History Analysis

Chair/Président: Candemir Cigsar
Organizer/Responsable: Candemir Cigsar

13:30-14:00 **Jerald F. Lawless** (University of Waterloo)
Assessing Interventions in Observational Event History Studies / Évaluation des interventions dans les études observationnelles de l'historique des événements

14:00-14:30 **Joan X. Hu** (Simon Fraser University) **Ken Peng** (Simon Fraser University) **Tim B. Swartz** (Simon Fraser University)
An Extended Hawkes Process Model for Recurrent Events / Modèle de processus de Hawkes étendu pour les événements récurrents

14:30-15:00 **Yi Xiong** (State University of New York at Buffalo) **Gary Chan** (University of Washington) **Malka Gorfine** (Tel Aviv University) **Li Hsu** (Fred Hutchinson Cancer Center)
Causal Inference in Cost-Effectiveness Analysis with Semi-competing Risks Data / Inférence causale pour les analyses de coût-efficacité avec des données de risques semi-concurrents

---

13:30-15:00 **Invited / Sur invitation** (abstract/résumé 333) **A 1045**

Navigating Academic Sabbatical (Panel)
Gérer son congé sabbatique (Table ronde)

Chair/Président: Kuan Liu
Organizer/Responsable: Kevin McGregor
Sponsor/Commanditaires: Committee on New Investigators/Comité des nouveaux chercheurs

13:30-15:00 **Bei Jiang** (University of Alberta) **Olli Saarela** (University of Toronto) **Paul Gustafson** (University of British Columbia) **Matthias Schonlau** (University of Waterloo)
Navigating Academic Sabbatical / Gérer son congé sabbatique

13:30-15:00 **Invited / Sur invitation** (abstract/résumé 334) **A 1043**

Bridging the Gap with Large Language Models (Panel)
Combler le fossé avec de grands modèles linguistiques (table ronde)

Chair/Président: Vahid Partovi Nia
Organizer/Responsable: Vahid Partovi Nia
Sponsor/Commanditaires: Business and Industrial Statistics Section/Groupe de statistique industrielle et de gestion

13:30-15:00 **Alejandro Murua** (Université de Montréal) **Pascal Poupart** (Vector Institute, University of Waterloo) **Pierre-Jérôme Bergeron** (Google) **Martin Lysy** (University of Waterloo)
Bridging the Gap with Large Language Models / Combler le fossé avec de grands modèles de langage

---

13:30-15:00 **Invited / Sur invitation** (abstract/résumé 335) **ED 2018B**

Q & A with the Director and Deputy Director of CANSSI
Questions-réponses avec le directeur et la directrice adjointe de l'INCASS

Chair/Président: Donald Estep
Organizer/Responsable: Donald Estep
Sponsor/Commanditaires: CANSSI/INCASS

13:30-15:00 **Donald Estep** (Simon Fraser University/CANSSI) **Andrea Benedetti** (McGill University)
CANSSI Programs and Plans / Programmes et plans de l'INCASS

13:30-15:00 **Invited / Sur invitation** (abstract/résumé 336) **IIC 2001**

Pierre Robillard Invited Address
Allocution du recipiendaire du Prix Pierre-Robillard

Chair/Président: Christian Léger
Organizer/Responsable: Christian Léger

13:30-15:00 **Qiuqi Wang** (Georgia State University)
Standard and comparative e-backtest based on elicitability / Backtests-e standard et comparatifs basés sur l'elicitabilité

13:30-15:00 **Invited / Sur invitation** (abstract/résumé 337) **ED 2018A**

Advancing Precision Medicine through Innovative Statistical Methods
Faire progresser la médecine de précision par des méthodes statistiques innovantes

Chair/Président: Cong Jiang
Organizer/Responsable: Cong Jiang

13:30-14:00 **Hengrui Cai** (University of California, Irvine)
Doubly Robust Interval Estimation for Optimal Policy Evaluation in Online Learning / Estimation d'intervalles doublement robuste pour l'évaluation optimale des politiques dans l'apprentissage en ligne

14:00-14:30 **Dylan Spicker** (University of New Brunswick)
Infinite and Irregular: Developments for Dynamic Treatment Regimes with Stochastic Decision Points / Infini et irrégulier : Développements pour les régimes de traitement dynamiques avec des points de décision stochastiques

---

13:30-15:00 **Invited / Sur invitation** (abstract/résumé 339) **C 2045**

Restoring Survey Response Rates and Mitigating the Nonresponse Error: Current Approaches and Recent Findings
Rétablir les taux de réponse aux enquêtes et atténuer l'erreur de non-réponse : approches actuelles et résultats récents

Chair/Président: Peter G. Wright
Organizer/Responsable: Peter G. Wright
Sponsor/Commanditaires: Survey Methods Section/Groupe des méthodes d'enquête

13:30-14:00 **Emma Troughton** (Statistics Canada) **Peter Wright** (Statistics Canada)
Recent Initiatives to Assess the Potential Nonresponse Error in Social Surveys / Initiatives récentes pour évaluer l'erreur potentielle due à la non-réponse des enquêtes sociales Ⓔ E

14:00-14:30 **France Lapointe** (Institut de la Statistique du Québec) **Éric Gagnon** (Institut de la Statistique du Québec)
Are Satistical Surveys of Individuals (Once Again) at a Crossroads? / Les enquêtes statistiques auprès des individus (encore) à la croisée des chemins ? Ⓕ E F

14:30-15:00 **Hélène Chaput** (Insee (France))
Mixed-mode collection of household surveys in France: difficulties and opportunities / Mixed-mode collection of household surveys in France : difficulties and opportunities Ⓕ E

13:30-15:00 **Invited / Sur invitation** (abstract/résumé 341) **A 2071**

Convergence of MCMC Algorithms
Convergence des algorithmes MCMC

Chair/Président: Jeffrey Negrea
Organizer/Responsable: Jeffrey S. Rosenthal

13:30-14:00 **Trevor Campbell** (The University of British Columbia) **Nikola Surjanovic** (University of British Columbia) **Saifuddin Syed** (Oxford University) **Alexandre Bouchard-Côté** (University of British Columbia)
An Exploration-agnostic Characterization of the Ergodicity of Parallel Tempering / Une caractérisation exploration-agnostique de l'ergodicité de l'atténuation parallèle Ⓔ E

14:00-14:30 **Gareth O. Roberts** (University of Warwick) **Jeffrey S. Rosenthal** (University of Toronto) **Nick Tawn** (University of Warwick)
Parallel Tempering Schemes and Robustness to Dimensionality / Schémas d'atténuation parallèle et robustesse à la dimensionalité Ⓔ E

14:30-15:00 **Jeffrey S. Rosenthal** (University of Toronto)
Experiments with MCMC Tempering Options / Expérimentations avec options de tempérage par chaîne de Markov Monte-Carlo (MCMC) Ⓔ E

13:30-15:00 **Contributed / Communications libres** (abstract/résumé 343) **A 2065**

Prediction Using Different Models
Prédiction à l'aide de différents modèles

Chair/Président: Julie Carreau

13:30-13:45 **Michael Guerzhoy** (University of Toronto) **Max Piasevoli** (Princeton University, Microsoft Corporation) **Tracy Qian** (University of Toronto)
Automatic Model Selection using Wasserstein Generative Adversarial Networks / Sélection automatique de modèles à l'aide de réseaux antagonistes génératifs de Wasserstein Ⓔ E

13:45-14:00 **Funmilola Mary Taiwo** (University of Manitoba)
Bayesian multiclass approach for predicting student dropout / Approche bayésienne multiclasse pour prédire le décrochage scolaire Ⓔ E

14:00-14:15 **Yuxuan Zhao** (University of Waterloo)
Inference for time-delay differential equations / Inférence pour les équations différentielles à retard

14:15-14:30 **Megan French** (Department of Statistics and Actuarial Science, University of Waterloo) **Ryan Browne** (University of Waterloo)
Block Diagonal Gaussian Mixture Models / Modèles de mélange gaussien en blocs diagonaux

14:30-14:45 **Robert Zimmerman** (University of Toronto) **David A. van Dyk** (Imperial College London) **Vinay L. Kashyap** (Harvard & Smithsonian) **Aneta Siemiginowska** (Harvard & Smithsonian)
Separating Flaring and Quiescent States in Active Coronae using State-Space Models / Séparation des états en éruption et en quiescence des couronnes actives à l'aide de modèles d'espace d'états

14:45-15:00 **Owen G. Ward** (Simon Fraser University)
Statistical Network Analysis with Aggregated Relational Data / Analyse statistique de réseaux avec des données relationnelles agrégées

---

13:30-15:00 **Contributed / Communications libres** (abstract/résumé 347) **C 2033**

Statistical Models for Clinical and Healthcare Data
Modèles statistiques pour les données cliniques et de santé

Chair/Président: Marie-Pierre Sylvestre

13:30-13:45 **Kehinde I. Olobatuyi** (University of Victoria) **Laura L.E. Cowen** (University of Victoria) **Patrick Brown** (University of Toronto) **Matthew Parker** (Simon Fraser University)
Multi-state Dynamic Capture-Recapture model for Big Data: Estimating undetected COVID-19 Cases in British Columbia, Canada / Modèle dynamique de capture et de recapture multi-états pour données volumineuses : estimation des cas de COVID-19 non détectés en Colombie-Britannique, Canada

13:45-14:00 **Razvan G. Romanescu** (University of Manitoba)
Epidemic Spread Dynamics on Associative Networks / Dynamique de propagation des épidémies sur les réseaux associatifs

14:00-14:15 **Jianchu Chen** (University of Waterloo) **Richard J. Cook** (University of Waterloo)
Cost-effective design of Survival Studies Involving Intermittent Observation of Time-Dependent Covariates / Conception économique d'études de survie comportant une observation intermittente de covariables dépendantes du temps

14:15-14:30 **Renny Doig** (Simon Fraser University) **Liangliang Wang** (Simon Fraser University)
Auxiliary-try Metropolis: Incorporating Auxiliary Variables into Multiple-try Metropolis / Metropolis essai-auxiliaire : intégration de variables auxiliaires à un Metropolis à essai multiple

14:30-14:45 **Audrey Béliveau** (University of Waterloo) **Xiangshan Kong** (University of Waterloo)
Generalized Fused Lasso for Treatment Pooling in Network Meta-Analysis / Fused lasso généralisé pour la mise en commun des traitements dans la méta-analyse en réseau

---

15:30-17:00 **Invited / Sur invitation** (abstract/résumé 350) **A 1043**

Statistical Challenges: Privacy, High-Dimensionality, and Real-World Applications
Défis statistiques : confidentialité, haute dimensionnalité et applications concrètes

Chair/Président: Zeinab Mashreghi
Organizer/Responsable: Zeinab Mashreghi
Sponsor/Commanditaires: Survey Methods Section/Groupe des méthodes d'enquête

15:30-16:00 **Anne-Sophie Charest** (Université Laval) **Mamadou Mbodj** (Université Laval) **Sébastien Gambs** (Université du Québec à Montréal)
Exploit Membership Inference attacks and Imputation strategies for Attribute disclosure from Estimated models / Exploiter les attaques d'inférence d'appartenance et les stratégies d'imputation pour la divulgation d'attributs à partir de modèles estimés

16:00-16:30 **Mehdi Dagdoug** (McGill University) **David Haziza** (University of Ottawa) **Esther Eustache** (Université de Neuchâtel)
High-dimensional Variance Estimation for Linear Model-assisted Estimation and Linear Imputation / Estimation de la variance en grande dimension pour l'estimation assistée par modèle linéaire et l'imputation linéaire

16:30-17:00 **Augustine Wigle** (University of Waterloo) **Audrey Béliveau** (University of Waterloo)
Estimating Provincial Methane Emissions from Complex Survey Data using a Multi-Stage Framework / Estimation des émissions de méthane à l'échelle provinciale à partir de données d'enquête complexes en utilisant un cadre à plusieurs phases

---

15:30-17:00 **Invited / Sur invitation** (abstract/résumé 352) **C 2045**

Recent Advances in Methods for Incomplete and Complex Survey Data
Progrès récents sur les méthodes des données d'enquête incomplètes et complexes

Chair/Président: Wendy Lou
Organizer/Responsable: Wendy Lou

15:30-16:00 **Trevor James Thomson** (Fred Hutchinson Cancer Center) **Joan X. Hu** (Simon Fraser University)
On Developing a Predictive Survival Model with Internal Time-varying Covariates / Développement d'un modèle de survie prédictif avec covariables internes variant avec le temps

16:00-16:30 **Zilin Wang** (Wilfrid Laurier University) **Mary Thompson** (University of Waterloo)
Modelling Missing-not-at-random for Mental Health Data from Complex Surveys / Modélisation de données manquantes de manière non aléatoire pour données d'enquêtes complexes sur la santé mentale

16:30-17:00 **Changbao Wu** (University of Waterloo)
Analysis of Complex Data through Combining Information from Multiple Sources / Analyse de données complexes par la combinaison d'informations tirées de plusieurs sources

---

15:30-17:00 **Invited / Sur invitation** (abstract/résumé 354) **ED 2018A**

Innovations in Statistical Modeling for Complex Data Structures
Innovations en modélisation statistique des structures de données complexes

Chair/Président: Cindy Feng
Organizer/Responsable: Cindy Feng

15:30-16:00 **Jiguo Cao** (Simon Fraser University) **Barinder Thind** (Simon Fraser University) **Kevin Multani** (Stanford University)
Functional Neural Networks / Réseaux neuronaux fonctionnels

16:00-16:30 **Muye Nanshan** (Simon Fraser University) **Nan Zhang** (Fudan University) **Jiguo Cao** (Simon Fraser University)
Online Functional Principal Component Analysis on a Multidimensional Domain with Dynamic Tuning / Analyse en composantes principales fonctionnelles en ligne sur un domaine multidimensionnel avec réglage dynamique

16:30-17:00 **Lam Ho** (Dalhousie University)
Modelling and Inferring Phenotypic Trait Evolution on Large Phylogenetic Trees / Modélisation et inférence de l'évolution des caractères phénotypiques sur de grands arbres phylogénétiques

---

15:30-17:00 **Invited / Sur invitation** (abstract/résumé 356) **A 1046**

Approaches to Teaching the Analysis of Large-Scale and Complex Data
Approches de l'enseignement de l'analyse de données complexes et à grande échelle

Chair/Président: Alexander Shestopaloff
Organizer/Responsable: Alexander Shestopaloff
Sponsor/Commanditaires: Statistical Education Section/Groupe d'éducation en statistique

15:30-15:52 **Xu (Sunny) Wang** (Wilfrid Laurier University) **Sukhjit Sehra** (Wilfrid Laurier University) **Devan G. Becker** (Wilfrid Laurier University)
A Six-Year Journey - Overview of Laurier Data Science Program - Collaboration and Creativity / Un parcours de six ans : Aperçu du programme de science des données à la Wilfrid Laurier University (WLU) – collaboration et créativité Ⓔ E

15:52-16:15 **Pierre Miasnikof** (University of Toronto) **Cristian Bravo** (University of Western Ontario) **Yuri Lawryshyn** (University of Toronto)
Statistics for networks and networks for statistics – why statisticians should know graphs / Statistique pour les réseaux et les réseaux pour la statistique - pourquoi les statisticiens devraient être familiers avec les graphes Ⓔ E

16:15-16:37 **G. Alexi Rodríguez-Arelis** (The University of British Columbia)
Multiclass Prediction and Inference: A Practical Approach / Prédiction et inférence multi-classes : une approche pratique Ⓔ E

16:37-17:00 **Varada Kolhatkar** (The University of British Columbia)
Making Machine Learning Approachable in Data Science Education / Rendre l'apprentissage automatique accessible dans l'enseignement de la science des données Ⓔ E

15:30-17:00 **Contributed / Communications libres** (abstract/résumé 359) **A 1049**

Advances in Functional Data Analysis
Progrès en analyse des données fonctionnelles

Chair/Président: Haixu Wang

15:30-15:45 **Jairo Diaz-Rodriguez** (York University) **Kelly Ramsay** (York University)
Differentially Private Boxplots / Boîtes à moustaches différentiellement confidentielles Ⓔ E

15:45-16:00 **Amanjot Bhullar** (University of Guelph) **Khurram Nadeem** (University of Guelph) **Ayesha Ali** (University of Guelph) **Evan D. G. Fraser** (University of Guelph)
DeepS^3: A Deep Learning Framework for Predicting Multi-Crop Land Suitability with Satellite Data / DeepS^3 : un cadre d'apprentissage profond pour prédire la qualité de la terre pour plusieurs récoltes avec des données satellites Ⓔ E

16:00-16:15 **Neve Loewen** (University of Manitoba) **Mohammad Jafari Jozani** (University of Manitoba)
Robust Regression Analysis with Nomination Sampling / Analyse de régression robuste avec échantillonnage nominatif Ⓔ E

15:30-17:00 **Contributed / Communications libres** (abstract/résumé 361) **A 2071**

Prediction and Learning
Prédiction et apprentissage

Chair/Président: Shirin Golchi

15:30-15:45 **Ajmery Jaman** (McGill University)
UPoSI Approach to Valid Post-Selection Inference for Penalized G-Estimation / Approche UPoSI de l'inférence post-sélection valide pour l'estimation G pénalisée Ⓔ E

15:45-16:00 **Minzee Kim** (University of Waterloo) **Joel A. Dubin** (University of Waterloo)
A Method for Improving Dynamic Prediction of Joint Models Using a Similarity-Based Approach / Une méthode pour améliorer la prédiction dynamique de modèles conjoints au moyen d'une approche basée sur la similarité

16:00-16:15 **Tatiana Krikella** (University of Waterloo) **Joel A. Dubin** (University of Waterloo)
Determination of Subpopulation Size for Similarity-Based Personalized Prediction Models that Jointly Optimize Discrimination and Calibration / Détermination de la taille de sous-population pour les modèles de prédiction personnalisés basés sur la similarité qui optimisent conjointement la discrimination et la calibration

16:15-16:30 **Li-Pang Chen** (National Chengchi University)
Variable Selection and Estimation for Length-Biased and Partly Interval-Censored Survival Data with Mismeasured Covariates / Sélection de variables et estimation pour les données de survie avec censure par intervalle et biais de longueur en présence de covariables mal mesurées

16:30-16:45 **Weifan Yan** (McGill University) **Jiayi Geng** (McGill Univerisity) **Hui Shen** (McGill Univerisity) **W. Evan Johnson** (Rutgers University) **Eric Kolaczyk** (McGill Univerisity)
Estimation of Bipartite Motif Frequencies in Ligand-Receptor Interaction Networks / Estimation des fréquences des motifs bipartis dans les réseaux d'interactions ligand-récepteur

SSC Presidential Invited Address
Allocution de l'invité de la présidente de la SSC

Chair/Président: Shirley E. Mills

Organizer/Responsable: Shirley E. Mills

Room/Salle: IIC 2001

Date: Monday June 3 / lundi 3 juin

Time/Heure: 08:30-09:50

Abstract/Résumé

[08:30-09:50]

Karen Kafadar (University of Virginia)

Statistics FOR Data Science: Combining Statistics and Exploratory Data Analysis

Statistique pour la science des données : Combiner la statistique et l'analyse exploratoire des données

Data science and machine learning algorithms are sometimes viewed as the only solution to analyze massive datasets. Yet concepts from classical statistics remain critical in such settings. Massive data often have clusters, exotic features, hidden patterns; they invite multiplicity of tests and require statistical thinking to obtain valid inferences. The central goals of data analysis are insight and valid inferences, and effective visualizations are valuable, so we need statistically-based algorithms, methods, and data displays that meet both objectives. Concepts of robustness are especially essential to these methods and displays, when sampling the dataset, estimating key quantities, and communicating insights and inferences. Indeed, classical statistics show that more data does not imply more confidence, especially when they do not represent their target populations. This talk will discuss some recently analyzed datasets and new displays that demonstrate that statistical methods in the 'data science' era remain critical for ensuring properly guided policies based on valid analyses of 'Big Data.

La science des données et les algorithmes d'apprentissage automatique sont parfois considérés comme la seule solution pour analyser des ensembles massifs de données. Pourtant, les concepts de la statistique classique restent essentiels dans de tels contextes. Les données massives présentent souvent des grappes, des caractéristiques exotiques, des modèles cachés ; elles invitent à une multiplicité de tests et nécessitent une réflexion statistique pour obtenir des inférences valides. Les objectifs principaux de l'analyse des données sont la compréhension et les inférences valides, et les visualisations efficaces sont précieuses. Nous avons donc besoin d'algorithmes, de méthodes et d'affichages de données basés sur les statistiques qui répondent à ces deux objectifs. Les concepts de robustesse sont particulièrement essentiels pour ces méthodes et ces affichages, lors de l'échantillonnage de l'ensemble des données, de l'estimation des quantités clés et de la communication des idées et des inférences. En effet, la statistique classique montre qu'un plus grand nombre de données n'implique pas une plus grande confiance, en particulier lorsque celles-ci ne représentent pas les populations cibles. Cet exposé présentera quelques ensembles de données récemment analysés et de nouveaux affichages qui démontrent que les méthodes statistiques de l'ère de la « science des données » restent essentielles pour garantir des politiques correctement guidées basées sur des analyses valides des « Big Data ».

Advances in Spatial and Spatiotemporal Modeling: Uncovering Complex Patterns and Enhancing Inference
Progrès en modélisation spatiale et spatiotemporelle : découverte de modèles complexes et amélioration de l'inférence

Chair/Président: Cindy Feng

Organizer/Responsable: Cindy Feng

Room/Salle: A 1045

Date: Monday June 3 / lundi 3 juin

Time/Heure: 10:20-11:50

Abstract/Résumé

[10:20-10:42]

Rob Deardon (University of Calgary) **Yirao Zhang** (University of Calgary) **Lorna Deeth** (University of Guelph)

Composite Spatial Epidemic Models: Computational Efficiency via Clustering

Modèles d'épidémies spatiales composites : efficacité de calcul grâce au regroupement

Individual-level models (ILMs) of infectious disease transmission allow the incorporation of different individual-level covariates such as spatial location and vaccination status. However, computational burden becomes a problem for large populations, especially when the risk of infection depends on spatial distance between susceptible and infected individuals. We introduce a novel computation time reduction technique via so-called composite ILMs, that divides the population into spatial clusters and model transmission assuming simple, computationally efficient mechanisms for infections between clusters. This approach facilitates a much faster and parallelizable approach to the likelihood computation. Model fitting will be conducted in a Bayesian framework using Markov chain Monte Carlo (MCMC) techniques. The models will be tested using simulated data and applied to data on measles in Germany and the UK 2001 foot-and-mouth disease epidemic.

Les modèles au niveau individuel de transmission des maladies infectieuses permettent d'intégrer différentes covariables au niveau individuel, comme la localisation spatiale et le statut vaccinal. Cependant, la complexité des calculs devient un problème pour les grandes populations, surtout lorsque le risque d'infection dépend de la distance spatiale entre les individus susceptibles d'être infectés et ceux qui le sont. Nous présentons une nouvelle technique de réduction du temps de calcul par le biais de composites au niveau individuel, qui divisent la population en grappes spatiales et modélisent la transmission selon des mécanismes simples et efficaces sur le plan du calcul pour les infections entre les grappes. Cette approche permet de calculer la vraisemblance beaucoup plus rapidement et simultanément. Les modèles seront ajustés dans un cadre bayésien à l'aide de techniques de Monte Carlo par chaînes de Markov. Ensuite, ils seront mis à l'essai avec des données simulées, puis appliqués aux données relatives à la rougeole en Allemagne et à l'épidémie de fièvre aphteuse au Royaume-Uni en 2001.

[10:42-11:05]

Mahmoud Torabi (University of Manitoba) **Charmaine B. Dean** (University of Waterloo) **Georges Bucyibaruta** (Imperial College London)

Innovative Strategies for Influenza Data Examination

Stratégies innovantes pour l'examen des données sur la grippe

We develop a discrete time compartmental model to describe the spread of seasonal influenza virus. As time and disease state variables are assumed to be discrete, this model is considered to be a discrete time, stochastic, Susceptible-Infectious-Recovered-Susceptible (DT-SIRS) model, where weekly counts of disease are assumed to follow a Poisson distribution. We allow the

Nous développons un modèle compartimental à temps discret pour décrire la propagation du virus de la grippe saisonnière. Le temps et les variables d'état de la maladie étant supposés discrets, ce modèle est considéré comme un modèle stochastique en temps discret, Susceptible-Infecté-Rétabli-Susceptible (DT-SIRS), dans lequel les comptes hebdomadaires de cas sont supposés suivre une distribution de Poisson. Le taux de transmission de la ma-

disease transmission rate to also vary over time, and the disease can only be reintroduced after extinction if there is a contact with infected individuals from other host populations. To capture the variability of influenza activities from one season to the next, we define the seasonality with a four-week period effect that may change over years. We examine three different transmission rates and compare their performance to that of existing approaches. The framework is applied in an analysis of the temporal spread of influenza in the province of Manitoba, Canada, 2012–2015.

ladie peut également varier dans le temps et la maladie ne peut être réintroduite après une extinction que s'il y a un contact avec des individus infectés provenant d'autres populations hôtes. Pour saisir la variabilité des activités grippales d'une saison à l'autre, nous définissons la saisonnalité avec un effet de période de quatre semaines qui peut changer au fil des ans. Nous examinons trois taux de transmission différents et comparons leurs performances à celles des approches existantes. Nous appliquons le modèle à une analyse de la propagation temporelle de la grippe dans la province du Manitoba de 2012 à 2015.

[11:05-11:27]

Patrick Brown (Unity Health Toronto) **Jamie Stafford** (University of Toronto)

A Historical Look at Geostatistical Models and Gaussian Markov Random Fields

Un regard historique sur les modèles géostatistiques et les champs aléatoires de Markov gaussiens

The relationship between Gaussian Markov Random Fields and Geostatistical models with a Matérn covariance function has been an essential part of many applications of spatial statistics over the past decade, the equivalence between the two processes is leveraged to produce a sparse precision matrix thereby overcoming the "curse of dimensionality". Lindgren et al's seminal 2011 paper started a sparse matrix revolution by giving a proof of the equivalence using stochastic differential equations. There is, however, a strain of research on the topic involving several papers by Julian Besag and going back as far as a paper by Lévy in 1948. This talk will 1) trace the evolution of the Matérn/GMRF duality over time; 2) explain the equivalence using matrix algebra and trigonometry, in a way which is accessible to those (including this presenter) who struggle with differential equations; and 3) discuss the potential for practical applications of the methodology in higher dimensions.

La relation entre les champs aléatoires de Markov gaussiens (GMRF) et les modèles géostatistiques avec une fonction de covariance de Matérn a été une partie essentielle de nombreuses applications de la statistique spatiale ces dix dernières années, l'équivalence entre les deux processus étant exploitée pour produire une matrice de précision peu dense, surmontant ainsi la « malédiction de la dimensionnalité ». L'article fondateur de Lindgren et al en 2011 a déclenché une révolution des matrices peu denses en apportant la preuve de l'équivalence à l'aide d'équations différentielles stochastiques. Il existe cependant une série de recherches sur le sujet, avec notamment plusieurs articles de Julian Besag et remontant jusqu'à un article de Lévy en 1948. Cet exposé 1) retracera l'évolution de la dualité Matérn/GMRF au fil du temps ; 2) expliquera l'équivalence à l'aide de l'algèbre matricielle et de la trigonométrie, d'une manière accessible à ceux (y compris le présentateur) qui ont des difficultés avec les équations différentielles ; et 3) discutera du potentiel d'applications pratiques de la méthodologie dans les dimensions supérieures.

[11:27-11:50]

Dirk Douwes-Schultz (McGill)

Markov Switching Zero-Inflated Space-Time Multinomial Models for Comparing Multiple Infectious Diseases

Modèles multinomiaux spatio-temporels à commutation de Markov avec excès de zéros pour la comparaison de maladies infectieuses multiples

The modeling of zero-inflated multinomial data across space and time is challenging due to the need to account for correlations across space, time and category in both the count and zero-inflated components of the model. Zero-inflated multinomial models for space-time data have not been considered. Here, we are interested in comparing the transmission dynamics of several co-circulating infectious diseases where some can

La modélisation spatio-temporelle des données multinomiales avec excès de zéros est difficile puisqu'il faut prendre en compte les corrélations entre l'espace, le temps et la catégorie à la fois dans les composantes de comptage et les composantes du modèle avec excès de zéros. Aucun modèle multinomial avec excès de zéros n'a été développé pour les données spatio-temporelles. Nous souhaitons ici comparer la dynamique de transmission de plusieurs maladies infectieuses qui circulent ensemble et dont certaines

be absent for long periods. We first assume there is a baseline disease that is well-established and always present in the region. The other diseases switch between periods of presence and absence in each area through a series of coupled Markov chains that account for disease interactions, covariates and disease spread from neighboring areas. Since we are only interested in comparing the diseases, we assume the cases of the present diseases in an area jointly follow an autoregressive multinomial model. We use the multinomial model to investigate whether there are associations between certain factors, such as temperature, and differences in the transmission intensity of the diseases. Inference is performed using efficient Bayesian Markov chain Monte Carlo methods based on jointly sampling all presence indicators. We apply the model to spatio-temporal counts of dengue, Zika and chikungunya cases in Rio de Janeiro during the first triple epidemic there.

peuvent être absentes pendant de longues périodes. Nous supposons tout d'abord qu'il existe une maladie de base bien établie et toujours présente dans la région. Les autres maladies passent d'une période de présence à une période d'absence dans chaque zone par le biais d'une série de chaînes de Markov couplées qui tiennent compte des interactions entre les maladies, des covariables et de la propagation des maladies à partir de zones voisines. Comme nous nous intéressons uniquement à la comparaison des maladies, nous supposons que les cas de maladies présentes dans une région suivent conjointement un modèle multinomial autorégressif. Nous utilisons ce modèle multinomial pour déterminer s'il existe des associations entre certains facteurs, tels que la température, et des différences dans l'intensité de transmission des maladies. L'inférence est réalisée à l'aide de méthodes bayésiennes efficaces de Monte Carlo en chaîne de Markov, basées sur l'échantillonnage conjoint de tous les indicateurs de présence. Nous appliquons le modèle aux décomptes spatio-temporels des cas de dengue, de Zika et de chikungunya à Rio de Janeiro pendant la première triple épidémie dans cette ville.

Chair/Président: Silvana Pesenti

Organizer/Responsable: Silvana Pesenti

Room/Salle: IIC 2001

Date: Monday June 3 / lundi 3 juin

Time/Heure: 10:20-11:50

Abstract/Résumé

[10:20-10:50]

Fangda Liu (University of Watelroo)

Distributional Uncertainty and Risk Sharing

Incertitude distributionnelle et partage des risques

The model uncertainty is of crucial importance when market participants are making risk management strategies. For a participant who adopts law-invariant risk measures for quantification, the study of the supremum of risk measure values can help the participant to better understand the performance of risk in the worst-case scenario. In this talk, we introduce several model uncertainty settings. The choices of risk measures, uncertainty sets, and transformations of the underlying risk play important roles in the characterization of the worst-case distribution.

L'incertitude du modèle est essentielle lorsque les participants au marché élaborent des stratégies de gestion des risques. L'étude des valeurs de mesure de la borne supérieure du risque peut aider un participant, qui adopte des distributions invariantes de risque pour la quantification, à mieux comprendre les résultats du risque dans le pire scénario. Nous présentons plusieurs modèles d'incertitude. Les choix des mesures de risque, des ensembles d'incertitude et des transformations du risque sous-jacent jouent un rôle important dans la caractérisation de la pire distribution.

[10:50-11:20]

Thai H. Nguyen (Université Laval)

Pareto-Optimal Investments and Contracting for Non-linear Payoffs

Investissements d'optimum de Pareto et passation de contrat pour des gains non linéaires

This paper explores financial and insurance contracts with non-linear payoffs by combining optimal contract design and dynamic portfolio planning. It avoids upfront parameter fixation, seeking to simultaneously optimize contract parameters and investment strategies in a Pareto-optimal manner. This approach sheds light on the implications of dynamic investment strategies, especially in contracts with non-linear elements like caps or investment guarantees, aiming to enhance their design and potential for Pareto improvements.

Cet article explore les contrats financiers et d'assurance avec des gains non linéaires en combinant la conception optimale de contrat et la planification dynamique de portefeuille. Nous évitons la fixation sur les paramètres initiaux, et cherchons plutôt à optimiser simultanément les paramètres de contrat et les stratégies d'investissement à la manière d'optimum de Pareto. Cette approche nous éclaire sur les conséquences des stratégies d'investissement dynamiques, tout particulièrement dans des contrats ayant des éléments non linéaires comme les actions privilégiées convertibles à taux variable ou les investissements garantis, et cherche à améliorer leur conception et leur potentiel relatif à l'optimisation de Pareto.

[11:20-11:50]

Ilaria Peri (Birkbeck, University of London) **Akif Ince** (Birkbeck, University of London) **Marlon Moresco** (Federal University of Rio Grande do Sul) **Silvana Pesenti** (University of Toronto)

Elictable Risk Functionals with Quasi-convex Score
Fonctionnelles de risque élicitables avec score quasi-convexe

We introduce a novel class of elicitable risk functionals defined on an Orlicz space, providing a generalization that encompasses a broad spectrum of risk measures. This inclusive class incorporates well-known risk measures such as expectiles, quantiles, M-quantiles, Lp-quantiles, lambda quantiles, and shortfall risk measures. Specifically, our approach involves quasi-convex scoring functions, allowing for a diverse range of scores, including non- continuous ones. We establish fundamental properties, including existence, finiteness, convexity, positive homogeneity, translation invariance, and monotonicity. Furthermore, we derive first-order conditions for these elicitable functionals, facilitating efficient computational methods. Additionally, we show that the quasi-convex scoring functions can be constructed using a distance function and a weight function, which allows for explicit jumps and asymmetries.

Nous introduisons une nouvelle classe de fonctionnelles de risque élicitables définies sur un espace d'Orlicz, sooit une généralisation qui englobe un large spectre de mesures de risque. Cette classe englobe des mesures de risque bien connues telles que les expectiles, les quantiles, les M-quantiles, les Lp-quantiles, les lambda-quantiles et les mesures de risque de manque à gagner. Plus précisément, notre approche fait appel à des fonctions de notation quasi-convexes, ce qui permet d'utiliser une gamme variée de scores, y compris des scores non continus. Nous en établissons les propriétés fondamentales, notamment l'existence, la finitude, la convexité, l'homogénéité positive, l'invariance de la translation et la monotonie. En outre, nous dérivons des conditions de premier ordre pour ces fonctionnelles élicitables, ce qui facilite l'utilisation de méthodes de calcul efficaces. Nous montrons également que les fonctions de notation quasi-convexes peuvent être construites à l'aide d'une fonction de distance et d'une fonction de poids, ce qui permet des sauts et des asymétries explicites.

Data Science and Analytics Section Keynote Lecture
Présentation principale du Groupe de science des données et analytique

Chair/Président: Nathaniel Tyler Stevens

Organizer/Responsable: Nathaniel Tyler Stevens

Room/Salle: A 1043

Date: Monday June 3 / lundi 3 juin

Time/Heure: 10:20-11:50

Abstract/Résumé

[10:20-11:50]

Xiaoli Meng (Harvard University)

Privacy, Data Privacy, and Differential Privacy

Vie privée, confidentialité des données et confidentialité différentielle

This talk traces the concept of privacy from a late 19th-century legal right—spurred by tabloid harassment—to today's digital age challenges. It highlights Differential Privacy (DP), a cryptography-based method designed to balance data privacy and utility in a quantified way. However, despite DP's advancements and explicit warnings (e.g., Kiferand Machanavajjhala, 2011, "No free lunch in data privacy"; Tschantz et al, 2022, "SoK: Differential privacy as a causal property"), misconceptions about its resistance to adversaries' prior knowledge persist. By revisiting Warner's (1965, JASA) randomized response mechanism, we argue that this misperception lies in treating data as static objects, rather than realizations of underlying, typically interdependent attributes or variables. We show how DP's effectiveness can falter when adversaries exploit interdependencies among individuals—similar to how quarantining only symptomatic individuals fails to stop an airborne disease. A holistic statistical perspective on joint modeling data is therefore as crucial for data privacy as for data analysis. (Joint work with James Bailie and Ruobin Gong.)

Cet exposé retrace l'évolution du concept de protection de la vie privée depuis un droit légal de la fin du 19e siècle, encouragé par le harcèlement dans les tabloïdes, jusqu'aux défis actuels de l'ère numérique. Il met en lumière la confidentialité différentielle (CD), méthode basée sur la cryptographie et conçue pour équilibrer la confidentialité et l'utilité des données de manière quantifiée. Cependant, malgré les progrès de la CD et les avertissements explicites (par exemple, Kifer et Machanavajjhala, 2011 « No free lunch in data privacy »; Tschantz et al, 2022, « SoK : Differential privacy as a causal property »), les idées fausses sur sa résistance aux connaissances préalables des adversaires persistent. En revisitant le mécanisme de réponse aléatoire de Warner (1965, JASA), nous soutenons que cette perception erronée réside dans le fait de traiter les données comme des objets statiques, plutôt que comme des réalisations d'attributs ou de variables sous-jacents, généralement interdépendants. Nous montrons comment l'efficacité de la CD peut s'affaiblir lorsque les adversaires exploitent les interdépendances entre les individus, de la même manière que la mise en quarantaine des seuls individus symptomatiques ne parvient pas à enrayer une maladie transmise par l'air. Une perspective statistique holistique sur les données de modélisation conjointe est donc aussi cruciale pour la protection de la vie privée que pour l'analyse des données. (Travail conjoint avec James Bailie et Ruobin Gong.)

Harnessing Statistical and Computational Models for Neuroscience
Exploiter les modèles statistiques et informatiques pour la neuroscience

Chair/Président: Reza Ramezan

Organizer/Responsable: Reza Ramezan

Room/Salle: C 2045

Date: Monday June 3 / lundi 3 juin

Time/Heure: 10:20-11:50

Abstract/Résumé

[10:20-10:50]

Reza Ramezan (University of Waterloo) **Farouk Nathoo** (University of Victoria) **Cedric Beualac** (Université du Québec à Montréal) **Michelle Miranda** (University of Victoria) **Jiguo Cao** (Simon Fraser University) **Liangliang Wang** (Simon Fraser University) **Mirsa Beg** (Simon Fraser University) **Yin Song** (University of Victoria) **Leno Rocha** (University of Victoria) **Sidi Wu** (Simon Fraser University) **Erin Gibson** (Simon Fraser University)

Neural Network Feature Extraction and Bayesian Spatial Modeling for Imaging Genetics

Extraction de caractéristiques d'un réseau neuronal et modélisation spatiale bayésienne pour l'imagerie génétique

Dealing with the high dimension of both neuroimaging data and genetic data is a difficult problem in the association of genetic data to neuroimaging. We tackle the latter problem with an eye toward developing solutions that are relevant for disease prediction. Our proposed solution uses neural networks to extract from neuroimaging data features that are relevant for predicting Alzheimer's Disease (AD) for subsequent relation to genetics. The neuroimaging-genetic pipeline we propose is comprised of image processing, neuroimaging feature extraction and genetic association steps. We present a neural network classifier for extracting neuroimaging features that are related with the disease. The proposed method is data-driven and requires no expert a priori selection of regions of interest. We further propose a spatial multivariate regression to relate the extracted features to genetics.

Le traitement simultané de la dimension élevée des données de neuro-imagerie et des données génétiques est un problème difficile dans l'association des données génétiques à la neuro-imagerie. Nous nous attaquons à ce dernier problème en vue de développer des solutions pertinentes pour la prédiction des maladies. Notre solution proposée exploite les réseaux neuronaux afin d'extraire à partir de données de neuro-imagerie des caractéristiques pertinentes dans la prédiction de la maladie d'Alzheimer (MA) pour la relation subséquente à la génétique. La liaison neuro-imagerie-génétique que nous proposons comprend le traitement de l'image, l'extraction de caractéristiques de la neuro-imagerie et les étapes d'association génétique. Nous présentons un classificateur de réseau neuronal pour extraire les caractéristiques de neuro-imagerie liées à la maladie. La méthode proposée est axée sur les données et ne nécessite pas de sélection a priori des régions d'intérêt par un expert. Nous proposons en outre une régression spatiale multivariée pour relier les caractéristiques extraites à la génétique.

[10:50-11:20]

Lloyd T. Elliott (Simon Fraser University)

Mediation Analysis shows effects for the LIFO Network in Brain Imaging Genetics

L'analyse de médiation montre des effets pour le réseau LIFO dans la génétique par imagerie cérébrale

Genome-wide association studies with pairwise associations between phenotypes and genetic variants provide a framework for brain imaging genetics. When a genetic variant correlates with both a disease and with an image-derived phenotype, the disease effect may be modulated

Les études d'association à l'échelle du génome avec des associations par paire entre phénotypes et variantes génétiques fournissent un cadre pour la génétique de l'imagerie cérébrale. Lorsqu'une variante génétique est corrélée à la fois à une maladie et à un phénotype dérivé de l'image, l'effet de la maladie peut être

by the image-derived phenotype. The LIFO (last in, first out) network is a brain network that can degenerate early and can be operationalized as an image-derived phenotype. LIFO is associated with Alzheimer's disease, and Alzheimer's disease is also associated with the MAPT genetic cluster. In collaboration with Oxford's Wellcome Centre for Integrative Neuroimaging (WIN), we analyze the LIFO-MAPT-Alzheimer's disease triangle and find an effect using VanderWeele mediation. The analysis is complicated by the non-binary nature of the genotype, and also by chromosome 17 inversion.

modulé par ce phénotype. Le réseau LIFO (last in, first out) est un réseau cérébral qui peut dégénérer précocement et qui peut être opérationnalisé en tant que phénotype dérivé de l'image. Le LIFO est associé à la maladie d'Alzheimer, et la maladie d'Alzheimer est également associée au groupe génétique MAPT. En collaboration avec le Wellcome Centre for Integrative Neuroimaging (WIN) d'Oxford, nous analysons le triangle LIFO-MAPT-maladie d'Alzheimer et trouvons un effet en utilisant la médiation de VanderWeele. L'analyse est compliquée par la nature non binaire du génotype et par l'inversion du chromosome 17.

[11:20-11:50]

Meixi Chen (University of Waterloo) **Martin Lysy** (University of Waterloo) **Reza Ramezan** (University of Waterloo)

Insights into Brain Dynamics: A Scalable Spike-Train Model for Neuronal Interactions

Perspectives sur les dynamiques cérébrales : un modèle extensible de trains d'impulsions nerveuses pour les interactions neuronales

Spike trains, which are successive electrochemical signals generated by nerve cells, can facilitate inference about the brain's state in a given environment. Inference about functional connectivity (FC), e.g. the statistical correlation between neurons based on spike trains, offers crucial insight on the interactions between different brain areas. The technological advancement in neural recording provides an abundance of data for statistical analyses. However, achieving both biological interpretability and computational scalability poses significant challenges in modelling FC. In this talk, we introduce a novel multi-neuron latent dynamics model based on the spike generation mechanism, coupled with an efficient approximate Bayesian inference procedure. To facilitate downstream analyses, we present a convenient test statistic for comparing inferred FCs. Application of our method to experimental data uncovers changes in FC in response to alcohol cues in the orbitofrontal cortex of rats.

Les trains d'impulsions nerveuses sont des signaux électrochimiques successifs générés par les cellules nerveuses et peuvent faciliter l'inférence de l'état du cerveau dans un environnement donné. L'inférence de la connectivité fonctionnelle (FC), p. ex. la corrélation statistique entre les neurones basée sur les trains d'impulsion nerveuse, procure une perspective importante sur l'interaction entre différentes régions du cerveau. L'avancée technologique dans l'enregistrement neuronal offre une abondance de données pour les analyses statistiques. Cependant, il est particulièrement difficile de réaliser une interprétation biologique et une extensibilité calculatoire dans la modélisation de FC. Lors de cet exposé, nous présentons un nouveau modèle de dynamiques latentes multineuronales basé sur le mécanisme de génération d'impulsion, accompagné d'une procédure efficace d'inférence bayésienne approximative. Pour faciliter les analyses en aval, nous présentons des variables à tester pratiques pour comparer les FC inférées. L'application de notre méthode à des données expérimentales couvre les changements dans la FC en réponse aux signaux d'alcool dans le cortex orbitofrontal de rats.

Inference for Autoregressive and Markov Models; A Memorial Session for Wolfgang Wefelmeyer
Inférence des modèles autorégressifs et de Markov ; session commémorative en l'honneur de Wolfgang Wefelmeyer

Chair/Président: Thomas Salisbury

Organizer/Responsable: Priscilla E. Greenwood

Room/Salle: ED 2018A

Date: Monday June 3 / lundi 3 juin

Time/Heure: 10:20-11:50

Abstract/Résumé

[10:20-10:50]

Priscilla E. Greenwood (The University of British Columbia)

Work of Wolfgang Wefelmeyer: Asymptotic Efficiency, Inference for Stochastic Processes, Non-parametric and Semiparametric Estimation

Travail de Wolfgang Wefelmeyer : efficacité asymptotique, inférence pour les processus stochastiques, estimation semi-paramétrique et non paramétrique

My talk will be an overview of the work of Wolfgang Wefelmeyer in asymptotic efficiency of estimators, inference for stochastic processes, non-parametric and semi-parametric estimation.

Mon exposé passera en revue le travail de Wolfgang Wefelmeyer concernant l'efficacité asymptotique des estimateurs, l'inférence pour les processus stochastiques et l'estimation semi-paramétrique et non paramétrique.

[10:50-11:20]

Ursula U. Müller (Texas A&M University)

Estimation for Markov Chains with Periodically Missing Observations

Estimation pour les chaînes de Markov avec observations périodiquement manquantes

When we observe a stationary time series with observations missing at periodic time points, we can still estimate its marginal distribution well. However, the dependence structure of the time series may not be recoverable at all, or the usual estimators may have much larger variance than in the fully observed case. We show how non-parametric estimators can often be improved by adding unbiased estimators. We consider a simple setting, first-order Markov chains on a finite state space, and an observation pattern in which a fixed number of consecutive observations is followed by an observation gap of fixed length, say workdays and weekends. In this talk I will focus on the simplest reasonable scenario, namely when every third observation is missing. The new estimators perform astonishingly well, as illustrated with simulations for this scenario. This talk is based on joint work with Anton Schick and Wolfgang Wefelmeyer.

Lorsqu'on étudie une série temporelle stationnaire avec des observations manquantes à des moments périodiques, il est possible de bien estimer sa distribution marginale. Cependant, la structure de dépendance de la série temporelle peut ne pas être récupérable du tout, ou les estimateurs habituels peuvent avoir une variance beaucoup plus grande que dans le cas d'observations complètes. Nous montrons comment améliorer les estimateurs non paramétriques dans bien des cas par l'ajout d'estimateurs sans biais. Nous considérons un cadre simple, des chaînes de Markov du premier ordre sur un espace d'état fini, et un modèle d'observation dans lequel un nombre fixe d'observations consécutives est suivi d'une lacune d'observation de longueur fixe (jours ouvrables et fins de semaine, par exemple). Dans cet exposé, je me concentrerai sur le scénario raisonnable le plus simple, à savoir lorsque chaque troisième observation est manquante. Les nouveaux estimateurs sont étonnamment performants, comme l'illustrent des simulations pour ce scénario. Cet exposé est basé sur un travail conjoint avec Anton Schick et Wolfgang Wefelmeyer.

[11:20-11:50]

Anton Schick (Binghamton University) **Wolfgang Wefelmeyer** (Universitaet zu Koeln)

Inference for Autoregressive and Markov Models; A Memorial Session for Wolfgang Wefelmeyer
Inférence des modèles autorégressifs et de Markov ; session commémorative en l'honneur de Wolfgang Wefelmeyer

Efficient Density Estimation in an AR(1) Model
Estimation efficace de la densité dans un modèle AR(1)

A class of plug-in estimators of the stationary density of an autoregressive model with autoregression parameter in the open interval (0,1) is studied. Two estimators of the innovation density are used, a standard kernel estimator and a weighted kernel estimator with weights chosen to mimic the condition that the innovation density has mean zero. Bahadur expansions are obtained for this class of estimators in the Banach space of integrable functions. These stochastic expansions establish root-n consistency in this space. It is shown that the density estimators based on the weighted kernel estimators are asymptotically efficient if an asymptotically efficient estimator of the autoregression parameter is used. Here asymptotic efficiency is understood in the sense of the Hajek–LeCam convolution theorem.

Notre étude porte sur une classe d'estimateurs plug-in de la densité stationnaire d'un modèle autorégressif avec un paramètre d'autorégression dans l'intervalle ouvert (0,1). Nous utilisons deux estimateurs de la densité d'innovation, soit un estimateur par noyau standard et un estimateur par noyau pondéré avec des poids choisis pour imiter la condition d'une densité d'innovation de moyenne nulle. Des extensions de Bahadur sont obtenues pour cette classe d'estimateurs dans l'espace de Banach de fonctions intégrables. Ces extensions stochastiques établissent une cohérence racine-n dans cet espace. Nous montrons que les estimateurs de la densité basés sur des estimateurs par noyau pondérés sont asymptotiquement efficaces lorsqu'un estimateur asymptotiquement efficace du paramètre d'autorégression est utilisé. Il faut ici comprendre l'efficacité asymptotique dans le sens du théorème de convolution de Hajek—LeCam.

Networking for Large Programs of Research (Panel)
Le réseautage dans les grands programmes de recherche (Table ronde)

Chair/Président: Thérèse A. Stukel

Organizer/Responsable: Thérèse A. Stukel

Room/Salle: ED 2018B

Date: Monday June 3 / lundi 3 juin

Time/Heure: 10:20-11:50

Abstract/Résumé

[10:20-11:50]

Lisa M. Lix (University of Manitoba) **Robyn Tamblyn** (McGill University) **Robert W. Platt** (McGill University) **Shelley Bull** (Lunenfeld-Tanenbaum Research Institute, Sinai Health, and University of Toronto)

Networking for Large Programs of Research

Le réseautage dans les grands programmes de recherche

Networking is the process of establishing and nurturing professional relationships to advance career goals. This is important for individual career advancement as well as to advance programs of research. Large research programs require a group of professionals from varying specialties with different skills who can work collaboratively on common projects. This invited panel assembles a diverse group of researchers in statistics and related STEM fields who have been successful at establishing successful national or provincial research networks and explores the attributes of a good network and the strategies used to achieve it. The benefits to trainees and junior investigators in these networks will also be explored.

Le réseautage est le processus qui consiste à établir et à entretenir des relations professionnelles afin de faire progresser ses objectifs de carrière. Cela est important aussi bien pour l'avancement de la carrière individuelle que pour faire progresser les programmes de recherche. Les grands programmes de recherche rassemblent un groupe de professionnels issus de diverses spécialités et dotés de compétences différentes, qui doivent être capables de travailler en collaboration sur des projets communs. Cette table ronde rassemble un groupe diversifié de chercheurs en statistique et autres domaines STIM connexes qui ont réussi à établir des réseaux de recherche nationaux ou provinciaux fructueux ; elle explore les attributs d'un bon réseau et les stratégies utilisées pour y parvenir. Nous examinerons également les avantages de ces réseaux pour les stagiaires et les jeunes chercheurs.

Chair/Président: Gyanendra Pokharel

Room/Salle: A 1049

Date: Monday June 3 / lundi 3 juin

Time/Heure: 10:20-11:50

Abstract/Résumé

[10:20-10:35]

Abigail McGrory (University of Toronto) **Anna Heath** (The Hospital for Sick Children)

Enhancing the Efficiency of Adaptive Platform Trials Through the Exploration of Alternative Treatment Ranking Methods

Augmentation de l'efficacité des essais de plateforme adaptatifs grâce à l'exploration de méthodes de classement de traitement de deuxième ligne

Adaptive platform trials (APTs) are a flexible and versatile clinical trial design. A common approach to compare treatments in an APT is to determine the posterior probability that an intervention is the best and compare it to a superiority threshold. This method, however, has been criticized and little research has been done to determine if alternative methods perform better. This study identified and assessed three alternative methods to evaluate multiple treatments in an APT. A simulation study was implemented across multiple trial designs to determine which ranking method performed best by comparing the power and expected sample size of each trial. Additionally, optimization algorithms were applied to identify which superiority threshold maximized the power of each trial. Results show that all methods perform well, but only when used at the optimal superiority threshold. Overall, this work will increase the efficiency of APTs and help patients access novel treatments faster.

Les essais de plateforme adaptatifs (APTs) représentent un type d'essai clinique polyvalent et flexible. Une approche fréquente pour la comparaison de traitements dans un APT est de déterminer la probabilité a posteriori qu'une intervention soit la meilleure et de la comparer à un seuil de supériorité. Cependant, cette méthode est critiquée et il existe peu de recherche pour déterminer si les autres méthodes sont supérieures. Cette étude a identifié et évalué trois autres méthodes pour évaluer plusieurs traitements dans un APT. Nous avons implanté une étude en simulations à travers plusieurs modèles d'essai afin de déterminer quelle méthode de classement fonctionnait le mieux en comparant la puissance et la taille d'échantillon attendue de chaque essai. De plus, nous avons appliqué des algorithmes d'optimisation pour identifier quel seuil de supériorité maximise la puissance de chaque essai. Les résultats démontrent que toutes les méthodes fonctionnent bien, mais seulement lorsqu'elles sont utilisées à un seuil de supériorité optimal. En général, ce travail améliorera l'efficacité des APTs et facilitera l'accès rapide à de nouveaux traitements pour les patients.

[10:35-10:50]

Zelalem Firisa Negeri (University of Waterloo) **Narayanaswamy Balakrishnan** (McMaster University)

Nonparametric statistical methods for diagnostic test meta-analyses

Méthodes statistiques non paramétriques pour les méta-analyses de tests diagnostiques

Outlying studies, commonly defined as those that result in an inflated overall test accuracy or heterogeneity parameter, are prevalent in meta-analyses of diagnostic test accuracy studies (DTAs). In meta-analyses of DTAs, random effects are commonly used to describe the variability in test accuracy beyond that predicted by the within-study variance. However, that may not sufficiently explain the whole variation when out-

Les études aberrantes, souvent définies comme celles dont la précision générale de test ou dont le paramètre d'hétérogénéité sont exagérés, sont répandues dans les méta-analyses d'études de précision de tests diagnostiques (DTAs). Dans ces méta-analyses, les effets aléatoires servent fréquemment à décrire la variabilité de la précision de test au-delà de celle prédite par la variance à l'intérieur de l'étude. Cependant, ce n'est pas toujours suffisant pour expliquer la variation dans son ensemble en présence

lying studies are present. Thus, the current standard and widely used statistical approaches may lead to misleading statistical inferences when such unusual studies are included in a meta-analysis. Moreover, these standard methods assume restrictive distributional assumptions that may be violated when a few studies are meta-analyzed, are computationally demanding as they use iterative numerical methods for parameter estimation, and may fail to converge for sparse data. We will develop and evaluate robust nonparametric bivariate random effects models for diagnostic test meta-analyses to overcome these limitations. We will illustrate the performance of the developed statistical methods using real-life and simulated meta-analytic data.

d'études aberrantes. Conséquemment, les approches standards actuelles et couramment employées peuvent mener à des inférences trompeuses quand de telles études inhabituelles sont incluses dans une méta-analyse. De plus, ces méthodes standards considèrent des hypothèses distributionnelles restrictives qui pourraient ne pas être respectées lorsque quelques études sont méta-analysées. Elles exigent aussi beaucoup de temps de calcul, car elles utilisent des méthodes numériques itératives pour estimer les paramètres, et peuvent donc échouer à converger pour des données éparses. Nous développerons et évaluerons des modèles non paramétriques robustes d'effets aléatoires bivariés pour les méta-analyses de tests diagnostiques afin de surmonter ces limites. Nous illustrerons la performance des méthodes statistiques conçues à l'aide des données de méta-analyses réelles et simulées.

[10:50-11:05]

Chen Chen (University of Toronto) **Aya A. Mitani** (University of Toronto)

A Joint Model of Hierarchical Data with Multivariate Skewed-t Distribution and Informative Cluster Size

Modèle conjoint de données hiérarchiques avec distribution multivariée asymétrique-t et taille de grappe informative

Pocket depth (PD) and clinical attachment loss (CAL) are important clinical measurements to assess the severity of periodontal disease. Previous studies have used a multivariate skewed-t distribution to model both PD and CAL, assuming data missing at random. However, PD and CAL are related to the number of teeth within a person, resulting in the issue of informative cluster size. As the measurements are clustered within teeth, we extended the multivariate skewed-t model to a joint model using a Bayesian estimation approach. The multivariate skewed-t model was reconstructed in a hierarchical structure, and a cluster size model assumed a binomial distribution with a logit link. The proposed model was evaluated in simulations and compared to the multivariate skewed-t model without the cluster size. We then applied the proposed joint model to dental data from 876 subjects in the San Juan Overweight Adults Longitudinal Study.

La profondeur des poches (PP) et la perte d'attache clinique (PAC) sont des mesures cliniques importantes pour évaluer la gravité de la maladie parodontale. Des études antérieures ont utilisé une distribution asymétrique-t multivariée pour modéliser à la fois la PP et la PAC, en supposant que les données manquantes le soient de façon aléatoire. Cependant, la PP et la PAC sont liées au nombre de dents d'une personne, ce qui pose le problème de la taille des grappes informatives. Comme les mesures sont regroupées à l'intérieur des dents, nous avons étendu le modèle multivarié asymétrique-t à un modèle conjoint en utilisant une approche d'estimation bayésienne. Nous avons reconstruit le modèle asymétrique-t multivarié dans une structure hiérarchique, et pour le modèle de taille de grappe nous avons supposé une distribution binomiale avec un lien logit. Nous avons évalué le modèle proposé avec des simulations et l'avons comparé au modèle multivarié asymétrique-t sans la taille de la grappe. Nous avons ensuite appliqué le modèle conjoint proposé aux données dentaires de 876 sujets de l'étude longitudinale des adultes en surpoids de San Juan.

[11:05-11:20]

Cody B. Halden (University of Ottawa) **Jemila Hamid** (University of Ottawa)

Performance of Iteratively Reweighted Growth Curve Model

Performance du modèle de courbe de croissance pondéré de manière itérative

Growth Curve Models are generalized ANOVA models for longitudinal or dose-response data. They involve within and between individual design matrices, yielding bilinear projections, termed bilinear regression models. Assuming multivariate normality, explicit likelihood solutions exist, optimal for small samples. Both

Les modèles de courbe de croissance sont des modèles ANOVA généralisés pour les données longitudinales. Ils impliquent des matrices de conception intra et interindividuelles, produisant des projections bilinéaires, appelées modèles de régression bilinéaire. En supposant une normalité multivariée, des solutions explicites de vraisemblance existent, optimales pour les petits échantillons.

least squares and maximum likelihood inference lead to weighted inference, by random weights related to the pooled sample covariance matrix. Despite methodological advances, the benefit of weighted inference and weighting choices remains unexplored. Our study conducted simulations to assess different weighting strategies' optimality and their improvement over unweighted inference. Results indicate that while weighting performs slightly better in moderate to large samples, it doesn't significantly outperform unweighted estimators. Weighting collapses in high-dimensionality, prompting new likelihood-based weights to address limitations and a proposed novel reweighting process.

L'inférence par moindres carrés et par maximum de vraisemblance conduit à une inférence pondérée, utilisant des poids aléatoires liés à la matrice de covariance de l'échantillon global. Malgré les avancées méthodologiques, l'avantage de l'inférence pondérée et le choix des poids restent inexplorés. Notre étude a réalisé des simulations pour évaluer l'optimalité des stratégies de pondération et leur amélioration par rapport à l'inférence non pondérée. Les résultats indiquent que la pondération, bien que légèrement meilleure dans les échantillons modérés, ne surpasse pas significativement les estimateurs non pondérés. La pondération s'effondre en haute dimension, ce qui conduit à l'élaboration de nouveaux poids basés sur la vraisemblance pour remédier aux limitations et à la proposition d'un nouveau processus de repondération.

[11:20-11:35]

Mamadou Yauck (Université du Québec à Montréal (UQAM))

A statistical Test for Detecting Homophily and Preferential Recruitment in Link-Tracing Sampling Surveys

Test statistique pour détecter l'homophilie et le recrutement préférentiel dans les enquêtes par sondage à dépistage de liens

Consider a network of N¿1 individuals sharing social connections. Homophily, or the tendency for individuals with similar characteristics to form social ties - or become neighbors -, is a common source of network dependence and can be a source of invalid inference when ignored. This talk considers a chain-referral sampling design on the social network, e.g. a sampling process in which social ties are explored from one neighbor to another. Preferential recruitment characterizes the tendency for individuals to recruit neighbors with similar traits. Distinguishing between homophily and preferential recruitment from a chain-referral sample is a challenging question and the subject of this talk. We present a new statistical test for detecting homophily and preferential recruitment in a chain-referral sample and investigate its performance.

Supposons un réseau d'individus N¿1 partageant des liens sociaux. L'homophilie ou la tendance des individus ayant des caractéristiques similaires à former des liens sociaux – ou à devenir voisins – est une source courante de dépendance de réseau et peut être une source d'inférence invalide si on n'en tient pas compte. Dans cette présentation, nous examinons un plan d'échantillonnage par renvoi en chaîne sur le réseau social, c'est-à-dire un processus d'échantillonnage dans lequel les liens sociaux sont explorés d'un voisin à l'autre. Le recrutement préférentiel désigne la tendance des individus à recruter des voisins présentant des caractéristiques similaires. Nous abordons la question difficile de la distinction entre l'homophilie et le recrutement préférentiel à partir d'un échantillon de référencement en chaîne. Nous présentons un nouveau test statistique pour détecter l'homophilie et le recrutement préférentiel dans un échantillon de référence en chaîne, puis nous analysons les résultats.

[11:35-11:50]

Sean Xinyang Feng (University of Toronto) **Aya A. Mitani** (University of Toronto)

Multivariate Joint Modeling for Clustered Data with Application to Periodontal Disease

Modélisation conjointe multivariée pour des données regroupées appliquée à la parodontopathie

Multivariate joint modeling that integrates multiple longitudinal data and a terminal event for clustered data has increased interest in medical research. In the study of periodontal disease, probing pocket depth and recession serve as tooth-level biomarkers that are associated with the risk of tooth loss. Considering the natural clustering of the teeth within individuals, we propose a clustered multivariate joint model to assess the effect of multiple tooth-level longitudinal biomarkers on the risk of tooth

La modélisation conjointe multivariée intégrant plusieurs données longitudinales et un événement final pour des données regroupées ont attiré l'attention en recherche médicale. Dans l'étude de la parodontopathie, le sondage de la profondeur de la poche parodontale et la récession représentent des biomarqueurs au niveau des dents associées au risque de perte de dent. Compte tenu du regroupement naturel de la dent chez les individus, nous proposons un modèle conjoint multivarié regroupé pour évaluer l'effet de plusieurs biomarqueurs longitudinaux au niveau de la

loss. We estimate our joint model using a Bayesian estimation approach. We showed that our method outperforms the alternative joint model that ignores the cluster effects. We further use our proposed joint model to make dynamic prediction on the probability of tooth loss within a specified time horizon, given a set of baseline and longitudinal predictors, and assess the performance of dynamic prediction for clustered data.

dent sur le risque de perte de dent. Nous estimons notre modèle conjoint à l'aide d'une approche d'estimation bayésienne. Nous avons démontré que notre méthode surpasse les autres modèles conjoints qui ignorent les effets de regroupement. De plus, nous employons le modèle conjoint proposé pour réaliser une prédiction dynamique de la probabilité de perte de dent à l'intérieur d'une période de temps déterminée. Puis en fonction d'un ensemble de prédicteurs longitudinaux et de base, nous évaluons la performance de la prédiction dynamique pour des données regroupées.

Chair/Président: Mohammad Kaviul Anam Khan

Room/Salle: A 2071

Date: Monday June 3 / lundi 3 juin

Time/Heure: 10:20-11:50

Abstract/Résumé

[10:20-10:35]

Mohammad Reza Fahimi (University of Toronto Dalla Lana School of Public Health) **Aya A. Mitani** (University of Toronto) **Osvaldo Espin-Garcia** (Western University)

Optimal Sampling Fractions in Two-Phase Designs with Ordinal Outcomes

Fractions d'échantillonnage optimales dans les plans à deux phases avec réponses ordinales

The two-phase design is a cost-effective approach useful when measuring certain variables for all study participants is prohibitive. Initially, outcomes and inexpensive variables are collected from everyone. Then, individuals are selected based on these data only and undergo measurement of the expensive variables. We propose a novel semi-parametric maximum likelihood approach for making inference in two-phase studies with ordinal outcomes. We consider four models namely proportional odds, adjacent category, stopping ratio, and stereotype regression. Our goal is to identify sampling fractions that minimize the asymptotic variance of the parameter of interest via constrained and stochastic optimization techniques. We compare the performance of the four studied ordinal models under the optimal sampling fractions against simple random sampling and balanced designs. We evaluate the proposed methods in extensive simulations and illustrate the application on a study of knee osteoarthritis.

Le plan en deux phases est une approche rentable utile lorsque mesurer certaines variables auprès de tous les participants d'étude est prohibitif. Dans un premier temps, des réponses et des variables peu coûteuses sont collectés auprès de tous les participants. Ensuite, les individus sont sélectionnés sur la base de ces seules données et les variables coûteuses sont mesurées. Nous proposons une nouvelle approche semi-paramétrique du maximum de vraisemblance pour faire des inférences dans les études en deux phases avec des réponses ordinales. Nous considérons quatre modèles, à savoir les probabilités proportionnelles, la catégorie adjacente, le ratio d'arrêt et la régression stéréotypée. Notre objectif est d'identifier les fractions d'échantillonnage qui minimisent la variance asymptotique du paramètre d'intérêt par le biais de techniques d'optimisation stochastique et sous contrainte. Nous comparons les performances des quatre modèles ordinaux étudiés pour les fractions d'échantillonnage optimales par rapport à l'échantillonnage aléatoire simple et aux plans équilibrés. Nous évaluons les méthodes proposées dans des simulations approfondies et illustrons l'application sur une étude de l'arthrose du genou.

[10:35-10:50]

Emily Somerset (University of Toronto)

Estimating and Forecasting Disease Trends from Wastewater Surveillance Data

Estimation et prévision des tendances d'une maladie à partir des données de surveillance des eaux usées

Wastewater-based surveillance has been proposed as a cost-effective and real-time method for disease monitoring at the community level. It offers a representative depiction of disease presence, irrespective of access to clinical testing and the sample size of symptomatic individuals undergoing testing. We introduce a general model framework to estimate and forecast disease activ-

La surveillance des eaux usées a été proposée comme méthode rentable et applicable en temps réel pour le suivi de maladies à l'échelle des communautés. Elle offre une description représentative de la présence d'une maladie, indépendamment de l'accès à des tests cliniques et de la taille de l'échantillon des sujets symptomatiques soumis à des tests. Nous présentons un cadre de modélisation général pour l'estimation et la prévision de

ity within communities. One advantage of this framework is its ability to distinguish between wastewater viral signals that are common across communities and those that are specific to each community. Each of these signals are simultaneously estimated with a functional form of their derivatives, to enhance forecasting accuracy. We present the theoretical basis for these methods and discuss results from applying them to publicly available data from Toronto wastewater treatment plants.

l'activité d'une maladie au sein des communautés. Un des avantages de ce cadre est sa capacité de distinguer les signaux viraux des eaux usées qui sont communs à l'ensemble des communautés de ceux qui sont spécifiques à chacune d'entre elles. Tous les signaux sont estimés simultanément avec une forme fonctionnelle de leurs dérivés, afin d'améliorer la précision des prévisions. Nous présentons le fondement théorique de ces méthodes et les résultats obtenus en les appliquant à des données publiques provenant de stations de traitement des eaux usées de Toronto.

[10:50-11:05]

Luke Hagar (University of Waterloo) **Nathaniel T. Stevens** (University of Waterloo)

Quantile Estimation for Sampling Distributions of Posterior Probabilities

Estimation des quantiles pour les distributions d'échantillonnage des probabilités a posteriori

To design Bayesian clinical studies, criteria for posterior-based operating characteristics – such as power and the type I error rate – are often carefully controlled to satisfy the reporting requirements of regulatory bodies. These posterior-based operating characteristics are typically assessed by estimating entire sampling distributions of posterior probabilities via simulation. We propose a scalable method to determine optimal sample sizes and decision criteria that maps posterior probabilities to low-dimensional conduits for the data. Our method leverages this mapping and large-sample theory to estimate quantiles of sampling distributions of posterior probabilities in a targeted manner. This targeted estimation approach prompts consistent sample size recommendations with fewer simulation repetitions than standard methods. We repurpose the posterior probabilities computed in that approach to efficiently investigate various sample sizes and decision criteria using contour plots.

Lors de la conception d'études cliniques bayésiennes, on contrôle souvent soigneusement les critères relatifs aux caractéristiques de fonctionnement basées sur l'analyse a posteriori (telles que la puissance et le taux d'erreur de type I) afin de satisfaire aux exigences des organismes de réglementation en matière de rapports. Ces caractéristiques d'opérationnalisation a posteriori sont généralement évaluées en estimant par simulation les distributions d'échantillonnage entières des probabilités a posteriori. Nous proposons une méthode évolutive pour déterminer les tailles d'échantillon et les critères de décision optimaux, qui associe les probabilités a posteriori à des conduits à petite dimension pour les données. Notre méthode s'appuie sur cette correspondance et sur la théorie des grands échantillons pour estimer de manière ciblée les quantiles des distributions d'échantillonnage des probabilités a posteriori. Cette approche d'estimation ciblée permet de recommander une taille d'échantillon cohérente avec moins de répétitions de simulations que les méthodes standard. Nous réutilisons les probabilités a posteriori calculées dans cette approche pour étudier efficacement diverses tailles d'échantillon et divers critères de décision à l'aide de graphiques en courbes de niveau.

[11:05-11:20]

Zijin Liu (University of Toronto) **Zhihui (Amy) Liu** (University Health Network) **Olli Saarela** (University of Toronto)

A Bayesian Joint Model for Mediation Analysis With Matrix-Valued Mediators

Modèle conjoint bayésien pour l'analyse de la médiation avec des médiateurs à valeur matricielle

Latent variable mediation models conceptualize mediators as dimension reductions of high-dimensional manifest (or indicator) variables. We propose a Bayesian joint mediation model for high-dimensional matrix-valued indicators of mediation, connecting latent features extracted from multilinear principal component analysis (MPCA) to exposure and outcome. For matrix-valued data, MPCA can reduce dimension more effec-

Les modèles de médiation à variables latentes considèrent les médiateurs comme des réductions de la dimensionnalité de variables manifestes (ou indicateurs) de haute dimension. Nous proposons un modèle bayésien de médiation conjointe pour les indicateurs de médiation à valeur matricielle de haute dimension, reliant les caractéristiques latentes extraites de l'analyse en composantes principales multilinéaire à l'exposition et au résultat. Pour les données à valeur matricielle, l'analyse en composantes

tively compared to conventional PCA. We also propose a quantity for identifying active indicators of mediation, mapping the mediated effects to the indicator variable matrix. Our proposed Bayesian estimation procedure produces probabilistic inferences on the mediation quantities of interest. Our work is motivated by, and illustrated through, a study of radiotherapy prescription dose effect for cancer on unplanned treatment interruptions. The radiation dose to organs at risk, summarized as a matrix of dose-volume histograms, mediates the prescription dose effect on the outcome.

principales multilinéaire peut réduire la dimension plus efficacement que l'analyse en composantes principales classique. Nous suggérons également une quantité pour identifier les indicateurs actifs de médiation en faisant correspondre les effets médiés à la matrice des variables indicatrices. La procédure d'estimation bayésienne que nous proposons fait des inférences probabilistes sur les quantités de médiation pertinentes. Nos travaux sont motivés et illustrés par une étude sur les effets de la dose de radiothérapie prescrite pour le cancer sur les interruptions de traitement non planifiées. La dose de rayonnement reçue par les organes à risque, exprimée sous la forme d'une matrice d'histogrammes dose-volume, est un médiateur de l'effet de la dose prescrite sur le résultat.

[11:20-11:35]

Shijie Min (University of Toronto)

A Copula-infused Graph Neural Network for Cell Type Classification in Single-cell RNA Sequencing Data

Réseau neuronal graphique basé sur des copules pour la classification des types de cellules dans des données de séquençage de l'ARN unicellulaire

Single-cell RNA sequencing (scRNA-seq) has advanced our understanding of cellular heterogeneity but the method of analyzing it challenged by the complexity in cell type classification. We propose a single-cell copula-based graph neural network model(scCopulaGNN) for the cell-type classification. The proposed method combines GCN' ability to capture representational information and the dependency modeling capabilities of copula functions. It applies the Gaussian copula model to capture the complex dependencies, integrates the copula model for cell-type classification. The performance of the model is evaluated against other five baseline models using five benchmark data sets. The scCopulaGNN demonstrates better performance as it outperforms baselines and achieves an accuracy of 95%. ROC-AUC and precision-recall AUC are measured at around 0.9 which shows it could effectively handle the variability in the data. The model could advance the understanding of cellular heterogeneity.

Le séquençage de l'ARN unicellulaire (scRNA-seq) nous a permis de mieux comprendre l'hétérogénéité cellulaire, mais la méthode pour l'analyser se complique en raison de la complexité de la classification des types de cellules. Nous proposons un modèle de réseau neuronal graphique unicellulaire basé sur des copules (scCopulaGNN) pour la classification des types de cellules. La méthode proposée combine la capacité d'un réseau convolutif graphique à saisir l'information représentationnelle et les capacités de modélisation de la dépendance des fonctions copules. Elle applique le modèle de copule gaussienne pour saisir les dépendances complexes et intègre le modèle de copule pour la classification des types de cellules. La performance du modèle est évaluée contre celle de cinq autres modèles de référence en utilisant cinq ensembles de données de référence. La performance du scCopulaGNN est meilleure, car elle surpasse celle des modèles de référence et son exactitude atteint 95 %. La mesure de l'aire sous la courbe ROC (AUC) et de l'AUC précision-rappel tourne autour de 0.9, ce qui indique que le modèle peut efficacement traiter la variabilité des données. Le modèle peut améliorer la compréhension de l'hétérogénéité cellulaire.

[11:35-11:50]

Amanda Qiu (University of Victoria) **Hong Wang** (SANOFI) **Luc Essermeant** (SANOFI) **Weiliang Qiu** (SANOFI) **Xuekui Zhang** (University of Victoria)

Novel Evaluation of Cell-Type Deconvolution Algorithms in Bulk RNAseq Data Analyses

Nouvelle évaluation d'algorithmes de déconvolution de type de cellule dans les analyses de grand nombre de données de séquençage de l'ARN

Estimation of cell type proportions from bulk RNA sequencing data is crucial for adjusting cellular het-

L'estimation des proportions d'un type de cellule à partir d'un grand nombre de données de séquençage de l'ARN (RNAseq) est

erogeneity in differential expression analysis. Many excellent methods, such as CIBERSORTx, MuSiC, and SCDC, have been developed. Benchmark studies have ranked these algorithms based on simulations with known cell-type proportions. However, in real-world bulk data, actual proportions are usually unknown, shifting focus towards result quality rather than method comparison. This underscores the absence of a method akin to goodness-of-fit for evaluating cell-type deconvolution without known proportions—a significant research gap that has not yet been addressed. We propose a novel indirect approach for evaluating cell-type deconvolution performance without known cell-type proportion information. Both simulation and real data analysis demonstrate this method's effectiveness in assessing the accuracy of estimated cell-type proportions, offering a valuable tool to researchers in the field.

cruciale pour ajuster l'hétérogénéité cellulaire dans les analyses d'expression différentielle. Plusieurs méthodes excellentes ont été conçues comme CIBERSORTx, MuSiC et SCDC. Les études de références ont classé ces algorithmes en se basant sur des simulations avec des proportions de types de cellule connues. Cependant, dans des données réelles, les proportions sont généralement inconnues, ce qui met l'accent sur la qualité des résultats plutôt que sur la comparaison des méthodes. Ceci met en évidence l'absence d'une méthode équivalente à l'adéquation pour évaluer la déconvolution de type de cellule sans connaître les proportions ; une lacune considérable de la recherche qui n'est pas encore résolue. Nous proposons une nouvelle approche indirecte pour évaluer la performance de déconvolution de type de cellule sans connaître les proportions de type de cellule. Des analyses faites sur des données réelles et simulées démontrent l'efficacité de cette méthode pour l'évaluation de la précision des proportions estimées de type de cellule. La méthode représente donc un outil pertinent pour les recherches dans le domaine.

Chair/Président: Tianyu Guan

Organizer/Responsable: Tianyu Guan

Room/Salle: A 1046

Date: Monday June 3 / lundi 3 juin

Time/Heure: 10:20-11:50

Abstract/Résumé

[10:20-10:50]

Shirley E. Mills (Carleton University)

Evaluating Player and Team Performance

Évaluer un joueur et la performance d'équipe

This talk presents statistical and machine learning methods to evaluate athletes and teams during individual matches and considers comparison across seasons. Application to basketball and/or hockey data is considered.

Cet exposé présente des méthodes statistiques et d'apprentissage automatique pour évaluer les athlètes et les équipes pendant des parties individuelles et étudie les comparaisons entre les saisons. Nous considérons leur application à des données sur le hockey et/ou le basketball.

[10:50-11:20]

Paramjit S. Gill (University of British Columbia Okanagan)

Tennis Analytics Based on Point-by-Point Data

Analyse du tennis basée sur des données point par point

A rally point in tennis may be scored by an ace serve, or a winning shot, or because of forced/unforced errors by the opponent player. We fit probabilistic models for outcomes of rallies using data from a large number of matches. We fit separate models for rallies started with the first or the second serve. Our model includes parameters representing the strength of individual players. Using the estimated parameters, we can simulate matches between any two players. We can simulate matches under some imaginary situations, such as a match between two players from different era. In tennis the serving player has very high chance of winning a rally point. We may ask what would happen if only one serve were allowed. If the second serve is not available, the players will be cautious not to default the serve and will serve as if they were delivering the second serve. Based on a large number of simulations, we can provide some insight to what would happen if the second serve were outlawed.

Au tennis, un point peut être marqué par un as, un coup gagnant ou par des erreurs forcées ou non forcées de l'adversaire. Nous ajustons des modèles probabilistes aux résultats d'échanges en nous appuyant sur des données provenant d'un grand nombre de matchs. Nous ajustons des modèles distincts pour les échanges commencés sur le premier ou le second service. Notre modèle comprend des paramètres représentant la force des joueurs individuels. En utilisant les paramètres estimés, nous pouvons simuler des matchs entre deux joueurs quelconques. Nous pouvons simuler des matchs dans certaines situations imaginaires, comme un match entre deux joueurs d'époques différentes. Au tennis, le joueur au service a de grandes chances de remporter le point. Nous pouvons nous demander ce qui se passerait si un seul service était autorisé. Si le second service n'est pas disponible, les joueurs feront attention à ne pas manquer leur service et serviront comme s'ils étaient au second service. Sur la base d'un grand nombre de simulations, nous pouvons donner un aperçu de ce qui se passerait si le second service était interdit.

[11:20-11:50]

Tim B. Swartz (Simon Fraser University)

Causal Inference Problems in Soccer using Tracking Data
Problèmes d'inférence causale dans le soccer et données de suivi

A frequent impediment to applied causal analysis is the identification and quantification of confounding variables. With the advent of tracking data in sport, there is often a realistic chance of dealing with the confounding variable problem. In this presentation, we consider three questions involving soccer tactics that are each approached using causal methods (a) what is the benefit of crossing the ball? (b) what is the benefit of playing with pace? and (c) what is the benefit associated with throw-in decisions? The problems each have a common structure, and we provide a template for approaching such problems. The differences between the problems lie in the nature of the independent variables, the dependent variables and the confounding variables.

L'identification et la quantification de variables confusionnelles constituent un obstacle fréquent à l'analyse causale appliquée. Mais l'avènement des données de suivi dans le sport offre une chance réaliste de traiter le problème des variables confusionnelles. Dans cette présentation, nous examinons trois questions relatives aux tactiques de soccer à l'aide de méthodes causales (a) Quel est l'avantage d'une frappe croisée ? b) Quel est l'avantage de la rapidité de jeu ? et c) Quel est l'avantage associé aux décisions de remise en jeu ? Les problèmes ont tous une structure commune et nous fournissons un modèle pour les aborder. Les différences entre les problèmes résident dans la nature des variables indépendantes, des variables dépendantes et des variables confusionnelles.

Chair/Président: Ana Carolina da Cruz

Room/Salle: A 2065

Date: Monday June 3 / lundi 3 juin

Time/Heure: 10:20-11:50

Abstract/Résumé

[10:20-10:35]

Malcolm Risk (University of Michigan) **Xu Shi** (University of Michigan) **Lili Zhao** (University of Michigan)

Distributed Kaplan-Meier Curves via the Influence Function

Courbes de Kaplan-Meier distribuées avec la fonction d'influence

Large, multi-center observational data are required to study rare events and exposures. However, sharing sensitive individual-level survival data such as event times and patient characteristics can require a lengthy approval process. Existing work on distributed survival analysis focuses on parametric and semi-parametric models rather than non-parametric Kaplan-Meier (KM) curves. We develop a privacy-preserving sequential distributed method for approximating KM curves by splines updated via the influence function, with confounder adjustment via inverse probability weighting and inference using the weighted log-rank test. Our method requires sharing only summary-level data (spline coefficients and knot locations), and we show equivalent inferential performance to KM analysis with pooled data in simulations. We use our method to examine incidence of blood clots after COVID-19 infection and COVID-19 vaccination using electronic health record data at Corewell Health and Michigan Medicine.

L'étude d'événements et d'expositions rares nécessite des données d'observation multicentriques de grande taille. Cependant, le processus d'approbation des données de survie sensibles au niveau individuel (temps d'événements et caractéristiques des patients) peut être long. Les travaux actuels sur l'analyse de survie distribuée portent sur des modèles paramétriques et semi-paramétriques plutôt que sur des courbes de Kaplan-Meier non paramétriques. Nous élaborons une méthode séquentielle et distribuée préservant la confidentialité pour l'approximation des courbes de Kaplan-Meier par des splines mises à jour avec la fonction d'influence. Nous adaptons les facteurs de confusion au moyen d'une pondération des probabilités inverses, et l'inférence se fait au moyen du test logarithmique par rangs pondéré. Notre méthode ne nécessite que l'échange de données sommaires (coefficients de spline et emplacements des nœuds). Nous présentons des résultats inférentiels équivalents à ceux de l'analyse de Kaplan-Meier avec des données de simulation regroupées. Nous utilisons notre méthode pour analyser l'incidence des caillots sanguins après l'infection par la COVID-19 et après la vaccination contre la COVID-19 avec les données des dossiers médicaux électroniques de Corewell Health et Michigan Medicine.

[10:35-10:50]

Shiyao Ying (University of Toronto) **Yun-Hee Choi** (Western University, London, Ontario) **Laurent Briollais** (Lunenfeld-Tanenbaum Research Institute, Mount Sinai Hospital, Toronto, Ontario)

Correlated Frailty Models with Kinship for Analysis of Time-to-Event Outcomes within Families

Modèles de fragilité corrélés avec la parenté pour l'analyse des résultats temporels au sein des familles

The study of family traits and their impact on health outcomes has been a popular research area, particularly in the context of understanding how genetic similarities affect the risk of diseases within families. The kinship matrix measures the pairwise relatedness of individuals

L'étude des caractéristiques familiales et de leur impact sur les résultats de santé est un domaine de recherche populaire, en particulier s'agissant de comprendre comment les similitudes génétiques affectent le risque de maladies au sein des familles. La matrice de parenté mesure la parenté par paire des individus

within the same family. Incorporating the kinship matrix into survival models for family data is difficult but this can be done through correlated frailty models. The popular method Coxme is based on a penalized partial likelihood approach but has never been tested for the analysis of complex pedigrees. Alternatively, a full likelihood approach can be proposed but is generally difficult to implement due to computational challenges. We compare these approaches using extensive simulations with or without ascertainment of the families through affected probands and also in an application to a series of families harboring BRCA1 mutations.

au sein d'une même famille. Il est difficile d'intégrer cette matrice dans les modèles de survie pour les données familiales, mais il est possible de le faire au moyen de modèles de fragilité corrélés. La méthode populaire Coxme est basée sur une approche de vraisemblance partielle pénalisée mais n'a jamais été testée pour l'analyse de pedigrés complexes. Une approche de vraisemblance totale peut également être proposée, mais elle est généralement difficile à mettre en œuvre en raison de problèmes de calcul. Nous comparons ces approches à l'aide de simulations approfondies avec ou sans vérification des familles par le biais de probants atteints, ainsi que dans le cadre d'une application à une série de familles porteuses de mutations BRCA1.

[10:50-11:05]

Amin Abed (University of Manitoba) **Mahmoud Torabi** (University of Manitoba) **Zeinab Mashreghi** (University of Winnipeg)

Modeling of Infectious Disease with Reinfection: Tuberculosis Transmission in Manitoba

Modélisation des maladies infectieuses avec réinfection : transmission de la tuberculose au Manitoba

The basic homogeneous SEIR (Susceptible-Exposed-Infected-Removed) model is commonly used for analyzing infectious diseases. The SEIR model lacks individual-level information like location and distance between susceptible and infected individuals. To address this limitation, a Geographically Dependent Individual-Level Model (GD-ILM) within an SEIR framework was previously developed. We expand the SEIR GD-ILM to accommodate infectious diseases involving loss of immunity, like Tuberculosis, resulting in SEIRS (Susceptible-Exposed-Infected-Removed-Susceptible) GD-ILM for when both regional and individual-level spatial data are available. To overcome computational complexity, we introduce a new approach based on the Monte Carlo Expectation Conditional Maximization algorithm for efficient parameter estimation. Focusing on Manitoba's health regions, we analyze individual-level Tuberculosis data (2011-2018) to predict infection rates over time and identify high-risk geographical areas.

Le modèle homogène de base SEIR (Susceptible-Exposé-Infecté-Retiré) est couramment utilisé pour analyser les maladies infectieuses. Or le modèle SEIR manque d'informations au niveau individuel, telles que la localisation et la distance entre les individus sensibles et infectés. Pour remédier à cette lacune, un modèle individuel dépendant de la géographie (GD-ILM) a été développé dans le contexte du modèle SEIR. Nous étendons ce modèle pour tenir compte des maladies infectieuses impliquant une perte d'immunité, comme la tuberculose, ce qui donne le GD-ILM SEIRS (Susceptible-Exposé-Infecté- Retiré-Susceptible) pour les cas où des données spatiales régionales et individuelles sont disponibles. Pour surmonter la complexité des calculs, nous introduisons une nouvelle approche basée sur l'algorithme de maximisation conditionnelle des espérances de Monte Carlo pour une estimation efficace des paramètres. En nous concentrant sur les régions sanitaires du Manitoba, nous analysons les données sur la tuberculose au niveau individuel (2011-2018) pour prédire les taux d'infection au fil du temps et identifier les zones géographiques à haut risque.

[11:05-11:20]

Mei Dong (University of Toronto) **Linbo Wang** (University of Toronto) **Wei Xu** (University of Toronto)

Robust Estimator for Average Treatment Effect with Continuous Instrumental Variables

Estimateur robuste de l'effet moyen du traitement avec des variables instrumentales continues

Instrumental variable (IV) approach is popular in estimating the average treatment effect (ATE) in the presence of unmeasured confounders. While methods for estimating ATE with binary IVs are well-established, challenges arise when the IV is continuous, as is of-

L'approche par variable instrumentale (IV) est très utilisée pour estimer l'effet moyen du traitement (EMT) en présence de facteurs de confusion non mesurés. Si les méthodes d'estimation de l'EMT avec des IV binaires sont bien établies, des problèmes se posent lorsque l'IV est continue, comme c'est souvent le cas en

ten the case in genetic epidemiology with polygenic risk scores. Current approaches dealing with continuous IV rely on structural equation modeling, yielding biased estimates when the outcomes are binary. In this work, we developed four novel estimators for ATE using continuous IV under the potential outcome framework. Of these, three estimators are consistent under different observed data models and one estimator is triply robust, that is, consistent in the union of these observed data. Simulation results showed that our proposed estimator is unbiased and robust to model misspecification. We further illustrate our approaches to estimate the causal effect of obesity in lung cancer patients' two-year mortality rate.

épidémiologie génétique avec des scores de risque polygéniques. Les approches actuelles concernant les IV continues reposent sur la modélisation par équations structurelles, ce qui donne des estimations biaisées lorsque les résultats sont binaires. Dans ce travail, nous avons développé quatre nouveaux estimateurs pour l'EMT en utilisant une IV continue dans le cadre des résultats potentiels. Parmi ceux-ci, trois estimateurs sont convergents sous différents modèles de données observées et un estimateur est triplement robuste, c'est-à-dire convergent dans l'union de ces données observées. Les résultats de simulations ont montré que l'estimateur proposé est sans biais et robuste à la mauvaise spécification du modèle. Nous illustrons ensuite nos approches pour estimer l'effet causal de l'obésité sur le taux de mortalité à deux ans des patients atteints de cancer du poumon.

[11:20-11:35]

Fatema Tuj Johara (University of Toronto Dalla Lana School of Public Health) **Eleanor M. Pullenayegum** (Hospital for Sick Children)

Methods of Quantifying Within Person Variability for Longitudinal Data With Irregular Observation

Méthodes de quantification de la variabilité intra-personnelle pour données longitudinales avec observations irrégulières

Variability in longitudinal outcomes is often perceived as a nuisance parameter in statistical models and is not usually estimated. However, within-subject variability may be informative. For example, pediatric systemic lupus erythematosus (SLE) is a chronic disease commonly involving kidney inflammation and is characterized by periods of flare and periods of disease quiescence. This makes modeling variability of kidney function clinically interesting. In this project, we aim to model outcome variability in the presence of irregular observation. We propose several estimating functions, each of which comprises a marginal mean model and variability models. The variability models to consider are sample variance and median absolute deviation (MAD) about mean. We obtain closed form expression of sandwich variance estimator. A simulation study compared variability estimators for estimating equations and confirmed a good performance in terms of their biases, variances and 95% coverages.

La variabilité des résultats longitudinaux est souvent perçue comme un paramètre de nuisance dans les modèles statistiques et n'est généralement pas estimée. Cependant, la variabilité intra-personnelle peut être informative. Par exemple, le lupus érythémateux systémique pédiatrique est une maladie chronique qui implique souvent une inflammation des reins et qui se caractérise par des périodes de poussée et des périodes de quiescence de la maladie. La modélisation de la variabilité de la fonction rénale est donc cliniquement intéressante. Dans ce projet, nous visons à modéliser la variabilité des résultats en présence d'observations irrégulières. Nous proposons plusieurs fonctions d'estimation, chacune comprenant un modèle de moyenne marginale et des modèles de variabilité. Les modèles de variabilité à considérer sont la variance de l'échantillon et l'écart absolu médian par rapport à la moyenne. Nous obtenons une expression de forme fermée de l'estimateur sandwich de la variance. Nous menons une étude de simulation pour comparer les estimateurs de variabilité et confirmons une bonne performance en termes de biais, de variances et de couvertures à 95 %.

[11:35-11:50]

Wensha Zhang (Dalhousie University) **Toby J. Kenney** (Dalhousie University) **Lam Ho** (Dalhousie University)

Detection of Evolutionary Shifts in Variance under an Ornstein–Uhlenbeck Model

Détection des changements évolutifs de la variance dans le cadre d'un modèle d'Ornstein-Uhlenbeck

Abrupt environmental changes can lead to evolutionary shifts in both optimal value and variance of descendants in trait evolution. Current methods mainly focus on detecting shifts in optimal values, with less attention

De brusques changements environnementaux peuvent entraîner des modifications de la valeur optimale et de la variance des descendants dans l'évolution des caractères. Les méthodes actuelles sont principalement axées sur la détection des changements dans

to variance. We use a multi-optima and multi-variance OU process model to describe the trait evolution process with shifts in both optimal value and variance. Furthermore, we propose a new method to detect the shifts in both variance and optimal values based on minimizing the loss function with L1 penalty; and implement our method in a new R package, ShiVa. We conduct simulations to compare our method with the two methods considering only shifts in optimal values (l1ou; PhylogeneticEM). Our method shows better predictive ability and includes far fewer false positive shifts in optimal value when shifts in variance exist. We applied our method to the cordylid data. ShiVa outperformed l1ou and phyloEM, exhibiting the highest log-likelihood and lowest BIC.

les valeurs optimales et portent moins d'attention à la variance. Nous utilisons un modèle de processus d'Ornstein-Uhlenbeck à optima et variance multiples pour décrire le processus d'évolution des traits avec des changements à la fois de la valeur optimale et de la variance. De plus, nous proposons une nouvelle méthode pour détecter les changements de variance et de valeurs optimales en minimisant la fonction de perte avec une pénalité L1, puis nous implémentons notre méthode dans un nouveau progiciel R, ShiVa. Nous effectuons des simulations pour comparer notre méthode avec les deux méthodes qui ne prennent en compte que les déplacements des valeurs optimales (l1ou et PhylogeneticEM). Notre méthode, appliquée aux données sur les cordylidés, présente une meilleure capacité de prédiction et inclut beaucoup moins de faux positifs dans la valeur optimale en cas de changements de la variance. ShiVa a surpassé l1ou et phyloEM, affichant le logarithme du rapport de vraisemblance le plus élevé et le critère d'information bayésien le plus bas.

Chair/Président: Armin Hatefi

Room/Salle: C 2033

Date: Monday June 3 / lundi 3 juin

Time/Heure: 10:20-11:50

Abstract/Résumé

[10:20-10:35]

Golshid Aflaki (HEC Montréal) **Jean-François Plante** (HEC Montreal) **Juliana Schulz** (HEC Montreal)

A New Bivariate Zero-Inflated Poisson Model

Un nouveau modèle Poisson bivarié à excès de zéros

There are numerous applications which involve modeling multi-dimensional count data. When such data exhibit an excess of zeros, common count models are no longer adequate. In this work, we propose a new bivariate zero-inflated Poisson model appropriate for modeling multivariate counts with a surplus of zeros. The proposed model construction is based on a mixture model approach involving a common mass at zero along with a Poisson random pair. The latter stems from a bivariate Poisson model wherein a positive correlation is induced via a comonotonic shock. The properties of the proposed model are detailed, and several estimation methods are explored. Comprehensive simulations as well as a real data illustration demonstrate the practical applications of the proposed model.

Il existe plusieurs applications nécessitant la modélisation multidimensionnelle de données de dénombrement. Lorsque de telles données comportent un excès de zéros, les modèles de dénombrement habituels peuvent s'avérer inadéquats. L'auteure propose un nouveau modèle Poisson bivarié avec surabondance de zéros convenant à de telles données. La construction du modèle proposé exploite une approche par mélange comportant une masse conjointe à zéro et une paire de variables Poisson à laquelle une dépendance positive est induite par un choc comonotone. L'auteure décrit les propriétés du modèle proposé et explore les différentes méthodes d'estimation de ses paramètres. Elle présente les résultats d'une importante simulation de Monte Carlo et illustre les capacités du modèle par une analyse de données réelles.

[10:35-10:50]

Pranath Pussella (Brock University) **Tianyu Guan** (Brock University) **Robert Nguyen** (University of New South Wales)

Simulation for Cricket: A Machine Learning Approach

Simulation pour le cricket : une approche d'apprentissage automatique

Cricket is the second most popular sport in the world with a significant presence in Commonwealth countries. Despite its popularity, cricket is underrepresented in the literature, especially in the domain of simulation. Simulation in cricket is challenging because of its complexity, dynamic nature, and data scarcity. In this research, we develop a simulation mechanism for cricket using machine learning techniques. The construction of the simulator is based on the availability of a detailed dataset from Cricket Australia. We employ machine learning to predict the outcome of a "delivery", the core element of gameplay, which can further be utilized for scorecard generation and match simulations. Our simulator's po-

Le cricket est le deuxième sport le plus populaire au monde et a une présence considérable dans les pays du Commonwealth. Malgré sa popularité, le cricket reste sous représenté dans la littérature, tout particulièrement dans le domaine de la simulation. La simulation au cricket est ardue en raison de la complexité du sport, de sa nature dynamique et du manque de données. Dans cette recherche, nous développons un mécanisme de simulation pour le cricket en utilisant des techniques d'apprentissage automatique. La construction du simulateur est basée sur l'apport d'un ensemble de données détaillées provenant de Cricket Australia. Nous employons l'apprentissage automatique pour prédire le résultat d'un « lancer », l'élément fondamental du jeu, qui peut de plus servir à la génération de cartes de pointage et de simulations de

tential is demonstrated by employing it to determine the optimal batting position of a given batter in a team in Twenty20 cricket. Additionally, we develop an interactive web platform to enable direct interaction with the simulator and the tool for optimizing batting positions.

parties. Nous démontrons le potentiel de notre simulateur en l'utilisant pour déterminant la position de frappe optimale d'un batteur donné dans une équipe pour une partie de cricket de format « Twenty20 ». De plus, nous développons une plateforme web interactive pour permettre une interaction directe avec le simulateur et les outils pour optimiser les positions de frappe.

[10:50-11:05]

David Awosoga (University of Waterloo)

Investigating Player Contribution in Volleyball Using Bayesian Spatiotemporal Data Analysis

Enquête sur la contribution des joueurs au volley-ball à l'aide d'une analyse de données spatio-temporelle bayésienne

Understanding player contributions is an important component of lineup construction, advance scouting, and performance evaluation. However, traditional methods utilize oversimplified percentages that fail to acknowledge latent variables and situational nuance. These shortcomings are addressed in this work via a Bayesian approach that incorporates player roles, lineup matchups, and additional context-specific information into its estimates. A Markov chain with inputs derived from multi-year spatiotemporal event data is used to provide continuously updating point scoring probabilities during a rally. Changes in this probability after each ball contact are used to divide credit between players. The results demonstrate an increased ability to differentiate between players and measure contribution in a stable manner over time. The resultant model outputs are illustrated in multiple case studies, with direct application to volleyball coaches, players, and the community at large.

Une bonne compréhension de la contribution des joueurs est une composante importante de l'alignement, du recrutement avancé et de l'évaluation de la performance. Les méthodes traditionnelles recourent toutefois à des pourcentages exagérément simplifiés qui échouent à reconnaitre les variables latentes et la nuance situationnelle. Cette étude porte sur ces lacunes en adoptant une approche bayésienne qui incorpore dans son estimation les rôles des joueurs, les confrontations et d'autres informations spécifiques au contexte. Une chaîne de Markov avec facteurs dérivés de données spatio-temporelles multiannuelles d'événements est utilisée pour fournir des probabilités d'attribution de points constamment mises à jour pendant une rencontre. Les changements à cette probabilité après chaque contact avec la balle sont utilisés pour répartir le crédit entre les joueurs. Les résultats indiquent une capacité accrue de différencier les joueurs et de mesurer la contribution de chacun de façon stable avec le temps. Les résultantes du modèle sont illustrées par plusieurs études de cas, avec une application directe aux entraîneurs, joueurs et la communauté en général de volley-ball.

[11:05-11:20]

Yuying Huang (University of Waterloo) **Samuel Wong** (University of Waterloo)

Sequential Design Strategy for Mean Response Surface Modeling of Expensive Stochastic Simulation with Heterogeneous Noise via Bayesian Framework

Stratégie de conception séquentielle pour la modélisation de surface de réponse moyenne de simulation stochastique coûteuse avec bruit hétérogène dans le cadre bayésien

Gaussian Processes (GP) is an effective surrogate model for globally accurate emulation of noisy computer simulations. With the goal of building surrogate model for expensive simulations with heterogeneous noise, we utilize a Bayesian framework paired with GP, from which a novel empirical integrated mean squared error-based sequential design strategy is proposed to approximate the mean response surface with a small fixed simulation budget. Through different synthetic examples, we show that the proposed strategy has the potential to achieve high predictive accuracy under a small budget compared

Les processus gaussiens (GP) sont des modèles de substitution efficaces pour l'émulation globalement précise de simulations bruitées. Avec l'objectif de construire un modèle de substitution pour les simulations coûteuses avec bruit hétérogène, nous utilisons un cadre bayésien couplé avec des GP à partir desquels nous proposons une nouvelle stratégie de conception séquentielle basée sur l'erreur quadratique moyenne intégrée empirique afin d'estimer la surface de réponse moyenne avec un petit budget de simulation fixe. À travers différents exemples synthétiques, nous prouvons que la stratégie proposée a le potentiel d'atteindre une précision prédictive élevée avec un petit budget par rapport aux

to existing state-of-the-art methods. We also demonstrate the performance of our strategy on a real-data simulation of finding the reliable region in podium building seismic design.

méthodes actuelles de fine pointe. Nous démontrons aussi la performance de notre stratégie sur une simulation avec des données réelles cherchant à trouver une région fiable pour la construction de podium selon un modèle sismique.

Strategies in Teaching Statistics and Evaluating Statistical Knowledge
Stratégies d'enseignement de la statistique et l'évaluation des connaissances statistiques

Chair/Président: Mark Reesor

Room/Salle: C 4036

Date: Monday June 3 / lundi 3 juin

Time/Heure: 10:20-11:50

Abstract/Résumé

[10:20-10:35]

Scott Andrew Robison (University of Calgary)

Interesting Courses, Need to be Interesting

Les cours intéressants doivent susciter l'intérêt

In the book After Virtue, the philosopher Alasdair MacIntyre distinguishes between external goods ("means to the end," the consequence/outcome of rote-practice, the result of circumstance and not inherently because of the activity) and internal goods ("the means/paths they take to end to the end," or intrinsic part of that practice; such goods, cannot be separated from activity). Grades are external goods. This means that, if a student can get away with "cheating" or another method to obtain a grade, without truly understanding the material, there is no reason not to do so! However, course outcomes and class engagement are internal goods, such understanding is the true purpose of education. Please consider, with me, how incorporating: reflective journaling, lecturing style, subject-matter history, replaceable grading, and providing the motivations for our material helps to deepen our student's learning and improve their opinions of our courses.

Dans le livre « After Virtue », le philosophe Alasdair MacIntyre distingue le bien externe (« la fin justifie les moyens », le résultat d'une pratique répétée, le produit de circonstances plutôt que d'une action en soi uniquement) et le bien interne (« les moyens justifient la fin », ou la partie intrinsèque d'une pratique ; ces biens sont indissociables de l'action). Les notes sont des biens externes. Par conséquent, si un étudiant peut « tricher » ou utiliser une autre méthode pour obtenir une note, sans vraiment comprendre le contenu du cours, il aurait toutes les raisons de le faire ! Cependant, les résultats de cours et l'investissement dans un cours sont des biens internes ; une telle compréhension est le véritable objectif de l'éducation. Veuillez examiner avec moi de quelles façons nous pouvons intégrer le journal réflexif, le style d'enseignement, l'histoire du sujet abordé et la notation alternative permet d'intéresser les étudiants au sujet de matière et davantage approfondir leurs apprentissages et leur donner une meilleure opinion des cours.

[10:35-10:50]

Tharshanna Nadarajah (McGill University)

Enhance Student Learning by Reviewing Previous Learning Regularly

Amélioration de l'apprentissage des étudiants par la révision régulière des cours précédents

Students' retention of knowledge can be strengthened by reviewing previous learning before each lesson. Keeping a daily review schedule is one of the most important strategies to improve learning. Consequently, they will gain a deeper understanding of syllabus material, make connections between topics, and improve their critical thinking skills. While it's a good idea to have a quick recap, how about giving students the opportunity to prepare and submit their own reviews? Students are asked to submit a summary of each week's

Les étudiants peuvent mieux mémoriser ce qu'ils ont appris en révisant avant chaque cours. La tenue d'un calendrier de révision quotidien est l'une des stratégies les plus importantes pour améliorer l'apprentissage. Les étudiants peuvent ainsi mieux comprendre la matière du programme, établir des liens entre les sujets et améliorer leurs capacités de réflexion critique. Même si un bref résumé est une bonne idée, pourquoi ne pas donner aux étudiants la possibilité de préparer et de soumettre leurs propres révisions ? On leur demande en effet de présenter chaque semaine un résumé des connaissances acquises, dans le cadre de

learning as part of the course evaluation. It is a weekly course assignment followed by a class pop quiz. Using a survey, we collected data on students' attitudes and practices related to reviewing previous learning regularly. This data was linked to course outcomes. According to the results, regular review of previous learning enhances student performance and is thought to be a helpful study technique by students.

l'évaluation du cours. Il s'agit d'un projet hebdomadaire suivi d'un questionnaire en classe. Nous avons recueilli, à l'aide d'une enquête, des données sur les attitudes et les pratiques des étudiants en ce qui concerne la révision régulière des cours précédents. Nous avons établi un lien entre ces données et les résultats du cours. D'après les résultats, la révision régulière des cours améliore les résultats des étudiants et ces derniers considèrent qu'il s'agit d'une méthode d'étude utile.

[10:50-11:05]

Danika M. Lipman (University of Calgary) **Scott Andrew Robison** (University of Calgary)

Creating a Positive Class Environment Through Assessment

Créer un environnement de classe positif au moyen de l'évaluation

Fostering engagement in a mathematical statistics class is a challenge. In our exploration of strategies to enhance participation and alleviate exam stress, we present three approaches to student assessment. 1) Post-test replacement opportunities, where collaborative problem-solving allows students to replace a test grade with a potential 70% credit. 2) Pre-test replacement opportunities enabling students to collectively tackle practice tests for a chance at a 70% grade replacement. 3) Low-stakes assessment through take-home pre-quizzes and quizzes with low weighting to encourage consistent learning without the burden of high-stakes exams. Recognizing that exam stress can hinder students from performing their best, the first two methods provide avenues for learning from mistakes. The third method promotes material mastery through low-pressure assessment opportunities. Our observations indicate increased class engagement contributing to an overall positive classroom atmosphere.

Il est ardu de favoriser l'engagement dans un cours de mathématique statistique. Nous avons exploré des stratégies afin d'améliorer la participation et réduire le stress lié aux examens, et présentons maintenant trois approches pour évaluer les étudiants. 1. Occasions de réévaluation après-test, durant lesquelles une résolution de problème collaborative permet aux étudiants de remplacer une note d'examen avec un potentiel de crédit de 70 %. 2. Occasions de réévaluation avant test permettant aux étudiants d'effectuer collectivement un examen de pratique pour avoir une chance de remplacement de note de 70 %. 3. Évaluations comptant pour peu sous la forme d'exercices (à la maison ou en classe) qui encouragent un apprentissage constant sans la pression de quelques examens comptant pour beaucoup. En reconnaissant que le stress lié aux examens peut nuire à la performance des étudiants, les deux premières méthodes leur procurent des occasions pour apprendre de leurs erreurs. La troisième méthode favorise la maîtrise du contenu de cours au moyen d'évaluations peu stressantes. Nos observations indiquent un engagement supérieur en classe contribuant à une atmosphère de classe généralement positive.

[11:05-11:20]

Lengyi Spectrum Han (The University of British Columbia)

Using WipeBooks to Increase Student Engagement

L'emploi de WipeBooks pour augmenter l'engagement étudiant

Students seem to become easily distracted by their own digital devices, and their eyes glaze over when complex statistical concepts and derivations are presented as a monologue in currently popular formats, such as lecture slides. When the students are able to discuss concepts with each other and to write things down in their own way, they become owners of the subject matter. I have begun experimenting with an alternative strategy, allowing the students to work in groups of three in the classroom, deriving expressions and learning the concepts through chatting with each other. Wipebooks, which are

Les étudiants semblent facilement être distraits par leur propre appareil numérique et ils ne portent plus attention lorsqu'on présente des dérivations et des concepts statistiques complexes par un monologue souvent accompagné des diapositives. Par contre, lorsque les étudiants peuvent discuter de concepts entre eux et prendre des notes à leur façon, ils s'approprient le contenu enseigné. J'ai commencé une expérience avec une stratégie différente qui permet aux étudiants de travailler en groupe de trois en classe, puis de dériver et apprendre les concepts par des discussions entre eux. WipeBooks, une version portable des tableaux blancs, représente un outil pratique pour les étudiants leur permettant d'écrire leurs

portable versions of whiteboards, are a convenient tool for the students to write their ideas and solutions on, as a group. Students leave the classroom with an air of satisfaction. In this presentation, I will demonstrate the use of Wipebooks for teaching concepts that come up in a regression course.

idées et solutions en groupe. Les étudiants sortent de la classe avec un air satisfait. Dans cette présentation, je ferai la démonstration de WipeBooks pour l'enseignement de concept présent dans un cours de régression.

[11:20-11:35]

Shahriar Shams (University of Toronto Scarborough) **Sotirios Damouras** (University of Toronto Scarborough)

Implementing Computer-Based Assessments in Statistics Courses.

Implémentation d'évaluations informatisées dans les cours de statistique

Statistics and Data Science courses typically have a large and growing computational component, making extensive use of data and statistical software like R or Python. Nevertheless, summative examinations in such courses do not always test statistical thinking and data analytic skills in an integrated and authentic manner. In this talk, we will share some initial observations from introducing computer-based exams in a first-year Introductory Data Science course and a fourth-year Time Series course. Students in these courses take midterm and final exams in a computer lab, where they are required to manipulate, explore, and analyze data interactively to answer specific or open-ended questions under realistic conditions. We will describe practical considerations for offering such exams and share experiences and student feedback.

Les cours de statistique et de science des données ont habituellement une composante computationnelle importante et croissante qui comporte une forte utilisation de données et de logiciels statistiques comme R ou Python. Malgré tout, les examens sommatifs dans ces cours n'évaluent pas toujours de manière intégrée et authentique la pensée statistique et les compétences en analyse des données. Dans le cadre de notre exposé, nous partageons quelques observations initiales sur l'intégration d'examens informatisés dans le cours de première année d'introduction à la science des données et à un cours de quatrième année sur les séries chronologiques. Les étudiants dans ces cours ont passé des examens à mi-session et finaux dans un laboratoire d'informatique où on leur a demandé de manipuler, explorer et analyser des données de façon interactive afin de répondre à des questions spécifiques ou ouvertes, sous des conditions réalistes. Nous élaborons aussi sur les considérations pratiques pour offrir de tels examens et partageons des expériences et les commentaires des étudiants.

[11:35-11:50]

Chelsea Uggenti (University of Waterloo) **Divya Lala** (University of Waterloo)

Reflecting on the First Year of our Teaching Assistant Program

Réflexion sur la première année de notre programme d'assistants d'enseignement

Most universities have teaching centers that offer programs geared towards the training and development of Teaching Assistants (TAs). Department or discipline specific TA programs, especially in statistics and actuarial science, tend to be scarce and/or informal. In 2023 the University of Waterloo's Department of Statistics and Actuarial Science created the TA Program which encompasses all aspects related to graduate teaching assistantships and offers the Foundations in University Teaching in Statistics and Actuarial Science certificate training program. Our program works in collaboration with the university's Centre for Teaching Excellence to train incoming and current graduate TAs by providing sequential levels of training with a unique focus on discipline-specific courses and material. Our goal

La plupart des universités disposent de centres d'enseignement qui proposent des programmes axés sur la formation et le développement des assistants d'enseignement. Les programmes d'assistanat spécifiques à un département ou à une discipline, en particulier en statistique et en sciences actuarielles, ont tendance à être rares et/ou informels. En 2023, le Département de statistique et de science actuarielle de l'Université de Waterloo a créé un programme d'assistanat qui englobe tous les aspects liés aux postes d'assistants d'enseignement pour les étudiants des cycles supérieurs et propose un programme de formation menant à un certificat en Fondements de l'enseignement universitaire en statistique et en science actuarielle. Notre programme travaille en collaboration avec le Centre d'excellence en enseignement de l'université pour former les assistants à l'enseignement, qu'ils soient nouveaux ou actuels, en proposant des niveaux

is to prepare graduate students with the skills to confidently and successfully perform TA duties including proctoring, grading, facilitating tutorials, creating and delivering material for an in-class lecture, and more. In this talk, we will share details about the creation and structure of our TA Program, highlight key observations gleaned from its first year of operation, explore future goals for the program, and discuss potential implementation at other universities/departments.

de formation séquentiels qui mettent l'accent sur les cours et le matériel spécifiques à la discipline. Notre objectif est de préparer les étudiants des cycles supérieurs à acquérir les compétences nécessaires pour remplir avec confiance et succès les fonctions d'assistant d'enseignement, y compris la surveillance, la correction, l'animation de tutoriels, la création et la présentation de matériels pour un cours en classe, et plus encore. Dans cet exposé, nous partagerons des détails sur la création et la structure de notre programme d'assistanat, nous mettrons en évidence les observations clés glanées au cours de sa première année de fonctionnement, nous explorerons les objectifs futurs du programme et nous discuterons de sa mise en œuvre potentielle dans d'autres universités/départements.

Chair/Président: Joel A. Dubin

Room/Salle: C 3053

Date: Monday June 3 / lundi 3 juin

Time/Heure: 10:20-11:50

Abstract/Résumé

[10:20-10:35]

Kyu Min Shim (University of Waterloo)

Variance Reduction with Model-Based Counterfactual Estimation

Réduction de la variance par estimation contrefactuelle basée sur un modèle

Reducing the variance of the average treatment effect estimator is a critical problem in the context of online controlled experiments. Recent developments in variance reduction utilize pre-experiment data to achieve significant variance reduction under the assumption that pre-experiment and in-experiment data are highly correlated. However, in settings such as e-commerce and social media where trends in data may change rapidly, the validity of such an assumption may be questionable. This work addresses this challenge with a two-stage modeling framework that exploits relationships between covariates and the outcome in both pre-experiment and in-experiment data. Inference is made by estimating the counterfactual outcome of each unit and performing a pairwise comparison. This method of inference is proven to be asymptotically unbiased, with an asymptotic variance that scales with the model's predictive accuracy though is never larger than that of the naive difference in means estimator.

Réduire la variance de l'estimateur de l'effet moyen du traitement est un problème critique dans le contexte des expériences contrôlées en ligne. Les développements récents en la matière utilisent des données pré-expérimentales pour parvenir à une réduction significative de la variance, en partant du principe que les données pré-expérimentales et expérimentales sont fortement corrélées. Cependant, dans des contextes tels que le commerce électronique et les médias sociaux, où les tendances des données peuvent changer rapidement, cette hypothèse n'est pas toujours valable. Nous relevons ce défi avec un cadre de modélisation en deux étapes qui exploite les relations entre les covariables et la réponse dans les données pré-expérimentales et expérimentales. L'inférence se fait en estimant le résultat contrefactuel de chaque unité et en effectuant une comparaison par paires. Nous prouvons que cette méthode d'inférence est asymptotiquement sans biais, avec une variance asymptotique qui évolue avec la précision prédictive du modèle, mais qui n'est jamais plus grande que celle de l'estimateur naïf de la différence des moyennes.

[10:35-10:50]

Trang Bui (University of Waterloo) **Stefan Steiner** (University of Waterloo) **Nathaniel T. Stevens** (University of Waterloo)

Analysis of Experiments on Networks with Binary Outcomes

Analyse d'expériences sur les réseaux avec réponses binaires

Thanks to the popularity of online experiments, in which the treatment assignment of one unit may influence the outcome of another, the problem of experimental design and analysis under network interference is receiving increasing attention. While experiments with binary outcomes are common in practice, most experimental outcome models proposed in the literature are built for continuous outcomes. In this presentation, we consider a class of binary models to analyze binary

Grâce à la popularité des expériences en ligne, dans lesquelles l'affectation de traitement d'une unité peut influencer la réponse d'une autre, on accorde plus d'intérêt au problème du design expérimental et à l'analyse avec une interférence du réseau. Les expériences ayant des réponses binaires sont fréquentes en pratique, mais la plupart des modèles expérimentaux proposés dans la littérature sont conçus pour des réponses continus. Dans le cadre de cette présentation, nous évaluons une classe de modèles binaires afin d'analyser les expériences de réseaux binaires qui per-

network experiments which allows the network effects to be modeled flexibly using nonlinear functions. Based on the model, we define causal quantities and hypothesis tests of interest and demonstrate how estimation and inference can be done via the maximum likelihood framework. Different specifications of the proposed model class are then applied to analyze a real-world agricultural insurance experiment. This example demonstrates the need for nonlinear network effect modeling.

mettent aux effets de réseau d'être modélisés de façon flexible au moyen de fonctions non linéaires. En nous basant sur le modèle, nous définissons les quantités causales et les tests d'hypothèses pertinents puis démontrons de quelle façon l'estimation et l'inférence peuvent être réalisées par l'entremise d'un cadre de maximum de vraisemblance. Nous appliquons ensuite différentes spécifications à la classe de modèle proposée dans le but d'analyser une expérience réelle sur l'assurance agricole. Cet exemple démontre le besoin pour la modélisation d'effet de réseau non linéaire.

[10:50-11:05]

Nathan Phelps (University of Western Ontario)

Challenges when Calibrating a Random Forest Fit to Undersampled Data

Problèmes de calibrage pour une forêt aléatoire ajustée à des données sous-échantillonnées

Imbalanced binary classification problems arise in many fields of study, such as wildfire, health care, and finance. When using random forests to learn from imbalanced data, it is common to subsample the majority class (i.e., undersampling) to create a (more) balanced dataset for the random forest to learn from. This skews the random forest's predictions, so those wanting meaningful probability estimates try to calibrate them. One way of doing this is to map the original predictions to new values based on the sampling rates for the majority and minority classes, which were used to create the training dataset. However, calibrating a random forest this way has surprising consequences. The result is a prevalence estimate that depends on both i) the sampling rates used; and ii) the number of predictors considered at each split in the random forest. We explain why these have an impact and show the potential changes in prevalence estimates based on different choices of these hyperparameters.

Des problèmes de classification binaire avec données déséquilibrées surviennent dans plusieurs domaines d'étude, tels que les feux de forêt, les soins de santé et la finance. Lorsque des forêts aléatoires sont utilisées pour un apprentissage à partir de données déséquilibrées, il est courant de sous-échantillonner la classe majoritaire afin de créer un ensemble de données (plus) équilibré pour ajuster la forêt. Cette stratégie biaise toutefois les prévisions de la forêt aléatoire, et donc si on souhaite estimer correctement les probabilités on essayera de les calibrer. Une façon de le faire est de faire correspondre les prévisions originales à de nouvelles valeurs en se basant sur les taux d'échantillonnage utilisés dans les classes majoritaire et minoritaire pour créer l'ensemble de données d'entrainement. Calibrer une forêt aléatoire de cette manière a cependant des conséquences étonnantes. Il en résulte une estimation de la prévalence qui dépend à la fois i) des taux d'échantillonnage utilisés et ii) du nombre de prédicteurs pris en compte à chaque division de la forêt aléatoire. Nous expliquons pourquoi ces éléments ont un impact et montrons les changements potentiels dans les estimations de prévalence en fonction de différents choix de ces hyperparamètres.

[11:05-11:20]

Yutong Lu (University of Toronto) **Yan Yi Li** (University of Toronto)

Knowledge Fusion of Large Language Models for Molecular Property Prediction

Fusion des connaissances des grands modèles de langage pour la prédiction des propriétés moléculaires

The Simplified Molecular Input Line Entry System (SMILES) is a chemical notation for detailing molecular structures, commonly used as input of the unstructured data for Large Language Models (LLMs) for molecular property prediction. To harness the unique strengths of different LLMs, we propose a novel statistical modeling method to fuse the outputs of multiple LLMs into a cohesive framework. We first train three distinct LLMs

Le système Simplified Molecular Input Line Entry System (SMILES) est une notation chimique permettant de détailler les structures moléculaires, couramment utilisée comme entrée des données non structurées pour les grands modèles de langage afin de prédire les propriétés moléculaires. Pour exploiter les forces uniques des différents grands modèles de langage, nous proposons une nouvelle méthode de modélisation statistique visant à fusionner les résultats de plusieurs grands modèles de langage dans un cadre

on a unified SMILES dataset and obtain preliminary predictions. These predictions are then used as inputs of second-level model, where we fuse the results from the first-level models and produce the final predictions. Alternatively, we assign subsets of one SMILES dataset to multiple LLMs and obtain the final prediction as a weighted combination of the results from different models. Our results show that the LLMs fusion architecture has better performance than standard deep learning baselines, showcasing its potential in advancing molecular property prediction.

cohérent. Nous commençons par entraîner trois grands modèles de langage distincts sur un ensemble de données SMILES unifié et obtenons des prédictions préliminaires. Ces prédictions sont ensuite utilisées comme entrées dans le modèle de deuxième niveau, où nous fusionnons les résultats des modèles de premier niveau et produisons les prédictions finales. Parallèlement, nous attribuons des sous-ensembles d'un jeu de données SMILES à plusieurs grands modèles de langage et obtenons la prédiction finale sous la forme d'une combinaison pondérée des résultats de différents modèles. Nos résultats montrent que l'architecture de fusion de grands modèles de langage est plus efficace que les modèles d'apprentissage profond standard, ce qui démontre son potentiel pour améliorer la prédiction des propriétés moléculaires.

[11:20-11:35]

Bahram Moeinianfar (University of Manitoba) **Mohammad Jafari Jozani** (University of Manitoba)

Elite-Driven Support Vector Machine (EDSVM): Building Classifiers with Insights from a Collective of Support Vectors

Machine vectorielle de support dirigée par les élites (EDSVM) : construction de classifieurs avec des informations provenant d'une collection de vecteurs de support

In SVM, the construction of classifiers relies on a set of observations known as support vectors, determined by the choice of loss functions. Each loss function results in a specific decision boundary and identifies a unique set of support vectors, leading to varied classification performances. We propose a novel SVM methodology that gives additional weight to a collection of elite observations that play key roles in constructing SVM decision boundaries under various loss functions. These elite observations are identified as support vectors in various SVM models. We construct new loss functions to highlight the role of these elite observations during the training of our SVM classifiers. The loss functions for EDSVM are designed to be classification-calibrated, ensuring that they theoretically sound while enhancing the model's focus on these elite observations. Theoretical findings are presented, alongside thorough numerical analysis, to assess EDSVM's efficacy across various datasets.

En matière de machine vectorielle de support (SVM), la construction de classifieurs dépend d'un ensemble d'observations, appelées vecteurs de support, qui sont déterminées par le choix des fonctions de perte. Chaque fonction de perte produit en une borne décisionnelle spécifique et identifie un ensemble unique de vecteurs de support, ce qui entraîne des performances de classification variées. Nous proposons une nouvelle méthodologie SVM qui donne un poids additionnel à une collection d'observations élites qui jouent des rôles clés dans la construction de bornes décisionnelles SVM sous des fonctions de perte variées. Ces observations sont identifiées comme vecteurs de support dans divers modèles SVM. Nous construisons de nouvelles fonctions de perte pour mettre en lumière le rôle de ces observations élites pendant l'entrainement de nos classifieurs SVM. Les fonctions de perte EDSVM sont conçues pour être calibrées par classification, ce qui assure qu'elles sont théoriquement sensées, tout en centrant davantage le modèle sur les observations élites. Nos résultats théoriques sont présentés en même temps qu'une analyse numérique rigoureuse afin d'évaluer l'efficacité de l'EDSVM dans divers ensembles de données.

Funding Opportunities for Statistics and Data Science Research
Opportunités de financement pour la recherche en statistiques et en science des données

Chair/Président: Saman Muthukumarana

Organizer/Responsable: Saman Muthukumarana

Room/Salle: C 3033

Date: Monday June 3 / lundi 3 juin

Time/Heure: 10:20-11:50

Abstract/Résumé

[10:20-11:50]

Donald Estep (Simon Fraser University/CANSSI) **Pascal Marchand** (Natural Sciences and Engineering Research Council of Canada) **Ilana Gombos** (Canadian Institutes of Health Research) **Katarina Dedovic** (Canadian Institutes of Health Research) **Heidi Crummel** (MITACS)

Funding Opportunities for Statistics and Data Science Research

Possibilités de financement pour la recherche en statistique et science des données

This session explores the diverse funding opportunities available for researchers in the fields of statistics and data science. The session brings together funding agencies NSERC, CIHR, Mitacs and CANSSI to discuss various opportunities and strategies for securing funding from these granting agencies. Attendees will gain valuable insights into current funding opportunities, learn about them, and discover resources that can help propel their research projects to new heights.

Cette session explore les diverses possibilités de financement disponibles pour les chercheurs en statistique et science des données. La session réunit les agences de financement CRSNG, IRSC, Mitacs et INCASS pour discuter des diverses possibilités et des stratégies pour obtenir un financement de ces agences. Les participants en tireront des informations précieuses sur les possibilités de financement actuelles et découvriront des ressources qui peuvent les aider à propulser leurs projets de recherche vers de nouveaux sommets.

Fairness and Discrimination in Insurance
Équité et discrimination en assurance

Chair/Président: Marie-Pier Côté

Organizer/Responsable: Marie-Pier Côté

Room/Salle: A 1043

Date: Monday June 3 / lundi 3 juin

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-14:00]

Carlos Andres Araiza Iturria (University of Waterloo) **Mary Hardy** (University of Waterloo) **Paul Marriott** (University of Waterloo)

Discrimination in Insurance Pricing

La discrimination dans la tarification de l'assurance

Discrimination is an ongoing problem in the insurance industry that persists, regardless of intentionality, when the insurer blinds the pricing process from the socially controversial or legally prohibited input. We contextualize the problem in P&C insurance considering the prevailing legislations in the United States and the European Union. We first analyze the largest publicly available database of police-reported motor vehicle traffic accidents in the United States. Then, we set the pricing problem under a causal framework to provide tools that can help identify sources of discrimination. This leads to a criteria we propose to obtain actuarially fair premiums that can mitigate discrimination. The aforementioned comes together in a microsimulation approach for auto insurance in the U.S. that approximates population statistics and reported accident data reported. Our results show that some traditional actuarial assumptions can result in unintended discriminatory consequences.

La discrimination est un problème récurrent dans l'industrie de l'assurance et qui persiste, indépendamment de l'intentionnalité, lorsque l'assureur s'assure que le processus de tarification n'a pas d'information sur les variables socialement controversées ou légalement interdites. Nous contextualisons le problème dans l'assurance des particuliers en tenant compte de la législation en vigueur aux États-Unis et dans l'Union européenne. Nous analysons d'abord la plus grande base de données publiquement accessible sur les accidents de la circulation routière signalés par la police aux États-Unis. Ensuite, nous formulons le problème de tarification dans un cadre causal pour fournir des outils pouvant aider à identifier les sources de discrimination. Cela conduit à des critères que nous proposons pour obtenir des primes actuariellement équitables qui peuvent atténuer la discrimination. Ceux-ci sont réunis dans une approche de microsimulation pour l'assurance automobile aux États-Unis qui approxime les statistiques populationnelles et les données d'accidents déclarés. Nos résultats montrent que certaines hypothèses actuarielles traditionnelles peuvent entraîner des conséquences discriminatoires non intentionnelles.

[14:00-14:30]

Chengguo Weng (University of Waterloo)

Optimal Prediction under Several Fairness Criteria

Prévision optimale selon plusieurs critères d'équité

In recent years, there has been a notable uptick in interest surrounding the fairness of insurance premium rating, coupled with increased regulatory scrutiny aimed at prohibiting the use of certain sensitive policyholder characteristics variables in insurance pricing. In my

Ces dernières années, on s'intéresse de plus en plus à améliorer l'équité de la tarification des primes d'assurance, parallèlement à une surveillance réglementaire accrue visant à interdire l'utilisation de certaines variables sensibles relatives aux caractéristiques des assurés. Dans mon exposé, je me pencherai sur les discus-

talk, I will delve into discussions concerning the optimal prediction of insurance premiums under various fairness criteria, including the consideration of uncorrelation and adherence to principles like the four-fifths rule. By examining these factors, we aim to foster a deeper understanding of how insurance premiums can be determined in a manner that prioritizes fairness and equity for policyholders.

sions concernant la prévision optimale des primes d'assurance pour divers critères d'équité, notamment la prise en compte de la non-corrélation et l'adhésion à des principes tels que la règle des quatre cinquièmes. En examinant ces facteurs, nous souhaitons mieux comprendre comment les primes d'assurance peuvent être déterminées d'une manière qui donne la priorité à la justice et à l'équité pour les assurés.

[14:30-15:00]

Marie-Pier Côté (Université Laval) **Olivier Côté** (Université Laval) **Arthur Charpentier** (Université du Québec à Montréal)

A Fair Price to Pay: Exploiting Causal Graphs for Fairness in Insurance

Un juste prix à payer : exploiter les graphes causaux pour l'équité en assurance

In many countries, insurance companies must not discriminate on some given policyholder characteristics. Omission of prohibited variables from models prevents direct discrimination, but fails to address proxy discrimination. To this end, multiple fairness methodologies exist but lead to different results. In this talk, we review causal inference notions and introduce a causal graph tailored for fairness in insurance. Exploiting these, we discuss potential sources of bias, formally define direct and indirect discrimination, and study the properties of fairness methodologies. A novel categorization of fair methodologies into five families (best-estimate, unaware, aware, hyperaware, and corrective) is constructed based on their expected fairness properties. A pedagogical example illustrates our findings on the interplay between our fair score families and sources of discrimination.

Dans plusieurs pays, les compagnies d'assurance ne doivent pas discriminer selon certains attributs des personnes assurées. L'omission des variables prohibées dans les modèles prévient la discrimination directe, mais pas la discrimination indirecte. Plusieurs méthodes ont été proposées pour rendre les modèles équitables, mais leurs résultats divergent. Dans cette présentation, on rappelle des notions d'inférence causale et on propose un graphe causal conçu pour l'équité en assurance. En tirant parti de ces outils, on discute des sources possibles de biais, on définit les discriminations directe et indirecte et on étudie les propriétés des méthodes équitables. On construit une nouvelle catégorisation des méthodes en cinq familles (précis, ignorant, conscient, hyperconscient, réparateur) selon leurs propriétés d'équité. On illustre nos constats à l'aide d'un exemple pédagogique, qui permet de souligner le lien entre les familles de scores équitables et les sources de discrimination.

Survey Methods Section Presidential Invited Address
Allocution de l'invité du président du Groupe des méthodes d'enquête

Chair/Président: Éric Gagnon

Organizer/Responsable: Éric Gagnon

Room/Salle: A 1045

Date: Monday June 3 / lundi 3 juin

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-15:00]

David Haziza (University of Ottawa) **Mehdi Dagdoug** (McGill University) **Camelia Goga** (Université de Bourgogne Franche Comté)

Statistical Inference in the Presence of Imputed Survey Data through Regression Trees and Random Forests

Inférence statistique en présence de données d'enquête imputées au moyen d'arbres de régression et de forêts aléatoires

In recent years, machine learning procedures have attracted much attention in National Statistical Offices. Item nonresponse in surveys is usually handled through some form of single imputation. Regression trees and random forests provide flexible tools for obtaining a set of imputed values. Belonging to the class of non-parametric methods, these methods have the ability to capture nonlinear trends in the data and tend to be robust to the non-inclusion of interactions or predictors accounting for curvature. In this presentation, we will discuss the properties of imputed estimators based on regression trees and random forests. We will also discuss a novel variance estimator based on the so-called reverse approach for variance estimation. We will present the results from a simulation study to assess the proposed methods in terms of bias and efficiency. Finally, the choice of hyper-parameters will be discussed.

Ces dernières années, les procédures d'apprentissage automatique ont suscité beaucoup d'intérêt de la part des bureaux nationaux de statistiques. La non-réponse partielle dans les enquêtes est généralement traitée par une forme d'imputation simple. Les arbres de régression et les forêts aléatoires constituent des outils flexibles permettant d'obtenir un ensemble de valeurs imputées. Appartenant à la classe des méthodes non paramétriques, ces méthodes ont la capacité de capturer des tendances non linéaires dans les données et tendent à être robustes à la non-inclusion d'interactions ou de prédicteurs associés à la courbure. Dans cette présentation, nous discuterons des propriétés des estimateurs imputés basés sur les arbres de régression et les forêts aléatoires. Nous discuterons également d'un nouvel estimateur de variance basé sur l'approche dite renversée pour l'estimation de la variance. Nous présenterons les résultats d'une étude par simulation visant à évaluer les méthodes proposées en termes de biais et d'efficacité. Enfin, nous discuterons du choix des hyperparamètres.

Chair/Président: Dehan Kong

Organizer/Responsable: Dehan Kong

Room/Salle: A 1046

Date: Monday June 3 / lundi 3 juin

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-14:00]

Annie Qu (University of California, Irvine) **Hansen Ye** (UC Irvine) **Wenzhuo Zhou** (UC Irvine) **Ruoqing Zhu** (UIUC)

Stage-Aware Learning for Dynamic Treatments

Apprentissage conscient des étapes pour les traitements dynamiques

Recent advances in dynamic treatment regimes (DTRs) provide powerful optimal treatment searching algorithms, which are tailored to individuals' specific needs and able to maximize their expected clinical benefits. However, existing algorithms could suffer from insufficient sample size under optimal treatments, especially for chronic diseases involving long stages of decision-making. To address these challenges, we propose a novel individualized learning method which estimates the DTR with a focus on prioritizing alignment between the observed treatment trajectory and the one obtained by the optimal regime across decision stages. By relaxing the restriction that the observed trajectory must be fully aligned with the optimal treatments, our approach substantially improves the sample efficiency and stability of inverse probability weighted based methods. The proposed learning scheme builds a more general framework which includes the popular outcome weighted learning framework.

Les récents progrès dans les régimes de traitement dynamique (DTR) offrent de puissants algorithmes de recherche de traitement optimal, adaptés aux besoins spécifiques des individus et capables de maximiser leurs avantages cliniques attendus. Cependant, les algorithmes existants pourraient souffrir d'un échantillonnage insuffisant sous des traitements optimaux, en particulier pour les maladies chroniques impliquant de longues étapes de prise de décision. Pour relever ces défis, nous proposons une nouvelle méthode d'apprentissage individualisée qui estime le DTR en mettant l'accent sur la priorisation de l'alignement entre la trajectoire de traitement observée et celle obtenue par le régime optimal à travers les étapes de décision. En relaxant la restriction selon laquelle la trajectoire observée doit être entièrement alignée avec les traitements optimaux, notre approche améliore substantiellement l'efficacité et la stabilité des méthodes basées sur le poids de probabilité inverse. Le schéma d'apprentissage proposé construit un cadre plus général qui inclut le populaire cadre d'apprentissage pondéré par les résultats.

[14:00-14:30]

Peter X. Song (University of Michigan)

Supervised Homogeneity Pursuit via Mixed Integer Optimization

Recherche d'homogénéité supervisée par optimisation en nombres entiers mixtes

Stratification is one statistical principle in data processing to mitigate the underlying population heterogeneity, which is typically handled by clustering when stratum labels are unknown. Many practical problems require post-clustering statistical learning that is challenged by the difficulty of uncertainty quantification. One solution to address this challenge is to perform a simultaneous

La stratification est un principe statistique du traitement des données visant à atténuer l'hétérogénéité sous-jacente de la population, qui est généralement traitée par regroupement en clusters en l'absence d'étiquettes pour les strates. De nombreux problèmes pratiques nécessitent un apprentissage statistique post-regroupement, mais il est souvent difficile de quantifier l'incertitude. Une solution alors est d'effectuer une opération si-

operation of clustering and estimation in data analyses. We propose a new paradigm of supervised homogeneity pursuit via mixed integer optimization, which provides a conceptually simple and computationally straightforward machinery with the use of suitable constraints in optimization. This toolbox has been then applied to solve several real-world problems arising from infectious disease surveillance, influence of environmental exposure to health, and risk factors for aging. Some algorithmic limitations worth future research will be discussed.

multanée de regroupement et d'estimation dans les analyses de données. Nous proposons un nouveau paradigme de recherche d'homogénéité supervisée par optimisation en nombres entiers mixtes, mécanisme conceptuellement simple et computationnellement direct lorsque des contraintes appropriées sont utilisées dans l'optimisation. Nous appliquons cette boîte à outils à la résolution de plusieurs problèmes réels liés à la surveillance des maladies infectieuses, à l'influence d'une exposition environnementale sur la santé et aux facteurs de risque du vieillissement. Nous discuterons de certaines limitations algorithmiques méritant des recherches futures.

[14:30-15:00]

Peijun Sang (University of Waterloo) **Yao Luo** (University of Toronto)

Penalized Sieve Estimation of Structural Models

Estimation pénalisée par tamis des modèles structurels

Existing methods for estimating structural models are often computationally or statistically inefficient, depending on how equilibrium conditions are imposed. We propose a class of penalized sieve estimators by approximating the solution with a linear combination of basis functions and imposing equilibrium conditions as a penalty in search of the best-fitting coefficients. Like the MPEC estimator, our estimators avoid solving the model repeatedly, apply to a broad class of models, and are consistent, asymptotically normal, and asymptotically efficient. On the other hand, they solve unconstrained optimization problems with fewer unknowns and produce standard errors akin to the conventional MLE. As an illustration, we apply our method to an entry game between Walmart and Kmart.

Dépendamment de la façon dont les conditions d'équilibre sont imposées, les méthodes existantes pour l'estimation des modèles structurels sont souvent inefficaces, tant sur le plan computationnel que statistique. Nous proposons une classe d'estimateurs tamis par approximation de la solution à l'aide d'une combinaison linéaire de fonctions de base et l'imposition de conditions d'équilibre comme pénalité dans la recherche des coefficients les mieux ajustés. À l'instar de l'estimateur de programmation mathématique avec contraintes d'équilibre (MPEC), nos estimateurs évitent de résoudre le modèle à répétition, s'appliquent à une vaste classe de modèles, sont cohérents, ainsi que d'une normalité et d'une efficacité asymptotiques. Par ailleurs, ils résolvent les problèmes d'optimisation sans contraintes avec moins d'inconnus et produisent des erreurs standards proches de celles de l'estimateur du maximum de vraisemblance (MLE) conventionnel. Pour illustrer notre méthode, nous l'appliquons à un jeu d'entrée entre Walmart et Kmart.

Reproducibility in Machine Learning and Statistics
Reproductibilité en apprentissage automatique et statistique

Chair/Président: Tiffany A. Timbers

Organizer/Responsable: Tiffany A. Timbers

Room/Salle: C 2045

Date: Monday June 3 / lundi 3 juin

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-14:00]

Rohan Alexander (University of Toronto)

Reproducibility and Code - Using LLMs to Translate Replication Packages to Enhance Credibility

Reproductibilité et code — utilisation des LLMs pour traduire les progiciels de réplication pour hausser la crédibilité

Social sciences such as economics and political science, as well as statistics, have put in place the requirement for replication packages to accompany applied papers in many top journals. Trisovic (2022) and others systematically evaluate the content of these packages to identify common errors that would prevent them from running. We use Large Language Models (LLMs) to refactor the code underpinning these replication packages into Python. We identify how, just when translating spoken languages there are some words that cannot be directly translated, our process identifies particular functions that cause considerable issues. We also identify how this refactoring identifies some issues in the original code. Our approach would be relatively easy to integrate into existing journal paper acceptance workflows and would enhance the credibility of papers.

Les sciences sociales comme l'économie et la science politique, ainsi que les statistiques, ont mis en place une exigence pour qu'un progiciel de réplication accompagne les documents écrits dans plusieurs revues importantes. Trisovic (2022) et d'autres évaluent systématiquement le contenu de ces progiciels pour trouver les erreurs fréquentes qui les empêcheraient de fonctionner. Nous utilisons les grands modèles de langage (LLMs) pour refactoriser le code à la base de ces progiciels de réplication dans Python. Tout comme certains mots ne peuvent pas être traduits littéralement dans la langue parlée, nous démontrons que nos processus découvrent certaines fonctions qui causent des problèmes considérables. Nous montrons aussi de quelle façon cet refactorisation détecte des problèmes dans le code d'origine. Notre approche serait relativement facile à intégrer dans les flux d'acceptation d'articles de revue et pourrait rehausser la crédibilité des revues.

[14:00-14:30]

Callandra Moore (The Hospital for Sick Children)

Reproducibility in Clinical NLP

Reproductibilité dans le traitement automatique des langues en clinique

In healthcare, the increasing volume of electronic health record data from hospitals provides a rich source of information on patient state, treatment, and outcome as semi-structured and unstructured text. Despite increasing use of natural language processing in research and clinical contexts, reproducibility of this work is hindered by privacy protections and institutional regulations preventing the sharing of clinical text. Though automated deidentification plays a pivotal role in mitigating these challenges by enabling the sharing of

En soins de santé, la hausse du volume de données des dossiers médicaux électroniques des hôpitaux procure une source riche en information sur l'état, le traitement et l'issue des patients sous la forme de texte semi-structuré et non structuré. Malgré l'utilisation croissante du traitement des langues naturelles en recherche et en contexte clinique, la reproductibilité de ce travail demeure entravée par la protection de la vie privée et les régulations institutionnelles qui préviennent le partage de documents cliniques. La dépersonnalisation automatisée joue un rôle important pour surmonter ces obstacles en permettant le partage

anonymized data, deidentification algorithms face their own reproducibility challenges. Accurately assessing and comparing deidentification algorithms is complex, with variations in evaluation techniques and anonymization standards across studies and institutions. Improved reproducibility can be achieved through standardized deidentification protocols, transparent reporting of methods, and open-access resources for sharing evaluation metrics.

de données anonymes, mais les algorithmes de dépersonnalisation comportent leurs propres défis de reproductibilité. Il est complexe d'évaluer avec précision et de comparer les algorithmes de dépersonnalisation, en raison des variations entre les techniques d'évaluation et les standards d'anonymisation à travers les études et les établissements. Une reproductibilité supérieure est réalisable par l'entremise de protocoles de dépersonnalisation standardisés, de descriptions transparentes sur les méthodes et de ressources libres pour partager les mesures d'évaluation.

Generative Artificial Intelligence and What's Next for the Teaching and Learning of Statistics (Panel)
Intelligence artificielle générative et l'avenir de l'enseignement et de l'apprentissage de la statistique (Table ronde)

Chair/Président: Alison L. Gibbs

Organizer/Responsable: Alison L. Gibbs

Room/Salle: A 2071

Date: Monday June 3 / lundi 3 juin

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-15:00]

David Riegert (Trent University) **Nathan Taback** (University of Toronto)

Generative Artificial Intelligence and What's Next for the Teaching and Learning of Statistics

Intelligence artificielle générative et l'avenir de l'enseignement et de l'apprentissage de la statistique

Generative artificial intelligence tools can write code to carry out a data analysis and write a report describing the findings. They can produce a lesson plan, assessment questions and a rubric. They can be trained to answer questions on any topic. They have the potential to disrupt many of our daily practices, raising many questions for instructors of statistics. Should we change what we teach? Should we change how we teach and how we assess? What does it mean to integrate the use of generative AI tools responsibly and ethically? The session will explore these questions through an open participatory conversation. The conversation will be informed by demonstrations and reflections from two statistics instructors who are artificial intelligence users: the first discussing the use of generative AI tools for carrying out data analysis and the second discussing the use of generative AI to support to create teaching materials.

Les outils d'intelligence artificielle générative peuvent écrire un code pour effectuer une analyse de données et rédiger un rapport en décrivant les résultats. Ils peuvent produire un plan de cours, des questions et une grille d'évaluation. Ils peuvent être formés pour répondre à des questions sur n'importe quel sujet. Ils ont le potentiel de perturber bon nombre de nos pratiques quotidiennes, ce qui soulève de nombreuses questions pour les professeurs de statistique. Devons-nous changer ce que nous enseignons ? Devons-nous changer notre façon d'enseigner et d'évaluer ? Que signifie intégrer l'utilisation d'outils d'IA générative de manière responsable et éthique ? La session explorera ces questions à travers une conversation participative ouverte. La conversation sera alimentée par des démonstrations et les réflexions de deux professeurs de statistique qui sont des utilisateurs d'intelligence artificielle : le premier discutera de l'utilisation d'outils d'IA générative pour effectuer des analyses de données et le second discutera de l'utilisation de l'IA générative pour soutenir la création de matériel d'enseignement.

Modern Approaches for Clustering Data
Approches modernes pour le regroupement des données

Chair/Président: Sanjeena Dang

Organizer/Responsable: Brian Franczak

Room/Salle: A 2065

Date: Monday June 3 / lundi 3 juin

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-13:52]

Mateen Shaikh (Thompson Rivers University)

Applicants of Sequences of Surrogate Functions in Statistcal Learning

Applications de séquences de fonctions de remplacement à l'apprentissage statistique

Previous work on sequences of surrogate functions provided methods of finding optima, with a greater chance of finding global optima, in optimization problems. Augmenting several likelihood-based approaches in statistical learning with these sequences afford an efficient means of addressing unsupervised learning problems embedded in other learning outcomes, be they supervised or unsupervised. Here, we present additional applicants of sequences of surrogate functions to regression (supervised) and clustering (unsupervised) to address problems often solved through less efficient search-and-score methods.

Les travaux antérieurs sur les séquences de fonctions de substitution ont fourni des méthodes de recherche d'optima, avec une plus grande chance de trouver des optima globaux, dans les problèmes d'optimisation. L'augmentation de plusieurs approches basées sur la vraisemblance dans l'apprentissage statistique avec ces séquences offre un moyen efficace d'aborder les problèmes d'apprentissage non supervisé intégrés dans d'autres résultats d'apprentissage, qu'ils soient supervisés ou non supervisés. Nous présentons ici d'autres candidats à l'utilisation de séquences de fonctions de substitution pour la régression (supervisée) et le regroupement (non supervisé) afin de résoudre des problèmes souvent résolus par des méthodes de recherche et de score moins efficaces.

[13:52-14:15]

John R.J. Thompson (The University of British Columbia) **Jesse Ghashti** (The University of British Columbia)

Kernel Metric Learning for Mixed-type Distance Shrinkage and Variable Selection

Apprentissage de mesure de noyaux pour le rétrécissement de distance de type mixte et la sélection de variables

The choice of metric for distance-based clustering of continuous and categorical mixed-type datasets can affect and limit the ability of a clustering algorithm to capture underlying non-linear grouping structures, particularly for high-dimensional data. In this talk, we discuss flexible kernel metrics that measure similarity and balance continuous and categorical variable types for distance calculations. We introduce maximum similarity cross-validation to select optimal bandwidths for balancing and selecting relevant variables, and demonstrate that kernel metric learning is a shrinkage method from a maximum dissimilarity metric to a uniform dissimilarity metric. We show improved clustering accuracy

Le choix de mesure pour un regroupement basé sur la distance d'ensembles de données de type mixte catégoriques et continues peut influencer et restreindre la capacité d'un algorithme de regroupement servant à capturer les structures de regroupement non linéaires sous-jacentes, tout particulièrement pour des données de grande dimension. Dans le cadre de cet exposé, j'aborderai les mesures de noyaux flexibles servant à mesurer la similarité et l'équilibre de variables catégoriques et continues pour les calculs de distance. Nous présentons une validation croisée à similarité maximale afin de sélectionner les bandes passantes optimales pour équilibrer et sélectionner les variables pertinentes. Puis nous démontrons que l'apprentissage de mesure de noyaux est une méthode de rétrécissement partant d'une mesure de différence

when utilized in distance-based clustering algorithms on simulated and real-world mixed-type datasets. We find that kernel metric learning can smooth out irrelevant variables for distance measurement, and balance variables important to dissimilarities within each mixed-type dataset.

maximale à une mesure de différence uniforme. Son emploi génère une précision de regroupement supérieure avec des algorithmes de regroupement basés sur la distance à partir d'ensemble de données réelles et simulées de type mixte. Nous trouvons que l'apprentissage de mesure de noyaux peut faire disparaître les variables non pertinentes de la mesure de distance et équilibrer les variables importantes et les différences entre chaque ensemble de données.

[14:15-14:37]

Paul David McNicholas (McMaster University) **Mackenzie Neal** (McMaster University)

Flexible Variable Selection for Clustering

Sélection de variables flexibles pour le regroupement

The subject of variable selection in clustering has been widely studied for around two decades. However, relatively little attention has been given to the case where the clusters may be skewed and/or heavy-tailed. An approach is introduced for the multivariate case and an extension to the matrix-variate case is also discussed. Simulated and real data are used for illustration.

La question de la sélection des variables dans le cadre du regroupement a été largement étudiée depuis une vingtaine d'années. Cependant, on a accordé relativement peu d'attention au cas où les grappes peuvent être asymétriques et/ou à forte queue. Une approche est introduite pour le cas multivarié et une extension au cas matriciel-varié est également discutée. Des données simulées et réelles sont utilisées à titre d'illustration.

[14:37-15:00]

Brian Franczak (MacEwan University)

Outlier Detection using an Asymmetric Laplace Distribution

Détection des valeurs aberrantes à l'aide d'une distribution de Laplace asymétrique

Classification can be defined as the process of assigning labels to unlabelled observations to create homogeneous groups of observations. In unsupervised classification, also known as clustering, no prior information about a potential group structure is available. Model-based clustering means that a finite mixture model is used for unsupervised classification. This talk will present an approach for performing model-based clustering of incomplete multivariate data sets that exhibit asymmetric features in the presence of outlying observations. This approach will model asymmetry directly while simultaneously performing imputation. We use an expectation-maximization (EM) based scheme for parameter estimation. At convergence, we use likelihood-based criteria like the Bayesian information criterion for model selection and assess classification performance using the adjusted Rand index. We demonstrate the effectiveness of the proposed models using simulated and real data sets.

La classification peut être définie comme le processus d'attribution d'étiquettes à des observations non étiquetées afin de créer des groupes d'observations homogènes. La classification non supervisée, également appelée « groupement de données par classe », ne comporte aucune donnée préalable sur la structure d'un groupe potentiel. Le groupement de données par classe basé sur un modèle signifie qu'un modèle de mélange fini est utilisé pour la classification non supervisée. Nous présenterons une approche permettant d'effectuer un groupement de données par classe basé sur un modèle à partir de données multivariées incomplètes qui présentent des caractéristiques asymétriques en présence d'observations aberrantes. Cette approche modélise directement l'asymétrie tout en effectuant simultanément l'imputation. Nous utilisons un schéma basé sur l'algorithme espérance-maximisation pour estimer les paramètres. Au moment de la convergence, nous utilisons des critères de vraisemblance, tels que le critère d'information bayésien pour la sélection des modèles, puis nous évaluons l'efficacité de la classification à l'aide de l'indice de Rand ajusté. Nous démontrons l'efficacité des modèles proposés avec des ensembles de données simulées et réelles.

Distinguished Educator Award Address
Allocution du recipiendaire du Prix d'excellence en enseignement

Chair/Président: Wesley S. Burr

Organizer/Responsable: Wesley S. Burr

Room/Salle: IIC 2001

Date: Monday June 3 / lundi 3 juin

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-15:00]

Jim B. Stallard (University of Calgary)

How to Clear a Room

Comment vider la salle

In my experience, the profession of "Statistician" is often viewed as "social pariah". Reactions to social situation inquiries of "Jim, what do you do for a living?" may be placed in the 'not positive" bin at a much higher frequency when compared to responses that would be classified as 'positive'. In my presentation, I will attempt to illustrate how we as educators and researchers of the Discipline can perhaps minimize many of the negative stereotypes associated with our profession.

D'après mon expérience, la profession de « statisticien » est souvent vue comme « paria social ». Les réactions à la question « Jim, que fais-tu dans la vie ? » peuvent être classées dans la catégorie « non positive » à une fréquence beaucoup plus élevée que les réponses qui seraient classées comme « positives ». Dans ma présentation, je tenterai d'illustrer comment, en tant qu'éducateurs et chercheurs de la discipline, nous pouvons peut-être minimiser bon nombre de stéréotypes négatifs associés à notre profession.

Chair/Président: Wendy Lou

Organizer/Responsable: Wendy Lou

Room/Salle: A 1049

Date: Monday June 3 / lundi 3 juin

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-14:00]

Motomi Mori (St. Jude Children's Research Hospital)

Clinical Trial Designs to Evaluate Gene and Cell Therapies in Rare Diseases

Conception d'essais cliniques pour évaluer les thérapies géniques et cellulaires dans les maladies rares

With the recent advancement in genome editing and mRNA therapeutic technologies, there is much excitement and hope for diseases once considered incurable. For example, last December the U.S. FDA approved Casgevy and Lyfgenia, cell-based gene therapies for the treatment of older children and adults with advanced sickle cell disease. More recently the U.S. FDA approved Amtagvi, T cell immunotherapy based on patient's own tumor T cells for advanced melanoma. These treatments are individualized according to each patient cells or genetic mutations, so healthy cells or correct genetic materials can be delivered. In this talk, I will use an example of Antisense Oligonucleotide (ASO) Therapy for rare genetic neurological diseases among children to describe a basket trial of many related N of 1 trials, evaluate operating characteristics, and discuss logistical considerations of such a trial as well as challenges of clinical evaluations of individually tailored therapies in rare diseases.

Les progrès récents des technologies d'édition du génome et de thérapie par l'ARNm suscitent beaucoup d'enthousiasme et d'espoir pour des maladies autrefois considérées comme incurables. Par exemple, en décembre dernier, la FDA a approuvé Casgevy et Lyfgenia, des thérapies géniques cellulaires pour le traitement des grands enfants et des adultes atteints de drépanocytose avancée. Plus récemment, la FDA a approuvé Amtagvi, une immunothérapie cellulaire basée sur les cellules T tumorales du patient, pour le traitement du mélanome avancé. Ces traitements sont individualisés en fonction des cellules ou des mutations génétiques de chaque patient, de sorte à fournir des cellules saines ou du matériel génétique correct. Dans cet exposé, j'utiliserai un exemple de thérapie par oligonucléotides antisens (ASO) pour des maladies neurologiques génétiques rares chez les enfants pour décrire un essai collectif de plusieurs essais N de 1 apparentés, évaluer les caractéristiques de fonctionnement et discuter des considérations logistiques d'un tel essai ainsi que des défis des évaluations cliniques des thérapies personnalisées pour les maladies rares.

[14:00-14:30]

Bingshu Chen (Queen's University) **Wenyu Jiang** (Queen's University) **Parisa Gavanji** (Queen's University)

Penalized Likelihood Ratio Test for a Biomarker Threshold Effect in Clinical Trials Based on Generalized Linear Models

Test de rapport de vraisemblance pénalisé pour l'effet de seuil d'un biomarqueur dans des essais cliniques basés sur des modèles linéaires généralisés

In a clinical trial, the responses to the new treatment may vary among patient subsets with different characteristics in a biomarker. It is often necessary to examine whether there is a cutpoint for the biomarker that divides the patients into two subsets of those with more favourable and less favourable responses. More generally, we approach

Dans un essai clinique, les réponses au nouveau traitement peuvent varier entre sous-ensembles de patients présentant des caractéristiques différentes pour un biomarqueur. Il est souvent nécessaire d'examiner s'il existe un seuil pour le biomarqueur qui divise les patients en deux sous-ensembles, ceux dont les réponses sont plus ou moins favorables. Plus généralement, nous abordons

this problem as a test of homogeneity in the effects of a set of covariates in generalized linear regression models. We propose a penalized likelihood ratio test to overcome the model irregularities. Under the null hypothesis, we prove that the asymptotic distribution of the proposed test statistic is a mixture of chi-squared distributions. In extensive simulation studies, we find that the proposed test works well in terms of size and power. We further demonstrate the use of the proposed method by applying it to clinical trial data from the Digitalis Investigation Group (DIG) on heart failure.

ce problème comme un test d'homogénéité des effets d'un ensemble de covariables dans les modèles de régression linéaire généralisée. Nous proposons un test de rapport de vraisemblance pénalisé pour surmonter les irrégularités du modèle. Sous l'hypothèse nulle, nous prouvons que la distribution asymptotique de la statistique de test proposée est un mélange de distributions chi-carré. Des études de simulation approfondies montrent que le test proposé fonctionne bien en termes de taille et de puissance. Nous démontrons également l'utilité de la méthode proposée en l'appliquant aux données des essais cliniques du Digitalis Investigation Group (DIG) sur l'insuffisance cardiaque.

[14:30-15:00]

Aaron Hudson (Fred Hutchinson Cancer Center) **Oliver Dukes** (Ghent University) **Mats Stensrud** (École Polytechnique Fédérale de Lausanne (EPFL)) **Riccardo Brioschi** (École Polytechnique Fédérale de Lausanne (EPFL))

A Nonparametric Test and Estimand for Qualitative Interactions

Test non paramétrique et estimation des interactions qualitatives

Qualitative effect heterogeneity occurs when treatment is beneficial for certain sub-groups and harmful for others. This specific type of heterogeneity is of clinical interest when treatment decisions will be tailored to individual characteristics. The problem of testing for qualitative heterogeneity has been well-studied when the comparison is made between finite subgroups. However, the problem is more challenging when the potential effect modifiers are continuous, and one wishes to infer heterogeneity under a nonparametric model. We propose a class of nonparametric tests for qualitative heterogeneity. Compared with some recent approaches, our proposal can incorporate a variety of structured assumptions on the conditional average treatment effect, extends to moderate/high-dimensional covariates and does not require sample splitting. The utility of the proposal is borne out in simulation studies and a re-analysis of data from a recent clinical trial.

L'hétérogénéité des effets qualitatifs survient lorsqu'un traitement est bénéfique pour certains sous-groupes, mais mauvais pour d'autres. Ce type spécifique d'hétérogénéité est d'un intérêt clinique lorsque les décisions en matière de traitement répondent à des caractéristiques individuelles. Le problème de tester l'hétérogénéité qualitative a été bien étudié lorsque la comparaison est faite entre des sous-groupes finis. Le problème est toutefois plus ardu lorsque les modificateurs de l'effet potentiel sont continus et que l'on veut inférer l'hétérogénéité sous un modèle non paramétrique. Nous proposons une classe de tests non paramétriques de l'hétérogénéité qualitative. Comparée à certaines approches récentes, notre proposition peut incorporer une diversité d'hypothèses structurées sur l'effet moyen conditionnel du traitement, s'étendre à des covariables de dimension modérée ou élevée, tout en ne nécessitant pas le fractionnement de l'échantillon. L'utilité de la proposition est confirmée dans des études de simulation et une nouvelle analyse des données d'un essai clinique récent.

Advanced Statistical Inference for Emerging Infectious Disease Epidemics
Inférence statistique avancée pour les épidémies de maladies infectieuses émergentes

Chair/Président: Lam Ho

Organizer/Responsable: Lam Ho

Room/Salle: ED 2018A

Date: Monday June 3 / lundi 3 juin

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-14:00]

Cindy Feng (Dalhousie University)

Spatial Generalized Additive Location Scale Models for Modeling Infectious Disease Risk

Modèles de position-échelle spatiaux additifs généralisés pour la modélisation du risque de maladies infectieuses

In spatial analysis, the focus has traditionally been on modeling spatial correlation in the residuals of the mean structure, often neglecting the consideration of spatial effects on other parameters. The implications of disregarding the modeling of these parameters and their impact on parameter estimates and model goodness of fit remain unclear. Motivated by a real-world application, our findings uncover spatial correlation not only in the mean but also in the variance of disease incidence. We also investigate the consequences of mismodeling the structure in various components of the regression model. Through simulation studies, we assess the sensitivity of parameter estimates to model misspecification. This research contributes to a deeper understanding of disease dynamics and emphasizes the need for robust modeling across all parameters.

En analyse spatiale, l'accent a traditionnellement été mis sur la modélisation de la corrélation spatiale dans les résidus de la structure moyenne, en négligeant souvent les effets spatiaux sur d'autres paramètres. Les implications de ne pas inclure une telle modélisation, et leurs impacts sur l'estimation des paramètres sont encore méconnus. Motivés par une application réelle, nous mettons en évidence une corrélation spatiale non seulement dans la moyenne mais aussi dans la variance de l'incidence de la maladie. Nous étudions également les conséquences d'une mauvaise modélisation de la structure dans divers composants du modèle de régression. Via des études de simulation, nous évaluons la sensibilité des estimations de paramètres à une mauvaise spécification du modèle. Ces travaux contribueront à une meilleure compréhension de la dynamique des maladies et soulignent la nécessité d'une modélisation robuste pour tous les paramètres.

[14:00-14:30]

Justin James Ian Slater (University of Guelph)

Overdispersed or Underreported? Inference for Infectious Disease Models with Underreported Case Counts

Surdispersion ou sous-déclaration ? Inférence pour des modèles de maladies infectieuses avec comptages de cas sous-déclarés

Infectious disease surveillance data often suffers from underreporting, posing challenges for studying disease dynamics at the population level. Traditionally, statisticians have approached this issue with skepticism, viewing the simultaneous estimation of reporting probability alongside other model parameters as infeasible without strong informative priors. In this talk, I will introduce Poisson Network Autoregressions (PNAR), statistical analogues to discrete-time susceptible-infectious-

Les données de surveillance des maladies infectieuses sont souvent sous-déclarées, ce qui rend difficile l'étude de la dynamique de la maladie au niveau de la population. Traditionnellement, les statisticiens ont envisagé ce problème avec scepticisme, percevant comme irréalisable l'estimation simultanée de la probabilité de déclaration et des autres paramètres du modèle sans de solides a priori informatifs. Je présente dans mon exposé des modèles autorégressifs de réseau Poisson (PNAR), des analogues statistiques de modèles suceptible-infectieux-guéris (SIR) à temps dis-

recovered (SIR) models, which leverage mechanistic information of disease spread to estimate reporting probability without relying on strong informative priors. Despite the promise of PNAR models, inference in the presence of underreporting poses significant challenges. I will discuss these challenges and present novel, practical Bayesian inference methods tailored for such models. Additionally, I will outline future research directions and advancements expected in the next 5 years.

cret, qui tire parti de l'information mécanistique de la propagation de la maladie pour l'estimation de la probabilité de déclaration sans recourir à de solides a priori informatifs. En dépit de la promesse des modèles PNAR, l'inférence en présence d'une sous-déclaration pose de sérieuses difficultés. Je discuterai de ces défis et présenterai aussi de nouvelles méthodes pratiques d'inférence bayésienne appropriées pour de tels modèles. De plus, je donnerai un aperçu des orientations de la recherche et des percées espérées dans les cinq prochaines années.

[14:30-15:00]

Jason Xu (Duke University)

Data-Augmented MCMC for Stochastic Epidemic Models

MCMC avec données augmentées pour des modèles épidémiques stochastiques

We propose novel data-augmented Markov Chain Monte Carlo strategies to enable exact Bayesian inference under the stochastic susceptible-infected-removed model and its variants. In the incidence data setting, where we are given only discretely observed counts of infection, significant challenges to inference arise due only a partially informative glimpse of the underlying continuous-time process. To account for the missing data while targeting the exact posterior of model parameters, we make use of latent variables that are jointly proposed from surrogates related to branching processes, carefully designed to closely resemble the SIR model. This allows several conditional sampling strategies that make classical MCMC ideas practical, surmounting the intractable observed data likelihood. The method extends to non-Markovian settings as well as tasks such as simultaneous change-point detection under time-varying transmission.

Nous proposons de nouvelles stratégies de Monte Carlo par chaînes de Markov avec données augmentées dans le but de réaliser une inférence bayésienne exacte selon le modèle compartimental stochastique (SIR) et ses variantes. Dans le cas de données d'incidence, dans lequel on a seulement le dénombrement d'infection discrètement observé, de nombreux problèmes importants surviennent pour l'inférence en raison d'un aperçu partiel du processus à temps continu sous-jacent. Pour tenir compte des données manquantes tout en ciblant la distribution a posteriori exacte des paramètres du modèle, nous exploitons des variables latentes qui sont conjointement proposées à partir de substitutions relatives aux processus d'embranchement, soigneusement conçus pour ressembler au modèle SIR. Cela donne plusieurs stratégies d'échantillonnage conditionnelles qui rendent les idées du MCMC pratiques, surmontant l'insoluble vraisemblance de données observées. La méthode s'étend aux cadres non-Markovien et aussi à des tâches comme la détection de points de rupture simultanée selon une transmission variant dans le temps.

Chair/Président: Johanna G. Nešlehová

Room/Salle: C 2033

Date: Monday June 3 / lundi 3 juin

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-13:45]

Jean-François Bégin (Simon Fraser University) **Barbara Sanders** (Simon Fraser University)

Benefit Volatility-targeting Strategies in Lifetime Pension Pools

Stratégies de ciblage de la volatilité de prestations dans les rentes viagères à paiements variables

Lifetime pension pools—also known as group self-annuitization plans and tontines—allow retirees to convert a lump sum into lifelong income, with payouts linked to investment performance and the pool's collective mortality experience. Existing literature has predominantly examined basic investment strategies like constant allocations and investments solely in risk-free assets. Recent studies, however, proposed volatility targeting, aiming to enhance risk-adjusted returns and minimize downside risk. Yet they only considered investment risk in the volatility target, neglecting the impact of mortality risk. This presentation thus aims to address this gap by investigating volatility-targeting strategies for both investment and mortality risks, offering a solution that keeps the risk associated with benefit variation as constant as possible through time. Practical investigations of the strategy demonstrate the effectiveness and robustness of the new dynamic volatility-targeting approach.

Les rentes viagères à paiements variables permettent aux retraités de convertir une somme forfaitaire en un revenu à vie, les versements étant liés à la performance des investissements et à l'expérience collective de la mortalité. La littérature existante a principalement examiné les stratégies d'investissement de base, telles que les allocations constantes et les investissements uniquement dans des actifs sans risque. Des études récentes ont toutefois proposé un ciblage de la volatilité, visant à améliorer les rendements ajustés au risque et à minimiser le risque de perte. Cependant, elles n'ont pris en compte que le risque d'investissement dans l'objectif de volatilité, négligeant l'impact du risque de mortalité. Cette présentation vise donc à combler cette lacune en étudiant les stratégies de ciblage de la volatilité pour les risques d'investissement et de mortalité, offrant une solution qui maintient le risque associé à la variation des prestations aussi constant que possible.

[13:45-14:00]

Kyran Cupido (Kyran Cupido) **Petar Jevtic** (Arizona State University) **Luca Regis** (University of Torino) **Kenneth Zhou** (Arizona State University)

Spatial Natural Hedging – A General Framework with Application to the Mortality of U.S. States

Couverture naturelle spatiale — un cadre général appliqué à la mortalité aux États-Unis

It is well known that coupling life and death benefits within an insurance portfolio may be a beneficial longevity risk reduction technique, especially when policies are underwritten in the same geographical region. Though desirable, the lack of available capacity of life insurance instruments in terms of underlying cohorts or duration of products underwritten within a given region can substantially constrain the use of nat-

Il est bien connu que coupler les prestations d'assurance-vie et de décès dans le contexte d'un portefeuille d'assurance représente une technique avantageuse pouvant réduire le risque de longévité, surtout lorsque les politiques sont émises dans la même région géographique. Malgré ses avantages, le manque de capacité relative aux outils d'assurance-vie en termes de cohortes sous-jacentes et de durée de produit dans une région donnée peut considérablement restreindre les entreprises d'assurance-vie à

ural hedging strategies for life insurance companies. This work investigates the implementation and effectiveness of natural hedging strategies when considering a spatial dimension. Starting from a well-known multipopulation mortality model, we evaluate the relevance of natural hedging strategies and their susceptibility to basis risk resulting from age, period, and spatial effects. In a practical numerical application using U.S. mortality data, we demonstrate the situation of a U.S.-based insurance company capable of selling policies across different states.

adopter des stratégies de couverture naturelle. Ce travail enquête sur l'implantation et l'efficacité des stratégies de couverture naturelle en tenant compte de la dimension spatiale. En commençant par un modèle reconnu de mortalité à multipopulation, nous évaluons la pertinence des stratégies de couverture naturelle et leur sensibilité au risque de base causé par l'âge, la durée et les effets spatiaux. À partir d'une application numérique pratique se servant de données de mortalité, nous démontrons la situation d'une entreprise d'assurance siégeant aux États-Unis et pouvant vendre des politiques dans différents états.

[14:00-14:15]

Barbara Sanders (Simon Fraser University) **Jean-François Bégin** (Simon Fraser University) **Nikhil Kapoor** (Simon Fraser University)

A New Approximation of Annuity Prices for Age–Period–Cohort Models

Nouvelle approximation des coûts des rentes pour des modèles âge-période-cohorte

We present a new general formula for estimating annuity prices within a wide range of stochastic mortality models. The formula is constructed using two building blocks: an approximation technique based on the Wentzel–Kramers–Brillouin method for calculating the sum of correlated lognormal random variables, and an approximate expression for the moment generating function of the lognormal distribution. Notably, this formula is applicable to virtually all age–period–cohort models where period effects are represented by vector autoregressive models. This broad assumption encompasses the majority of existing stochastic mortality models in the literature. Through a numerical illustration, we also demonstrate the reliability and precision of our new method in determining annuity prices.

Nous présentons une nouvelle formule générale pour l'estimation des coûts des rentes dans un large éventail de modèles de mortalité stochastiques. La formule est construite à l'aide de deux blocs constitutifs : une technique d'approximation basée sur la méthode Wentzel-Kramers-Brillouin pour le calcul de la somme des variables aléatoires log-normales corrélées, ainsi qu'une expression d'approximations de la fonction génératrice des moments de la distribution log-normale. À noter que cette formule est applicable à pratiquement tous les modèles âge-période-cohorte dans lesquels les effets de la période sont représentés par des modèles à vecteur autorégressif. Cette hypothèse générale englobe la plupart des modèles de mortalité stochastiques dans la documentation. À l'aide d'une illustration numérique, nous montrons aussi la fiabilité et la précision de notre nouvelle méthode à déterminer les coûts des rentes.

[14:15-14:30]

Dante Mata Lopez (Université du Québec à Montréal (UQAM))

On an optimal dividend problem with a concave bound on the dividend rate

Problème de dividende optimal avec borne concave du taux de dividende

We study a version of De Finetti's optimal dividend problem driven by a diffusion. In our version, the control strategies are assumed to have an absolutely continuous density, which is bounded above by an increasing, concave function. Under mild assumptions on the drift and diffusion coefficients, we provide sufficient conditions to show that an optimal strategy exists and lies within the set of generalized refraction strategies. In addition, we are able to characterize the optimal refraction threshold in our setting. Joint work with Hélène Guérin,

Notre étude porte sur une version du problème de dividende optimal de De Finetti piloté par une diffusion. Notre version suppose que les stratégies de contrôle ont une densité absolument continue, dont la borne supérieure est une fonction concave croissante. En partant d'hypothèses légères sur les coefficients de dérive et de diffusion, nous fournissons des conditions suffisantes pour montrer qu'une stratégie optimale existe et se trouve dans l'ensemble des stratégies de réfraction généralisées. De plus, nous sommes capables de caractériser le seuil optimal de réfraction dans notre configuration. Travail en collaboration avec Hélène Guérin, Jean-

Jean-François Renaud and Alexandre Roch. François Renaud et Alexandre Roch.

[14:30-14:45]

Hélène Cossette (Université Laval) **Etienne Marceau** (Université Laval) **Benjamin Côté** (Université Laval)

Risk Models Defined on a Family of Tree-Based Markov Random Fields with Poisson Marginals

Modèles de risque définis sur une famille de champs aléatoires de Markov arborescents avec des distributions marginales de Poisson

A new family of tree-based Markov random fields for a vector of discrete counting random variables is presented. Within this family, the marginal distributions are Poisson and the dependence structure is encrypted on a tree. The proposed family of tree-based Markov random fields for a vector of discrete counting Poisson random variables has many advantages, such as being able to find explicit expressions for the joint probability mass function, the joint probability generating function and to study the dependence properties within this family.

Nous présentons une nouvelle famille de champs aléatoires de Markov arborescents pour un vecteur de variables aléatoires de comptage discrètes. Dans cette famille, les distributions marginales sont de Poisson et la structure de dépendance est cryptée sur un arbre. La famille proposée de champs aléatoires de Markov arborescents pour un vecteur de variables aléatoires de Poisson à comptage discret présente de nombreux avantages, tels que la possibilité de trouver des expressions explicites pour la fonction de masse de probabilité conjointe, la fonction génératrice de probabilité conjointe et d'étudier les propriétés de dépendance au sein de cette famille.

[14:45-15:00]

Etienne Marceau (Université Laval)

Risk Sharing and Risk Allocation: Generating Function Approach

Partage et répartition des risques : méthode de génération de fonctions

Consider a risk portfolio with the aggregate loss random variable S defined as the sum of the n individual loss random variables (rvs) X1, . . . , Xn. Expected allocations are essential for risk-sharing and risk allocation. One uses expected allocations for computing contributions under the conditional mean risk-sharing rule and contributions under the Euler risk allocation principle. This paper introduces an ordinary generating function for expected allocations for each rv X1, . . . , Xn, given aggregate loss random variable S, assuming X1, . . . , Xn are discrete rvs. Using properties of ordinary generating functions, we obtain closed-formed solutions to risk allocation problems. We present an efficient algorithm to recover expected allocations using FFT algorithm. Ultimately, it provides a new practical tool to efficiently compute contributions under the conditional mean risk-sharing rule and contributions under the Euler risk allocation principle.

Envisageons un portefeuille de risques avec une variable aléatoire de perte globale S définie comme la somme des variables des n pertes individuelles (rvs) X1, . . . , Xn. Les allocations attendues sont essentielles pour la répartition et le partage des risques. On les utilise pour calculer les contributions selon la règle de partage des risques moyen conditionnel et selon le paradigme d'allocation des risques d'Euler. Cet article présente une fonction de génération ordinaire pour les allocations attendues pour chaque rv X1, . . . , Xn, étant donné une variable aléatoire de perte globale S, en supposant que les risques X1, . . . , Xn sont discrets. Premièrement, nous fournissons une relation simple entre la fonction génératrice ordinaire pour les allocations attendues et la fonction génératrice de probabilité. En exploitant les propriétés des fonctions génératrices ordinaires, nous révélons de nouvelles solutions fermées aux problèmes d'allocation des risques. Ensuite, nous présentons un algorithme efficace pour récupérer les allocations attendues en utilisant la transformée de Fourier rapide, fournissant ainsi un nouvel outil pratique pour calculer efficacement les allocations attendues, selon la règle de partage des risques moyen conditionnel et selon le paradigme d'allocation des risques d'Euler.

Chair/Président: Yunhong Lyu

Room/Salle: C 4036

Date: Monday June 3 / lundi 3 juin

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-13:45]

Yi Meng Chang (University of Toronto Dalla Lana School of Public Health) **Petros Pechlivanoglou** (The Hospital for Sick Children) **Olli Saarela** (University of Toronto) **Eleanor M. Pullenayegum** (The Hospital for Sick Children)

Issues and Bias in Phase-of-Care Costing When Estimating Attributable Healthcare Costs of a Disease

Problèmes et biais relatifs au calcul des coûts par phase de soin lors de l'estimation des coûts de soins de santé imputables à une maladie

Attributable healthcare costs of a disease are used to guide healthcare spending priorities. Phase-of-care costing, a popular costing method, divides care into clinically relevant phases, i.e. initial phase post diagnosis, continuing phase for stable disease, and terminal phase before death. This method deploys matching of cases and controls to estimate disease attributable healthcare costs; however, it is not based on principles of causal inference. Using a target trial framework we identified two methodological flaws: firstly, selection bias in the control subjects, whereby the controls die increasingly early as the disease prevalence increases; and secondly, bias due to conditioning on a mediating variable when cases and controls are matched on death age in the terminal phase. Simulations were used to show the magnitude of bias. By highlighting issues in phase-of-care costing, we suggest exploring alternative estimation methods that are based on principles of causal inference.

Les coûts de soins de santé imputables à une maladie servent à orienter les priorités en matière de dépenses en soins de santé. Le calcul des coûts par phase de soins, une méthode courante, divise les soins en phases cliniquement pertinentes. Par exemple : la phase initiale après un diagnostic, la phase continue pour une maladie stable et la phase terminale avant le décès. Cette méthode utilise l'appariement des cas et des témoins afin d'estimer les coûts de soins de santé imputables à une maladie, mais elle ne se base pas sur les principes de l'inférence causale. En utilisant un cadre d'essai ciblé, nous avons découvert deux défauts méthodologiques : le premier est le biais de sélection des sujets témoins, car les témoins décèdent de plus en plus tôt quand la prévalence de la maladie augmente ; et le deuxième est le biais dû au conditionnement sur une variable médiatrice lorsque les cas et les témoins sont appariés sur l'âge du décès dans la phase terminale. Nous avons employé des simulations pour démontrer l'envergure des biais. En soulignant les problèmes du calcul de coût par phase de soin, nous suggérons d'explorer d'autres méthodes d'estimation basées sur les principes d'inférence causale.

[13:45-14:00]

Tessa Reimer (University of Manitoba) **Alexandre Leblanc** (University of Manitoba) **Saman Muthukumarana** (University of Manitoba)

Bayesian Analysis of Batting Outcomes from Major League Baseball Using a Nested Dirichlet Prior Distribution

Analyse bayésienne des succès au bâton de la Ligue majeure de baseball à l'aide d'une distribution a priori de Dirichlet imbriquée

Bayesian methods and sports analytics have become popular areas of research, but their intersection remains under-considered in Statistics. While the Nested Dirichlet distribution has been proposed as a conjugate prior to Multinomial data commonly produced in sports set-

Les méthodes bayésiennes et les analyses sportives sont devenues des domaines de recherche populaires, mais leur intersection est sous-estimée en statistique. Même si la distribution a priori de Dirichlet imbriquée a été proposée comme a priori conjuguée aux données multinomiales couramment produites dans

tings, many properties of this distribution and its generalizability have not been explored. In this presentation, we propose a different parameterization for the Nested Dirichlet distribution and explain how it can be used to derive the posterior and posterior predictive distributions for the Multinomial Nested Dirichlet model. We also demonstrate the generalizability of our parameterization, an important improvement over the approaches that can be found in the literature. Finally, we will also present how we used this model to analyze 2023 MLB batting outcome data to produce model-based batting metrics with appropriate uncertainty quantification. It was found that our model provides an adequate fit to the batting data.

un contexte sportif, plusieurs propriétés de cette distribution et sa généralisabilité n'ont pas été étudiées. Nous proposons une paramétrisation différente de la distribution de Dirichlet imbriquée et expliquons comment celle-ci peut être utilisée pour dériver les distributions a posteriori et prédictives a posteriori du modèle multinomial de Dirichlet imbriquée. Nous montrons également la généralisabilité de notre paramétrisation, une amélioration importante par rapport aux approches qui figurent dans la littérature. Nous concluons en présentant notre façon d'appliquer ce modèle à l'analyse des données de succès au bâton de la Ligue majeure de baseball de 2023 afin de produire des mesures au bâton avec une quantification appropriée de l'incertitude. Il a été constaté que notre modèle fournit un ajustement adéquat aux données au bâton.

[14:00-14:15]

Martin Lysy (University of Waterloo)

PFJAX: Differentiable Particle Filtering in Python

PFJAX : Filtrage de particule différentiable dans Python

State-space models are commonly used to describe a wide range of scientific phenomena. In order to estimate the parameters of these models, the latent states are typically integrated over using particle filtering techniques. However, due to their computationally intensive nature, particle filters are difficult to implement without sacrificing simplicity for performance. Here we present PFJAX, a Python library for particle filtering which attempts to provide the best of both. This is achieved by implementing several particle filters and resamplers in JAX, a near drop-in replacement for NumPy which offers both (i) a high-performance just-in-time (JIT) compiler and (ii) a simple and powerful automatic differentiation engine. The latter is used to implement various differentiable particle filters, enabling the use of efficient gradient-based algorithms for parameter inference. The use and usefulness of PFJAX is demonstrated on several examples of state space models.

Les modèles spatio-temporels servent couramment à décrire un large éventail de phénomènes scientifiques. Afin d'estimer les paramètres de ces modèles, les états latents sont généralement intégrés grâce à des techniques de filtrage de particule. Cependant, en raison de la nature intensive de leur calcul, les filtres de particule sont difficiles à intégrer sans sacrifier la simplicité pour la performance. Nous présentons ici PFJAX, une bibliothèque Python pour le filtrage de particule qui tente de tirer le meilleur des deux parties. Cela est possible en intégrant plusieurs filtres de particules et de rééchantilloneurs dans JAX, un quasi-remplaçant pour NumPy qui offre à la fois un compilateur « juste à temps » (JIT) de haute performance et un moteur de différentiation automatique simple et puissant. Ce dernier sert à implanter plusieurs filtres de particule différentiables, ce qui permet l'exploitation d'algorithmes efficaces basés sur un gradient pour l'inférence de paramètre. Nous démontrons l'utilisation et l'utilité de PFJAX à partir de plusieurs exemples de modèles spatio-temporels.

[14:15-14:30]

Jeffrey Negrea (University of Waterloo) **Jun Yang** (University of Copenhagen) **Haoyue Feng** (Boston University) **Daniel Roy** (University of Toronto) **Jonathan Huggins** (Boston University)

Statistical Inference with Stochastic Gradient Algorithms

Inférence statistique avec algorithmes de gradient stochastique

Recent work in machine learning has couched stochastic gradient algorithms as methods for asymptotic posterior sampling based on heuristic arguments. Our work puts those heuristics on rigorous foundations, fully characterizing the asymptotic properties of stochastic

Un travail de recherche récent en apprentissage automatique a formulé des algorithmes de gradient stochastique comme méthodes d'échantillonnage a posteriori asymptotique sur la base d'arguments heuristiques. Nous avons établi pour ces arguments heuristiques des fondements rigoureux, en caractérisant complètement

gradient methods used as sampling algorithms. We present a functional Bernstein–von Mises-like theorem for the scaling limit of the paths of stochastic gradient algorithms, showing they converge to an Ornstein-Uhlenbeck process as the number of observations increases. This approach validates some previous hypotheses while falsifying others. Using the large sample asymptotics, we demonstrate how to properly tune SGAs for various desiderata, including matching the asymptotics of the posterior distribution, the bagged posterior, and the MLE.

les propriétés asymptotiques des méthodes de gradient stochastique utilisées comme algorithmes d'échantillonnage. Nous présentons un théorème fonctionnel de type Bernstein-von Mises pour la limite d'échelonnage des voies des algorithmes de gradient stochastique, montrant qu'elles convergent vers un processus Ornstein-Uhlenbeck à mesure que le nombre d'observations augmente. Cette approche valide certaines hypothèses antérieures et en invalide d'autres. En utilisant les propriétés asymptotiques d'échantillon de grande taille, nous montrons comment affiner adéquatement l'ajustement du gradient symplectique (SGA) pour divers desiderata, y compris la correspondance entre les propriétés asymptotiques de la distribution a posteriori, les "bagged posteriors" et l'estimateur du maximum de vraisemblance (MLE).

[14:30-14:45]

Lulu Zhang (University of New Brunswick) **Renjun Ma** (University of New Brunswick) **Guohua Yan** (University of New Brunswick) **Xifen Huang** (Yunnan Normal University)

A New Logistic Model with Subject-Specific and Serially Correlated Time-Specific Distribution-Free Random Effects on the Unit Interval for Intensive Longitudinal Binary Data

Nouveau modèle logistique avec effets aléatoires de distribution libre à sujet spécifique et à temps spécifiques sériellement corrélés sur l'intervalle unité pour des données binaires longitudinales à forte intensité

Various beta-binomial mixed effects models have been developed in recent years for longitudinal binary data; however, these approaches rely heavily on the parametric specification of beta and normal random effects. Furthermore, their incorporation of normal random effects into beta-binomial models has been done at the sacrifice of certain computational convenience and clear interpretation with beta-binomial models. In this work, we introduce a new model that incorporates subject-specific and serially correlated time-specific distribution-free random effects on the unit interval into logistic regression multiplicatively with fixed effects. This new multiplicative model facilitates interpretation of random effects on the unit interval as risk modifiers. This multiplicative model setup also eases the model derivation and random effects prediction. A quasi-likelihood approach has been developed in the estimation of our model. Our results are robust against random effects distributions.

Ces dernières années, divers modèles bêta-binomiaux à effets mixtes ont été développés pour les données binaires longitudinales. Ces approches s'appuient toutefois grandement sur la spécification paramétrique des effets aléatoires bêta et normaux. De plus, l'incorporation d'effets aléatoires normaux dans des modèles bêta-binomiaux se fait au détriment d'une certaine commodité computationnelle et d'une interprétation claire obtenues avec les modèles bêta-binomiaux. Nous présentons un nouveau modèle qui incorpore à une régression logistique multiplicative à effets fixes des effets aléatoires de distribution libre à sujet spécifique et à temps spécifiques sériellement corrélés sur l'intervalle unité. Ce nouveau modèle multiplicatif facilite l'interprétation des effets aléatoires sur l'intervalle unité comme modificateurs du risque. La configuration de ce modèle facilite également la dérive du modèle et la prédiction des effets aléatoires. Une approche de quasi-vraisemblance a été élaborée pour l'estimation de notre modèle. Nos résultats sont robustes contre les distributions des effets aléatoires.

[14:45-15:00]

Dingding Hu (University of Waterloo)

ROC Curve Analysis under Non-ignorable Verification Bias

Analyse de la courbe ROC en présence d'un biais de vérification non négligeable

The Receiver Operating Characteristic (ROC) curve is widely used as a statistical tool to exhibit the diagnostic ability of a binary classifier. In diagnostic accuracy stud-

La courbe ROC (Receiver Operating Characteristic) est largement utilisée comme outil statistique pour démontrer la capacité de diagnostic d'un classificateur binaire. Dans les études de précision

ies, the status of the patient, disease or non-disease, often serves as the outcome in the statistical model. However, to estimate the ROC curve, it is a practical challenge that the disease status of some patients remain unverified. In this paper, we focus on the ROC curve analysis under non-ignorable verification bias. We consider parametric models for both disease and verification status, and estimate the parameters through a maximum likelihood approach.We derive the point estimate for the Area Under the ROC Curve (AUC) and its asymptotic distribution, from which we construct the confidence interval. Simulation studies are ran to compare our proposed method to other competitive methods and to test the validity of the confidence interval estimation.

diagnostique, l'état du patient, malade ou non, sert souvent de réponse dans le modèle statistique. Cependant, pour estimer la courbe ROC, le fait que l'état pathologique de certains patients ne soit pas vérifié constitue un défi pratique. Dans cet article, nous nous concentrons sur l'analyse de la courbe ROC en présence d'un biais de vérification non négligeable. Nous considérons des modèles paramétriques pour l'état pathologique et de vérification et estimons les paramètres par une approche de maximum de vraisemblance. Nous dérivons l'estimation ponctuelle de l'aire sous la courbe ROC (AUC) et sa distribution asymptotique, à partir de laquelle nous construisons l'intervalle de confiance. Nous menons des études de simulation pour comparer la méthode proposée à d'autres méthodes concurrentes et pour tester la validité de l'estimation de l'intervalle de confiance.

Chair/Président: Wei Liu

Room/Salle: C 3053

Date: Monday June 3 / lundi 3 juin

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-13:45]

Zeinab Mashreghi (University of Winnipeg)

Bootstrap Resampling Methods for Survey Data in R

Méthodes de rééchantillonnage bootstrap pour des données d'enquête avec R

Bootstrap resampling methods have attracted significant attention in the estimation of variance for estimators derived from survey data. While numerous bootstrap techniques have been proposed to address various design features like stratification and clustering, there remains a need for practical implementation tools. To bridge this gap, I introduce an R package tailored for implementing bootstrap techniques in survey data analysis. This package includes a collection of specialized functions, crafted to accommodate diverse survey sampling scenarios, including stratified simple random sampling and stratified two-stage cluster sampling. By utilizing these tools, researchers can obtain accurate bootstrap variance estimates for the cases of population totals, means, and quartiles, and generate bootstrap samples for thorough analyses, ensuring reliable conclusions from survey data.

Une grande attention a été portée aux méthodes de rééchantillonnage bootstrap pour l'estimation de la variance des estimateurs dérivés de données d'enquête. Même si bon nombre de techniques bootstrap ont été proposées pour aborder diverses caractéristiques de conception comme la stratification et le regroupement, on a encore besoin d'outils pratiques d'implémentation. Pour combler cette lacune, je présente un package R conçu sur mesure pour implémenter des techniques bootstrap dans l'analyse de données d'enquête. Ce package comprend une série de fonctions spécialisées destinées à s'inscrire dans différents scénarios d'échantillons d'enquête, y compris l'échantillonnage aléatoire simple stratifié et l'échantillonnage stratifié en grappes à deux étapes. En utilisant ces outils, les chercheurs peuvent obtenir une estimation exacte de la variance bootstrap dans les cas de totaux, moyennes et quartiles de population et générer des échantillons bootstrap pour des analyses rigoureuses, assurant ainsi la fiabilité des conclusions tirées des données d'enquête.

[13:45-14:00]

Gradon Nicholls (University of Waterloo)

Model-Assisted Double-Coding of Open-Ended Survey Questions with Large Language Models

Double codage assisté par modèle de questions d'enquête ouvertes à l'aide de grands modèles de langage

Open-ended questions allow survey respondents to provide answers in their own words, but the resulting textual data often needs to be manually coded to facilitate analysis. Having two independent coders ("double-coding") has been proposed to improve coding quality, but disagreements among the two coders must then be resolved (e.g. through discussion or a third coder). The objective of the current research is to determine to what extent double-coding can improve coding quality when one of the human coders is replaced by model predictions. Clearly, it is not sufficient for the machine coder

Les questions ouvertes permettent aux personnes interrogées de répondre avec leurs propres mots, mais les données textuelles qui en résultent doivent souvent être codées manuellement pour faciliter l'analyse. Il a été proposé d'utiliser deux codeurs indépendants (« double codage ») pour améliorer la qualité du codage, mais les écarts entre les deux codeurs devaient ensuite être résolus (p. ex., par une discussion ou un troisième codeur). L'objectif de la présente étude est de déterminer dans quelle mesure le double codage peut améliorer la qualité du codage lorsque l'un des codeurs humains est remplacé par des prédictions de modèles. De toute évidence, le codeur machine ne doit pas se contenter d'imi-

to exactly mimic the human coder's behaviour–that is, it must (at least occasionally) catch mistakes made by the human coder. In this context, we investigate the utility of pre-trained Large Language Models (LLMs) combined with text data augmentation techniques to increase the number of examples seen during model training.

ter exactement le comportement du codeur humain. En effet, il doit (au moins occasionnellement) repérer les erreurs commises par le codeur humain. Dans ce contexte, nous analysons l'utilité des grands modèles de langage pré-entraînés associés à des techniques d'augmentation des données textuelles afin d'accroître le nombre d'exemples observés au cours de l'entraînement du modèle.

[14:00-14:15]

Michael John Ilagan (McGill University) **Carl F. Falk** (McGill University)

A Mixture Model for p-values to Detect Bots in Likert-type Questionnaire Data

Un modèle de mélange pour les valeurs p afin de repérer les robots dans des données de questionnaire de type Likert

In the social sciences, compensating online participants to answer questionnaires is a common practice. However, such practice incentivizes rapid completion of many questionnaires, introducing the risk of "random" responding, such as by bots. Toward safeguarding data quality, we consider the problem of classifying respondents as random vs. non-random, unsupervised. In prior work, for Likert-type response data, we proposed a sensitivity-calibrated classifier based on a permutation test on bias-corrected outlier statistics. While sensitivity calibration is valuable, it does not imply high accuracy. In the present work, to improve classification accuracy, we propose a mixture model for permutation test p-values. In a simulation study, we sampled non-random responders from a real dataset and introduced random responders. The Bayes classifier from our fitted mixture model was more accurate than the 95% sensitivity calibrated classifier when random responders were the minority.

En sciences sociales, il est courant d'indemniser des participants en ligne pour répondre à un questionnaire. Cependant, de telles pratiques encouragent un remplissage rapide de nombreux questionnaires, ce qui comporte le risque de réponse « aléatoire », par des robots par exemple. Afin de garantir la qualité des données, nous abordons le problème de classification non supervisé des répondants entre les aléatoires et les non aléatoires. Dans un travail antérieur, pour des données de réponse de type Likert, nous avons proposé un classifieur calibré selon la sensibilité basé sur un test de permutation sur des statistiques de valeurs aberrantes corrigées pour le biais. Bien que la calibration de la sensibilité de calibration soit utile, elle n'implique pas une précision élevée. Dans le cadre de ce travail, nous proposons un modèle de mélange pour les valeurs p de test de permutation, afin d'améliorer la précision de classification. Dans le cadre d'une étude de simulations, nous avons échantillonné des répondants non aléatoires à partir d'un ensemble de données et y avons inséré des répondants aléatoires. Le classificateur bayésien de notre modèle de mélange ajusté s'est avéré plus précis que le classificateur calibré avec une sensibilité de 95 % lorsque les répondants aléatoires étaient minoritaires.

[14:15-14:30]

Atefeh Kheirollahi (Memorial University of Newfoundland) **Nan Zheng** (Memorial University of Newfoundland) **Yildiz Yilmaz** (Memorial University of Newfoundland)

Accounting for Age Measurement Errors in Fish Growth Model Estimation using Length-stratified Age Sampling Data

Prise en compte des erreurs de mesure de l'âge dans l'estimation du modèle de croissance de poissons à l'aide de données d'échantillonnage par âge stratifiées selon la longueur

Fish growth models are commonly estimated using length-at-age data. This data is widely gathered through length-stratified age sampling (LSAS), a response-selective sampling scheme. The data may contain age measurement errors (MEs). We propose a methodology that accounts for LSAS and age MEs to enhance the precision of fish growth estimates. The proposed methods use the empirical proportion likelihood for LSAS and the structural errors in variables for age MEs. To ensure the reliability of our estimates, we provide a measure

Les modèles de croissance de poissons sont généralement estimés à l'aide de données sur la longueur en fonction de l'âge. Ces données sont largement collectées par le biais d'échantillons par âge stratifiés selon la longueur (EASL), un plan d'échantillonnage à réponse sélective. Or ces données peuvent contenir des erreurs de mesure de l'âge (EM). Nous proposons une méthodologie qui prend en compte les EASL et les EM de l'âge afin d'améliorer la précision des estimations de croissance de poissons. Les méthodes proposées utilisent la proportion de vraisemblance empirique pour les EASL et les erreurs structurelles sur les variables pour

of uncertainty for parameter estimates and standardized residuals for model validation. We employ a continuation ratio-logit model to model the age distribution, which is consistent with the random nature of the true age distribution. The simulation study shows that neglecting age MEs can lead to significant bias in growth estimation. However, our new approach performs well regardless of the magnitude of age MEs and accurately estimates standard errors of parameter estimates. Real data analysis demonstrates the effectiveness of the proposed methods.

les EM de l'âge. Pour garantir la fiabilité de nos estimations, nous fournissons une mesure de l'incertitude pour les estimations des paramètres et des résidus standardisés pour la validation du modèle. Nous utilisons un modèle ratio-logit de continuation pour modéliser la distribution de l'âge, ce qui est cohérent avec la nature aléatoire de la véritable distribution de l'âge. L'étude de simulation montre que le fait de négliger les EM de l'âge peut entraîner un biais important dans l'estimation de la croissance. Cependant, notre nouvelle approche donne de bons résultats quelle que soit l'ampleur des EM de l'âge et estime avec précision les erreurs standard des estimations des paramètres. L'analyse de données réelles démontre l'efficacité des méthodes proposées.

Teaching Statistics With a Data-Centric Perspective
Enseigner la statistique dans une perspective centrée sur les données

Chair/Président: Harsha Harsha Perera

Room/Salle: ED 2018B

Date: Monday June 3 / lundi 3 juin

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-13:45]

Michael Wallace (University of Waterloo)

Real-world Data Analysis in an Introductory Statistics Course: Assignments using Data from the Stanford Open Policing Project

Analyse de données réelles dans un cours d'introduction aux statistiques : travaux réalisés à l'aide des données du Stanford Open Policing Project

Crafting engaging assignments for statistics students is challenging, and can be especially so at the introductory level. While students routinely express an interest in gaining 'real-world' data analysis experience, this is often difficult to achieve within the constraints of formal assessments for a large enrolment. There is also increasing demand for teaching to address equity issues, such as those relating to sex, gender, and racial identities. In this talk, we will describe the use of data from the Stanford Open Policing Project in assignments for a large introductory statistics course at the University of Waterloo. We will discuss how the dataset was prepared with each student given a unique sample, assignment structure and grading logistics, the opportunities for enhanced learning these data offer, as well as the limitations and setbacks encountered along the way. A slide deck, and links to a sample dataset, will be available at https://mpwallace.github.io prior to the talk.

Il est particulièrement difficile de concevoir des travaux stimulants pour les étudiants suivant un cours d'introduction aux statistiques. Bien que les étudiants manifestent régulièrement leur intérêt pour acquérir une expérience d'analyse de données du « monde réel », peu d'entre eux y parviennent en raison des contraintes imposées par les examens officiels. De plus, la demande pour que l'enseignement aborde les questions d'équité (liées au sexe, au genre et aux identités raciales) est de plus en plus forte. Dans cette présentation, nous décrirons l'utilisation des données du Stanford Open Policing Project dans les travaux d'un cours d'introduction aux statistiques à la University of Waterloo. Nous présenterons la façon dont l'ensemble des données a été préparé (chaque étudiant recevant un échantillon unique), la structure des travaux et la logistique de l'évaluation, les possibilités d'amélioration de l'apprentissage qu'offrent ces données, ainsi que les limites et les échecs observés en cours de route. Avant la conférence, un jeu de diapositives et des liens vers un échantillon de données seront disponibles au https ://mpwallace.github.io.

[13:45-14:00]

Katherine Daignault (University of Toronto)

In-Depth Exploration of Linear Regression Concepts through Self-Paced LearnR Modules

Explorations des concepts de régression linéaire à travers les modules LearnR

Methods of Data Analysis 1 is a required course for the nearly 4000 undergraduates enrolled in the statistics program at the University of Toronto. The course addresses the theory and application of linear regression analysis. Dramatic variation exists among the students in their experience with inference and programming, creating a steep learning curve for certain students. To support students in strengthening these skills, learnR

Méthodes d'analyse de données 1 est un cours obligatoire pour les près de 4000 étudiants de premier cycle inscrits au programme de statistiques à l'Université de Toronto. Le cours aborde l'analyse de régression linéaire, tant sur le plan théorique que dans son application à l'aide de R. Les compétences prérequises des étudiants en termes d'inférence et programmation varient énormément créant une courbe d'apprentissage abrupte pour certains d'entre eux. Pour soutenir les étudiants dans le ren-

modules were created that merge low-stakes, guided R code practice with illustrations of commonly misunderstood course concepts. All students in the course benefit from completing these modules as they promote an intuitive understanding of the difficult concepts prior to formal theoretical introduction of these topics. Students with a weaker background in programming additionally strengthen their coding skills working their own pace. This session will elaborate on the rationale for this project, demonstrate the modules, and summarize early student feedback.

forcement de ces compétences, des modules «learnR» ont été développés, combinant des exercices guidés de code R avec de la pratique des concepts du cours fréquemment mal compris. Ces modules comptent pour peu de la note finale afin de ne pas pénaliser les étudiants avec plus faible prérequis. Tous les étudiants du cours bénéficient de la réalisation de ces modules, car ils favorisent une compréhension intuitive des concepts difficiles avant l'introduction théorique de ces sujets. Les étudiants ayant une faible expérience en programmation peuvent renforcer leurs compétences en programmation à leur propre rythme tout en progressant dans le cours. Cette session discutera davantage de l'implantation de ce projet, tout en présentant des exemples et résumera les premières rétroactions des étudiants.

[14:00-14:15]

Wanhua Su (MacEwan University)

Applying Statistical Learning Methods to Complex Survey Data

Application des méthodes d'apprentissage statistique aux données d'enquêtes complexes

Statistical learning methods are gaining popularity in handling complex survey data due to their attractive strengths in predictive accuracy, feature selection, and modeling flexibility. Complex survey datasets are characterized by thousands of observations, hundreds of variables, a stratified and/or clustered structure, missing values, invalid/valid skips, and associated weights. All these features make applying statistical learning methods challenging. In this talk, we will share some practical guidelines for applying statistical learning techniques to complex survey data in dimension reduction, feature selection, dealing with missing values and valid skips, and extremely imbalanced class distribution. The effects of incorporating sampling weights in tree-based models will be illustrated by simulations and two national complex surveys.

Les méthodes d'apprentissage statistique sont de plus en plus populaires pour le traitement des données d'enquêtes complexes en raison de leurs avantages en terme de précision, de sélection de variables et de flexibilité dans la modélisation. Les ensembles de données d'enquêtes complexes se caractérisent par des milliers d'observations, des centaines de variables, une structure stratifiée et/ou en grappes, des valeurs manquantes, des sauts invalides/valides et des pondérations associées. Toutes ces caractéristiques rendent l'application des méthodes d'apprentissage statistique difficile. Dans cet exposé, nous partagerons quelques indications pratiques pour l'application des techniques d'apprentissage statistique aux données d'enquête complexes en ce qui concerne la réduction de la dimension, la sélection des variables, le traitement des valeurs manquantes et des sauts valides, et la distribution extrêmement déséquilibrée des classes. Nous illustrerons les effets de l'incorporation du poids d'échantillonnage dans les modèles basés sur les arbres par des simulations et une application à deux enquêtes nationales complexes.

[14:15-14:30]

Bethany J.G. White (University of Toronto) **Jastaranpreet Singh** (University of Toronto)

Going Hybrid: Transforming an Introductory Statistics Course for Life Sciences

Enseignement hybrique : transformer un cours d'introduction à la statistique pour les sciences de la vie

A statistics course was co-developed by a statistician and an immunologist at the University of Toronto in 2018 and it has been team-taught from this multidisciplinary perspective since. The course learning outcomes, activities and assessments align with the needs of participating life sciences programs and were informed by evidence-based practices in statistics and life

Un cours de statistique a été élaboré conjointement par un statisticien et un immunologiste à l'Université de Toronto en 2018, et il est enseigné depuis en équipe dans cette perspective multidisciplinaire. Les résultats d'apprentissage, les activités et les évaluations du cours s'alignent sur les besoins des programmes de sciences de la vie qui y participent et s'appuient sur des pratiques éprouvées en matière d'enseignement de la statistique et des sciences de la vie

science education (see Tong et al., 2022). When the course was abruptly converted from in-person to online in 2020 in response to the pandemic, we found some of the learning activities and assessments worked quite well, and perhaps even better, online. Therefore, we recently redesigned this course with a mix of online and in-person components to leverage the strengths of online and in-person formats and enhance students' learning experiences. In this talk, I will share our experience with the move from in-person to hybrid, outline the course design, and discuss what went well, and what we will reconsider for next time.

(voir Tong et al., 2022). Lorsqu'en 2020, en réponse à la pandémie, le cours a été brusquement converti d'un cours en personne à un cours en ligne, nous avons constaté que certaines des activités d'apprentissage et des évaluations fonctionnaient très bien, voire mieux, en ligne. C'est pourquoi nous avons récemment remanié ce cours en y associant des éléments en ligne et en présentiel, afin de tirer parti des atouts des deux formats et d'améliorer l'expérience d'apprentissage des étudiants. Dans cet exposé, je vous ferai part de notre expérience du passage d'un cours en présentiel à un cours hybride, je vous présenterai la conception du cours, je discuterai des points positifs et de ceux que nous reconsidérerons la prochaine fois.

[14:30-14:45]

Samantha-Jo Caetano (University of Toronto) **Emily Somerset** (University of Toronto) **Andrea Portt** (University of Toronto)

Flexible Deadlines for Written Assessments

Échéances flexibles pour les évaluations écrites

Stress levels of post-secondary students are on the rise as students balance adjustment to post-secondary life and academics (Linden, 2020, Linden, 2022). Moreover, students in statistics courses tend to feel anxious when tasked with communication assessments, as their training is often self-viewed as more technical and objective (Wilkins, 2015). In hopes of reducing student stress, a flexible-late-policy was invoked in a large, third-year undergraduate statistics course. A survey was delivered to the students to gauge their feelings and attitudes towards written assessments, the flexible-late-policy, and their usage of generative AI as a support for writing. Results show that 95% (n=336) of students appreciated the flexible-late-policy and 97% (n=341) used the policy at some point. Additionally, subsequent results show that students who used the flexible-late-policy performed marginally worse on assessments and were more likely to use generative AI to support their writing.

Les niveaux de stress chez les étudiants postsecondaire sont à la hausse en raison de l'ajustement dans l'équilibre entre la vie postsecondaire et les études (Linden, 2020, Linden, 2022). De plus, les étudiants en statistiques ont tendance à ressentir de l'anxiété pour les évaluations en communication, car leur formation est souvent autoperçue comme plus technique et objective (Wilkins, 2015). Dans l'espoir de réduire le stress chez les étudiants, une politique de retard flexible fut invoquée dans un grand cours de statistiques de troisième année du premier cycle. Les étudiants ont répondu à une enquête cherchant à évaluer leurs sentiments et leur attitude concernant les évaluations écrites, la politique de retard flexible et leur utilisation de l'IA générative en guise de soutien à l'écrit. Les résultats montrent que 95 % (n=336) des étudiants ont apprécié la politique de retard flexible et que 97 % (n=341) d'entre eux ont profité de la politique. En outre, des résultats ultérieurs démontrent que les étudiants qui ont bénéficié de la politique de retard flexible ont eux des notes légèrement plus faibles aux examens et étaient plus susceptibles d'adopter l'IA générative pour soutenir leur écriture.

[14:45-15:00]

Sohee Kang (University of Toronto Scarborough)

Giving Students Choice: A Flexible Weight for Final Project

Donner le choix aux étudiants : une pondération flexible pour les projets finaux

In statistics education, there has been a notable shift towards collaborative and active learning methodologies. Central to this evolution is the recognition of the role of final projects in maximizing student engagement and mastery of course materials. In this talk, we share our findings from offering final projects as an option in the

En enseignement de la statistique, un virage marqué a été effectué vers les méthodologies d'apprentissage collaboratif et actif. Au cœur de cette évolution se trouve la reconnaissance du rôle des projets finaux pour maximiser l'engagement des étudiants et la maîtrise du contenu des cours. Dans notre exposé, nous partageons ce que nous avons découvert en offrant les projets finaux comme

second course of introductory statistics, investigating whether students who opt to participate in such projects achieve higher grades in the final exam (as evidence of mastery of learning) compared to their counterparts. We also explore the motivations behind students' choices regarding their engagement in these final projects. By understanding these motivations, we hope to tailor instructional strategies to better meet the needs and preferences of students, thereby enhancing overall learning experiences and outcomes.

option au deuxième cours d'introduction à la statistique, afin de savoir si les étudiants qui décident de participer à de tels projets obtiennent de meilleures notes à l'examen final (preuve de la maîtrise de l'apprentissage), comparativement à ceux qui s'en abstiennent. Nous nous intéressons également à ce qui motive le choix des étudiants relativement à leur engagement dans ces projets finaux. En comprenant ces motivations, nous espérons concevoir des stratégies d'enseignement qui répondent mieux aux besoins et préférences des étudiants, ce qui améliorera les expériences et les résultats d'apprentissage en général.

Case Study 1: Examining Graduate Student Perspectives on Quality of Supervision, Program and University Experiences
Étude de cas 1: Examen des perspectives des étudiants diplômés sur la qualité de la supervision, du programme et de l'expérience universitaire

Chair/Président: Chel Hee Lee

Organizer/Responsable: Chel Hee Lee

Room/Salle: CSF Whale Atrium

Date: Monday June 3 / lundi 3 juin

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-15:00]

Amanda Qiu (University of Victoria) **Jingtong Hu** (University of Victoria) **Jindi Huang** (University of Victoria) **Mingyang Chen** (University of Victoria)
University of Victoria 1
Université de Victoria 1

. .

[13:30-15:00]

Hunter Pozzebon (University of Toronto Dalla Lana School of Public Health) **Harieswar Sundaram** (University of Toronto) **Xueer Bi** (University of Toronto) **XiaoXuan Han** (University of Toronto)
University of Toronto 1
Université de Toronto 1

. .

[13:30-15:00]

Aadesh Warren Nunkoo (University of Prince Edward Island) **Paul Alexander Seward** (University of Prince Edward Island) **Mobasherah Falak** (University of Prince Edward Island) **Maleeha Haris** (University of Prince Edward Island)
University of Prince Edward Island
Université de l'Île-du-Prince-Édouard

. .

[13:30-15:00]

Winner Pathak (University of Manitoba) **Avanthi Moragammana Gedara** (University of Manitoba) **Thimani Dananjana Ranathungage** (University of Manitoba) **Jervis Gallanosa** (University of Manitoba)
University of Manitoba 1
Université du Manitoba 1

. .

[13:30-15:00]

Helen Bian (McGill University) **Rubiya Akter** (McGill University) **Qicheng Zhao** (McGill University)
McGill University 1
Université McGill 1

. .

[13:30-15:00]

Case Study 1: Examining Graduate Student Perspectives on Quality of Supervision, Program and University Experiences

Étude de cas 1: Examen des perspectives des étudiants diplômés sur la qualité de la supervision, du programme et de l'expérience universitaire

Sara Haroon (Carleton University) **Christiana Koebel** (Carleton University) **Yuliya Nesterova** (Carleton University)
Carleton University
Université Carleton

. .

[13:30-15:00]

Jingwen Ji (University of Toronto) **Ruiyang Wang** (University of Toronto) **Ruochen Zhao** (University of Toronto) **Yanyue Zhang** (University of Toronto)
University of Toronto 2
Université de Toronto 2

. .

[13:30-15:00]

Larry Dong (University of Toronto Dalla Lana School of Public Health) **George Stefan** (University of Toronto) **Fatema Tuj Johara** (University of Toronto Dalla Lana School of Public Health)
University of Toronto 3
Université de Toronto 3

. .

Case Study 2: Predicting Length of ICU Stay in People with Acute Traumatic Spinal Cord Injury
Étude de cas 2 : Prévision de la durée du séjour en USI des personnes souffrant d'une lésion traumatique aiguë de la moelle épinière

Chair/Président: Chel Hee Lee

Organizer/Responsable: Chel Hee Lee

Room/Salle: CSF Whale Atrium

Date: Monday June 3 / lundi 3 juin

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-15:00]

Alexandra Mossman (University of Waterloo) **Bryn Candles** (University of Waterloo) **Megan French** (University of Waterloo)
University of Waterloo
Université de Waterloo

. .

[13:30-15:00]

Siqi Cheng (McGill University) **Sebastian Garneau** (McGill University) **Yu Gu** (McGill University)
McGill University 2
Université McGill 2

. .

[13:30-15:00]

Alysha Cooper (University of Guelph) **Patrick McMillan** (University of Guelph) **Madeline Ward** (University of Calgary)
University of Guelph / University of Calgary
Université de Guelph / Université de Calgary

. .

[13:30-15:00]

Nasim Feizinazhadgheshlaghi (University of Manitoba) **Elham Afzali** (University of Manitoba) **Funmilola Mary Taiwo** (University of Manitoba) **Bahram Moeinianfar** (University of Manitoba)
University of Manitoba 3
Université du Manitoba 3

. .

[13:30-15:00]

Linke Li (University of Toronto Dalla Lana School of Public Health) **Jasper Zhongyuan Zhang** (University of Toronto) **Ziqian Zhuang** (University of Toronto) **Mei Dong** (University of Toronto)
University of Toronto 4
Université de Toronto 4

. .

[13:30-15:00]

Myron Moskalyk (University of Toronto Dalla Lana School of Public Health) **Zhaoyu Ding** (University of Toronto) **Yan Yi**

Li (University of Toronto) **Jinyu Luo** (University of Toronto)
University of Toronto 5
Université de Toronto 5

. .

[13:30-15:00]

Amin Abed (University of Manitoba) **Md. Hasan** (University of Manitoba) **Narges Amiri** (University of Manitoba) **Justin Dyck** (University of Manitoba)
University of Manitoba 4
Université du Manitoba 4

. .

[13:30-15:00]

Nam-Anh Tran (McGill University) **Kent Lu** (McGill University) **Mingchi Xu** (McGill University)
McGill University 3
Université McGill 3

. .

[13:30-15:00]

Priyonto Saha (University of Toronto Dalla Lana School of Public Health) **Xueying Han** (University of Toronto) **Youxue Ren** (University of Toronto) **Yucheng Jiang** (University of Toronto)
University of Toronto 6
Université de Toronto 6

. .

[13:30-15:00]

Xiao Yan (University of Toronto Dalla Lana School of Public Health) **Yixiao Chen** (University of Toronto)
University of Toronto 7
Université de Toronto 7

. .

[13:30-15:00]

Hao He (University of Ottawa) **Xiao Liang** (University of Ottawa) **Yuewen Pan** (University of Ottawa) **Chang Qu** (University of Ottawa)
University of Ottawa
Université d'Ottawa

. .

[13:30-15:00]

Yacine Marouf (University of Toronto) **Yutong Lu** (University of Toronto) **Hongyan Chen** (University of Toronto) **Haiqi Yang** (University of Toronto)
University of Toronto 8
Université de Toronto 8

. .

[13:30-15:00]

Chen Chen (University of Toronto) **Hongyu Chen** (University of Toronto) **Kiara Wu** (University of Toronto) **Zunaira Mehmood** (University of Toronto)

University of Toronto 9
Université de Toronto 9

. .

[13:30-15:00]

Abdulaziz Sherif (University of Toronto Dalla Lana School of Public Health) **Feifan Xiang** (University of Toronto) **Rachel Yeung** (University of Toronto) **Yankai Feng** (University of Toronto)
University of Toronto 10
Université de Toronto 10

. .

[13:30-15:00]

Brynn O'Connell (MacEwan University) **Alex Lyndon** (MacEwan University) **Adrian Neumann** (MacEwan University) **Stuart Dovey** (MacEwan University)
MacEwan University
Université MacEwan

. .

[13:30-15:00]

Jiachen Pan (University of Western Ontario) **Chengqian Xian** (University of Western Ontario) **Jingyu Tu** (University of Western Ontario) **Xianglong Fu** (University of Western Ontario)
University of Western Ontario 1
Université de Western Ontario

. .

[13:30-15:00]

Ellen Song (Western University) **Yiming Hu** (University of Western Ontario) **Yini Cheng** (University of Western Ontario) **Hanrui Dou** (University of Western Ontario)
University of Western Ontario 2
Université de Western Ontario 2

. .

[13:30-15:00]

Balage Don Harshani Hiranthika De Silva (University of Manitoba) **Samuel Morrissette** (University of Manitoba) **Ashani N. Wickramasinghe** (University of Manitoba) **Nayanthi Karunanayake** (University of Manitoba)
University of Manitoba 2
Université du Manitoba 2

. .

Chair/Président: Chengguo Weng

Organizer/Responsable: Chengguo Weng

Room/Salle: A 1043

Date: Monday June 3 / lundi 3 juin

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-16:00]

Frédéric Godin (Concordia University) **Andrei Neagu** (Concordia University) **Leila Kosseim** (Concordia University) **Clarence Simard** (Université du Québec à Montréal)

Deep Hedging under Imperfect Liquidity

Stratégie de couverture profonde en présence de liquidité imparfaite

This work explores the optimization of hedging strategies for financial options in the presence of imperfect liquidity for the underlying asset. A deep reinforcement learning approach is used to obtain the solution to the problem of minimizing global hedging losses risk under a given illiquidity market impact model. Numerical investigations reveal that the discrepancies between the optimal policy and delta hedging benchmarks are complex and materially driven by various interacting (and sometimes competing) parameters and state variables, such as market depth and impact resilience parameters, the underlying asset drift, the hedging portfolio value, time-to-maturity and past hedging positions. Such complexity highlights the need to rely on sophisticated optimization schemes such as deep reinforcement learning to uncover the optimal policy.

Nous explorons l'optimisation de stratégies de couverture d'options financières en présence d'un avoir ayant une liquidité imparfaite. Un algorithme d'apprentissage par renforcement profond est utilisé afin d'obtenir la solution au problème de couverture incorporant un modèle d'impact de marché. Des études numériques révèlent que les différences entre la police de couverture optimale et les standards de couverture delta sont complexes et matériellement influencés par plusieurs variables d'états et paramètres intéragissant (et parfois compétionnant), tels que les paramètres de profondeur de marché et de résilience des impacts, le taux de croissance du sous-jacent, la valeur du portefeuille de couverture, le temps restant avant maturité et les positions de couverture passées. Une telle complexité illustre la nécessité d'utiliser des méthodes d'optimisation sophistiquées telles que l'apprentissage par renforcement profond afin d'obtenir la police optimale.

[16:00-16:30]

Shu Li (Western University)

Use of Prediction Bias in Active Learning for Variable Annuity Portfolio Valuation

Utilisation d'un biais de prédiction dans l'apprentissage actif pour l'évaluation de portefeuille de rente variable

Active learning, a promising alternative to metamodeling for efficient VA portfolio assessment, allows a predictive regression model to be improved adaptively by augmenting the data for model training with an informative sample. A successful practice of active learning requires an effective metric to measure the informativeness of data. In this talk, we address that prediction bias can be nonneglible in large VA portfolio valuation and investigate the effectiveness of prediction bias

L'apprentissage actif est une option prometteuse par rapport à la métamodélisation pour l'évaluation efficace de portefeuille VA et permet au modèle de régression prédictif de s'améliorer de façon adaptative en augmentant les données pour l'éducation du modèle avec un échantillon informatif. Une pratique réussie de l'apprentissage actif nécessite une mesure efficace pour mesurer la qualité de l'information des données. Lors de cet exposé, nous abordons le fait que le biais de prédiction peut être non négligeable dans l'évaluation d'un grand portefeuille de VA, puis exami-

in the modeling and sampling stages of active learning. Experimental results show that bias-based sampling can be as effective as traditional ambiguity-based sampling, whereas its effectiveness depends on the severeness of bias of the predictive model.

nons l'efficacité du biais de prédiction dans la modélisation et les étapes d'échantillonnage de l'apprentissage actif. Les résultats expérimentaux démontrent que l'échantillonnage basé sur un biais peut être aussi efficace qu'un échantillonnage traditionnel basé sur une ambiguïté, bien que son efficacité dépende de la gravité du biais du modèle prédictif.

[16:30-17:00]

Himchan Jeong (Simon Fraser University) **Hashan Peiris** (Simon Fraser University) **Jae-Kwang Kim** (Iowa State University) **Hangsuck Lee** (Sungkyunkwan University)

Integration of Traditional and Telematics Data for Efficient Insurance Claims Prediction

Intégration de données traditionnelles et télématiques pour la prédiction efficace de réclamations d'assurance

While driver telematics has gained attention for risk classification in auto insurance, scarcity of observations with telematics features has been problematic, which could be owing to either privacy concerns or favorable selection compared to the data points with traditional features. To handle this issue, we apply a data integration technique based on calibration weights for usage-based insurance with multiple sources of data. It is shown that the proposed framework can efficiently integrate traditional data and telematics data and can also deal with possible favorable selection issues related to telematics data availability. Our findings are supported by a simulation study and empirical analysis in a synthetic telematics dataset.

Alors que la télématique des conducteurs a attiré l'attention pour la classification des risques dans l'assurance automobile, la rareté des observations dotées de fonctionnalités télématiques a posé problème, ce qui pourrait être dû soit à des problèmes de confidentialité, soit à une sélection favorable par rapport aux points de données dotés de fonctionnalités traditionnelles. Pour résoudre ce problème, nous appliquons une technique d'intégration de données basée sur des poids d'étalonnage pour une assurance basée sur l'utilisation avec plusieurs sources de données. Il est démontré que le cadre proposé peut intégrer efficacement les données traditionnelles et les données télématiques et peut également traiter d'éventuels problèmes de sélection favorables liés à la disponibilité des données télématiques. Nos résultats sont étayés par une étude de simulation et une analyse empirique dans un ensemble de données télématiques synthétiques.

Presenting your technical work to non-statistical audiences
Présenter son travail technique à des publics non statisticiens

Chair/Président: Eleanor M. Pullenayegum

Organizer/Responsable: Eleanor M. Pullenayegum

Room/Salle: ED 2018A

Date: Monday June 3 / lundi 3 juin

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-16:00]

Aasthaa Bansal (University of Washington)

A Framework for Personalizing the Timing of Surveillance Testing

Cadre de personnalisation d'un calendrier de tests de surveillance

Frequent surveillance testing is recommended and routinely conducted in several disease settings. Although surveillance tests provide information about current disease status and present an opportunity to detect disease progression early, the complications and costs of frequent tests may not be justified for patients who are at low risk of progression. I will discuss our recently developed Personalized Risk-Adaptive Surveillance (PRAISE) framework, a method for embedding dynamic predictions into a sequential decision-making framework to determine the time point at which the next collection of patient data would be most valuable, as well as a preliminary application of the framework to develop more cost-effective surveillance strategies in cystic fibrosis. I will begin with a short talk aimed at statistical audiences, followed by a presentation of the same work targeted to non-statistical audiences, and end with general tips on presenting technical work to non-statistical audiences.

Des tests de surveillance fréquents sont recommandés et effectués systématiquement dans plusieurs contextes pathologiques. Bien que ces tests fournissent des informations sur l'état actuel de la maladie et offrent la possibilité d'en détecter rapidement la progression, les complications et le coût de tests fréquents en dissuader l'utilisation pour les patients qui présentent un faible risque de progression. Je présenterai le cadre de surveillance personnalisée adaptée au risque (PRAISE) que nous avons récemment mis au point, méthode permettant d'intégrer des prédictions dynamiques dans un cadre décisionnel séquentiel afin de déterminer le moment où la prochaine collecte de données sur le patient serait la plus utile, ainsi qu'une application préliminaire de ce cadre à l'élaboration de stratégies de surveillance plus rentables dans le cadre de la mucoviscidose. Je commencerai par un bref exposé destiné à un public de statisticiens, suivi d'une présentation du même travail destinée à un public non statisticien, et je terminerai par des conseils généraux sur la présentation d'un travail technique à un public non statisticien.

[16:00-16:30]

Aya A. Mitani (University of Toronto) **Sean Xinyang Feng** (University of Toronto) **Elizabeth Kaye** (Boston University)

Modelling Time-Varying Risk Factors of Tooth Loss: Results From Joint Model Compared With Extended Cox Regression Model

Modélisation des facteurs de risque de perte de dents variant dans le temps : résultats du modèle conjoint comparés à ceux du modèle de régression de Cox étendu

Periodontal disease is a serious gum infection that can lead to tooth loss. Understanding the risk factors for tooth loss and building clinical prediction models for future tooth loss have been continuing efforts among oral health researchers. However, previous studies have only

La maladie parodontale est une infection grave des gencives qui peut entraîner la perte de dents. Ainsi, les chercheurs en santé bucco-dentaire s'efforcent de comprendre les facteurs de risque de la perte de dents et d'élaborer des modèles de prédiction clinique de la perte future de dents. Cependant, dans les études précédentes,

used covariates obtained from baseline to estimate long-term tooth loss despite the availability of routinely collected data. Therefore, we illustrated the use of joint models for longitudinal and survival data to estimate risk factors for tooth loss as a function of time-varying endogenous periodontal biomarkers. Through a simulation study and application to a longitudinal study of dental disease, we showed that the joint model can incorporate time-varying periodontal biomarkers to accurately estimate the hazard of tooth loss. In this talk, I will first present the work with technical details aimed towards a statistical audience and later present the same work aimed towards a non-statistical audience.

on utilisait uniquement des covariables obtenues à partir de la période initiale pour estimer la perte de dents à long terme, et ce malgré la présence de données régulièrement collectées. Nous avons donc illustré les modèles conjoints des données longitudinales et de survie pour estimer les facteurs de risque de perte de dents en fonction de biomarqueurs parodontaux endogènes variant dans le temps. Nous avons démontré, à l'aide d'une étude de simulation et d'une application à une étude longitudinale des maladies dentaires, que le modèle conjoint peut intégrer des biomarqueurs parodontaux variables dans le temps permettant d'estimer avec précision le risque de perte de dents. Je présenterai d'abord les travaux avec des détails techniques à un public de statisticiens et je les présenterai ensuite de manière plus générale à un public de non-statisticiens.

Embedding Equity, Diversity, and Inclusion in Statistical Research and Practice (Panel)
Intégrer l'équité, la diversité et l'inclusion dans la recherche et la pratique statistiques (Table ronde)

Chair/Président: Michael Wallace

Organizer/Responsable: Michael Wallace

Room/Salle: C 2045

Date: Monday June 3 / lundi 3 juin

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-17:00]

Josée Dupuis (McGill University) **Tolulope Sajobi** (University of Calgary) **Bei Jiang** (University of Alberta)

Embedding Equity, Diversity, and Inclusion in Statistical Research and Practice

Intégrer l'équité, la diversité et l'inclusion dans la recherche et la pratique statistiques

In recent years there has been a substantial increase in awareness of issues pertaining to equity, diversity, and inclusion (EDI). While EDI receives considerable focus across society as a whole, its importance within the field of statistics in particular cannot be understated. Embedding EDI principles in our work is essential to secure funding, design effective studies, conduct accurate and equitable analyses, and produce meaningful, reliable, and ethical conclusions. In this panel, Josée Dupuis, Tolulope Sajobi, and Bei Jiang will discuss how they approach research through an EDI lens, ranging from applying for funding, study design and data collection, to analysis and dissemination of results. The panel will focus on providing tangible and actionable advice, as well as offering broader principles that should underly all statistical research and practice.

Ces dernières années, les questions relatives à l'équité, à la diversité et à l'inclusion (EDI) ont revêtu une importance accrue. Alors que l'EDI reçoit une attention considérable dans l'ensemble de la société, son importance dans le domaine de la statistique en particulier ne peut être sous-estimée. Il est essentiel d'intégrer les principes d'EDI dans notre travail pour obtenir des financements, concevoir des études efficaces, mener des analyses précises et équitables et produire des conclusions significatives, fiables et éthiques. Dans cette table ronde, Josée Dupuis, Tolulope Sajobi et Bei Jiang expliqueront comment ils abordent la recherche sous l'angle de l'EDI, depuis la demande de financement jusqu'à l'analyse et la diffusion des résultats, en passant par la conception des études et la collecte de données. La session s'attachera à fournir des conseils concrets et réalisables, ainsi qu'à proposer des principes plus larges qui devraient sous-tendre toutes les recherches et pratiques statistiques.

Chair/Président: Jean Vaillancourt

Organizer/Responsable: Jean Vaillancourt

Room/Salle: A 1049

Date: Monday June 3 / lundi 3 juin

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-16:00]

Deli Li (Lakehead University) **Andrew Rosalsky** (University of Florida, USA)

Some Limit Theorems for Negative Quadrant Dependent Random Variables

Des théorèmes limites pour des variables aléatoires négativement dépendantes d'un quadrant

Negatively associated random variables play a significant role in probability theory. Such random variables inherit some properties enjoyed by independent random variables as do many other dependence structures; independent random variables are negatively associated and negatively associated random variables are negatively quadrant dependent. In this talk, we present some limit theorems for negatively quadrant dependent random variables established by us, and then we review the latest research trends on this topic.

Les variables aléatoires associées négativement ont un rôle important dans la théorie de probabilité. Elles possèdent des propriétés dont profitent les variables aléatoires indépendantes tout comme plusieurs autres structures de dépendance. Les variables aléatoires indépendantes sont associées négativement, et donc négativement dépendantes à un quadrant. Dans le cadre de cette présentation, nous présentons des théorèmes limites pour des variables aléatoires dépendantes négativement d'un quadrant préétabli. Puis nous passerons en revue les dernières tendances de recherche sur ce sujet.

[16:00-16:30]

Hélène Guérin (Université du Québec à Montréal)

Elephant Random Walk, Polya Urns and Asymptotic Behavior

La marche aléatoire de l'éléphant, les urnes de Polya et leur comportement asymptotique

The elephant random walk has been introduced in the early 2000s by physicists. This random walk keeps the memory of its entire history at all times, and its behavior depends on a memory parameter. Three regimes of asymptotic behavior have been identified: diffusive, critical and super-diffusive. Several variants of this walk have been studied, but in this presentation, we will focus on the classical elephant walk. A link with Polya urns will be introduced, enabling us to obtain a fixed-point equation for the limiting random variable that appears in the super-diffusive regime. From the fixed-point equation we'll deduce information about the law of this asymptotic variable. Depending on the time available, we will then return to the finite-time walk and introduce polynomials related to its characteristic function, in order to study its law at each instant.

La marche aléatoire de l'éléphant a été introduite au début des années 2000 par des physiciens. Cette marche aléatoire garde à chaque instant la mémoire de tout son passé et son comportement dépend d'un paramètre de mémoire. Trois régimes de comportement asymptotique ont été observés : diffusif, critique et super-diffusif. Plusieurs variantes de cette marche ont été étudiées mais dans cette présentation on se concentrera sur la marche de l'éléphant classique. On introduira un lien avec les urnes de Polya, ce qui nous permettra d'obtenir une équation de point fixe sur la variable aléatoire limite qui apparait dans le régime super-diffusif. De l'équation de point fixe on déduira des informations sur la loi de cette variable limite. Si le temps le permet, on reviendra par la suite à la marche en temps fini et on introduira des polynômes en lien avec sa fonction caractéristique dans le but d'étudier la loi à chaque instant.

[16:30-17:00]

Lucile Laulin (Université Paris Nanterre) **Alice Contat** (Université Sorbonne Paris Nord)

Scaling Limit for Amnesic Step-Reinforced Random Walks

Limite d'échelle pour les marches aléatoires renforcées amnésiques

A step-reinforced random walk is a self-interacting random walk that each time either repeats one of its former steps chosen uniformly or takes a step independently from its past. We introduce a variation of the step-reinforced random walk with general memory, which can be interpreted as amnesia. Our main purpose is to establish a version of Donsker's invariance principle for such amnesic step-reinforced random walks in the diffusive regime. While for the standard step-reinforced walk, the limit arising is a so-called noise reinforced Brownian motion, we show that for the amnesiac version, the Gaussian process is actually the sum of a noise-reinforced Brownian motion and a (not independent) Brownian motion.

Une marche aléatoire renforcée est un processus en auto-interaction, qui à chaque instant, soit répète l'un des pas précédent choisi de manière uniforme, soit effectue un nouveau pas indépendant. Nous introduisons une généralisation de la mémoire qui peut être interprétée comme de l'amnésie et nous voulons établir un principe d'invariance pour ce type de modèle. Alors que dans le cas standard de la marche renforcée, la limite qui apparaît et le mouvement Brownien renforcé, nous montrons que dans ce cas général on obtient la somme d'un mouvement Brownien renforcé et d'un mouvement Brownien qui sont dépendants.

Past, Present, and Future of Statistical Education
Le passé, le présent et l'avenir de l'enseignement de la statistique

Chair/Président: Yildiz Yilmaz

Organizer/Responsable: Yildiz Yilmaz

Room/Salle: A 2071

Date: Monday June 3 / lundi 3 juin

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-16:00]

Richard J. Cook (University of Waterloo)

Reflections on Some Experiential Graduate Training Programs in Biostatistics

Réflexions sur certains programmes universitaires de formation expérimentale en biostatistique

There is an increasing need for biostatisticians to engage in data-intensive health research in Canada and around the world. To work effectively upon graduation it is best to provide students with experiences working in a team environment during their training in order to equip them with practical knowledge and relevant soft skills. In addition to a strong foundation in statistics, proficiency in data analysis and scientific computing skills, an ability to communicate in a team research environment is essential for biostatisticians to make meaningful contributions. This talk will recap some experiences in designing and managing three biostatistics training programs over the past three decades, with each designed to increase capacity and capability in specific areas of health research in Canada. Considerations for the design of future effective training programs will also be highlighted.

Le Canada et le monde entier ont de plus en plus besoin de biostatisticiens pour participer à la recherche en santé nécessitant un grand volume de données. Pour que les étudiants puissent travailler efficacement après l'obtention de leur diplôme, il est préférable de leur donner l'occasion de travailler en équipe au cours de leurs études afin qu'ils acquièrent des connaissances pratiques et des compétences non techniques pertinentes. Outre de solides bases en statistiques, des compétences en analyse de données et en calcul scientifique, une capacité à communiquer dans un milieu de recherche en équipe est essentielle pour que les biostatisticiens puissent apporter une contribution significative. Dans cette présentation, nous passerons en revue certaines expériences de conception et de gestion de trois programmes d'études en biostatistiques au cours des trente dernières années, chacun d'entre eux étant destiné à renforcer les connaissances et les compétences dans des domaines de recherche spécifiques de la santé au Canada. Nous examinerons également les questions relatives à la conception de futurs programmes de formation efficaces.

[16:00-16:30]

Alison L. Gibbs (University of Toronto)

An Undergraduate Program in Statistics: Past Influences, Current Curriculum, and Future Questions

Programme de statistique de premier cycle universitaire : influences du passé, programme actuel et questions sur l'avenir

In 1988, David Moore asked "Should Mathematicians Teach Statistics?" Twenty years later, Brown and Kass countered with a call to rethink statistical training, asking "What is Statistics?" Consideration of both these questions has had significant impact on the evolution of the curriculum of the undergraduate program in statistics at the University of Toronto, which currently has enrollment of over 4,000 students. We will examine the

En 1988, David Moore demandait si les mathématiciens devaient enseigner la statistique. Vingt ans plus tard, Brown et Kass ont répliqué par un appel à repenser la formation en statistique en demandant ce qu'est la statistique. La prise en considération de ces questions a eu un impact marqué sur l'évolution du programme de statistique de premier cycle à l'Université de Toronto, auquel plus de 4 000 étudiants sont inscrits. Nous examinons la structure actuelle du programme et ses résultats sur le plan de l'apprentis-

current structure and learning outcomes of the program, as well as some additional questions that influenced its design. And as we approach another twenty years, we'll also consider what questions we should be asking now.

sage, en plus de voir certaines autres questions qui ont influé sur sa conception. Près de vingt ans après l'interrogation de Brown et Kass, nous considérons les questions que nous devons maintenant nous poser.

[16:30-17:00]

Donald Estep (Simon Fraser University/CANSSI) **Donald Estep** (Canadian Statistical Sciences Institute and Simon Fraser University)

The Vision for a CANSSI Research Training Library: Providing Canadian Students with Cutting Edge Statistical Knowledge and Skills

Objectif d'une bibliothèque de formation à la recherche de l'INCASS : fournir aux étudiants canadiens des connaissances et des compétences statistiques de pointe

Developed at the CANSSI National Retreat in 2020, the vision for a CANSSI Research Training Library (CRTL) is to provide students at all Canadian universities and colleges access to advanced preparation for research. It is motivated by the fact that students in various universities have limited access to research-level coursework, affecting their preparation for future careers and their competitiveness in seeking employment. A CRTL course package would include recorded lectures, lecture notes, assignments and assessments, projects, data sets, and auxiliary materials. The packages could be used in a variety of ways, e.g., as a basis for a reading course (organized locally) or for use by a faculty member teaching a course for the first time. Packages would be offered in English and French. The selection of course packs and developers would be guided by CANSSI and the SSC. In this talk, we will describe the idea of the CTRL and issues involved with implementation.

Élaborée lors de la retraite nationale de l'INCASS en 2020, l'objectif d'une bibliothèque de formation à la recherche de l'INCASS est de fournir aux étudiants de toutes les universités et de tous les collèges du Canada la possibilité de se préparer intensivement à la recherche. Ce projet est motivé par le fait que les étudiants de diverses universités ont un accès limité aux cours de recherche, ce qui nuit à leur préparation à une future carrière et à leur compétitivité lorsqu'ils cherchent un emploi. Un module de cours de la bibliothèque de formation à la recherche comprendrait des conférences enregistrées, des notes de cours, des travaux et des évaluations, des projets, des ensembles de données et du matériel connexe. Les modules, proposés en anglais et en français, pourraient être utilisés de différentes manières, par exemple comme base d'un cours de lecture (organisé localement) ou par un membre du corps enseignant qui donnerait un cours pour la première fois. La sélection des modules de cours et des développeurs sera guidée par l'INCASS et la Société statistique du Canada. Dans cette présentation, nous décrirons l'objectif de la bibliothèque de formation à la recherche et les problèmes liés à sa mise en œuvre.

Innovative Strategies in High-Dimensional Data Analysis with Applications to Business and Industry
Stratégies innovantes en analyse de données à haute dimension avec les applications aux entreprises et à l'industrie

Chair/Président: Armin Hatefi

Organizer/Responsable: S. Ejaz Ahmed

Room/Salle: A 2065

Date: Monday June 3 / lundi 3 juin

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-16:00]

Anand N Vidyashankar (George Mason University) **Crissa Marshburn** (McKesson Corporation)

Assessing Privacy and Security Risk and Mitigation Strategies

Évaluation des risques pour la vie privée et la sécurité et stratégies d'atténuation

Privacy and security laws continuously evolve, with states adopting and refining existing guidelines. Recently, within the U.S., Delaware passed a comprehensive data privacy law, joining twelve other states to provide consumers with privacy rights. These privacy laws tend to include additional security requirements and recommendations. Data warehouses that process personally identifiable information (PII) adopt disclosure control mechanisms to adhere to federal and state privacy and security regulatory guidelines. Thus, using metrics that integrate privacy and security vulnerabilities is beneficial. In this presentation, we describe a new class of metrics that are also key performance indicators of security and privacy policies. We study the statistical properties of the proposed metrics and provide an uncertainty assessment that facilitates the development of policies and procedures for data sharing.

Les lois relatives à la protection de la vie privée et à la sécurité évoluent constamment, les États adoptant et affinant les lignes directrices existantes. Récemment, aux États-Unis, le Delaware a adopté une loi complète sur la confidentialité des données, rejoignant ainsi douze autres États pour garantir les droits des consommateurs en matière de protection de la vie privée. Ces lois sur la protection de la vie privée ont tendance à inclure des exigences et des recommandations supplémentaires en matière de sécurité. Les entrepôts de données qui traitent des informations personnelles identifiables (PII) adoptent des mécanismes de contrôle de la divulgation afin de respecter les directives réglementaires fédérales et nationales en vigueur. Il est donc utile d'utiliser des mesures qui intègrent les vulnérabilités en matière de protection de la vie privée et de sécurité. Dans cette présentation, nous décrivons une nouvelle catégorie de mesures qui sont également des indicateurs clés de performance des politiques de sécurité et de protection de la vie privée. Nous en étudions les propriétés statistiques et fournissons une évaluation de l'incertitude qui facilite le développement de politiques et de procédures pour le partage des données.

[16:00-16:30]

Yi Li (University of Michigan)

Penalized Deep Partially Linear Cox Models

Modèles de Cox partiellement linéaires profonds pénalisés

Partially linear Cox models gain popularity for survival analysis by dissecting the hazard function into parametric and nonparametric components, allowing for the effective incorporation of both well-established risk factors (such as age and clinical variables) and emerg-

Les modèles de Cox partiellement linéaires gagnent en popularité pour l'analyse de survie. En effet, ils décomposent la fonction de risque en composants paramétriques et non paramétriques, ce qui permet d'intégrer efficacement les facteurs de risque avérés (âge et variables cliniques) et nouveaux (caractéristiques de l'image) dans

ing risk factors (e.g., image features) within a unified framework. However, when the dimension of parametric components exceeds the sample size, the task of model fitting becomes formidable, while nonparametric modeling grapples with the curse of dimensionality. We propose a novel Penalized Deep Partially Linear Cox Model (Penalized DPLC), which incorporates the SCAD penalty to select important texture features and employs a deep neural network to estimate the nonparametric component of the model. We prove the convergence and asymptotic properties of the estimator and compare it to other methods through extensive simulation studies, evaluating its performance in risk prediction and feature selection.

un cadre unifié. Cependant, lorsque la dimension des composants paramétriques dépasse la taille de l'échantillon, l'ajustement du modèle se révèle formidable, tandis que la modélisation non paramétrique pose problème en ce qui concerne la dimensionnalité. Nous proposons un nouveau modèle de Cox partiellement linéaire pénalisé, qui intègre la pénalité SCAD pour sélectionner les caractéristiques de texture importantes et utilise un réseau neuronal profond pour estimer le composant non paramétrique du modèle. Nous démontrons la convergence et les propriétés asymptotiques de l'estimateur et nous le comparons à d'autres méthodes à l'aide d'études de simulation approfondies, en évaluant son efficacité pour la prédiction des risques et la sélection des caractéristiques.

[16:30-17:00]

S. Ejaz Ahmed (Brock University)

Post-Shrinkage Strategies in Semiparametric Models for High-Dimensional Data Application

Stratégies post-rétrécissement dans les modèles semi-paramétriques pour l'application à des données de forte dimension

We present post-shrinkage strategy for the regression parameters of semiparametric models. The regression parameter vector is partitioned into two sub-vectors: the first sub-vector gives the predictors of interest, i.e., main effects (treatment effects), and the second sub-vector is for variables that may or may not needed to be controlled. We establish both theoretically and numerically that the proposed shrinkage strategy which combines two semiparametric estimators computed for the full model and the submodel outshines the full model estimation. A data example is given. We extend this strategy to high-dimensional data (HDD) analysis. For HDD analysis many penalized methods were introduced for simultaneous variable selection and parameters estimation when the model is sparse. We propose a high-dimensional shrinkage strategy to improve the prediction performance of a submodel. We demonstrate that the proposed strategy performs uniformly better than the existing methods in many cases.

Nous présentons une stratégie post-rétrécissement des paramètres de régression des modèles semi-paramétriques. Le vecteur des paramètres de régression est divisé en deux sous-vecteurs : le premier sous-vecteur contient les prédicteurs d'intérêt, à savoir les effets principaux (effets de traitement), et le second sous-vecteur contient les variables qu'il peut être nécessaire de contrôler ou non. Nous démontrons à la fois de manière théorique et numérique que la stratégie de rétrécissement proposée, qui combine deux estimateurs semi-paramétriques calculés pour le modèle complet et le sous-modèle, est meilleure que l'estimation du modèle complet. Nous présentons également un exemple de données. Nous étendons cette stratégie à l'analyse des données de forte dimension. Pour l'analyse de ces données, il existe de nombreuses méthodes pénalisées pour la sélection simultanée des variables et l'estimation des paramètres lorsque le modèle est peu dense. Nous proposons une stratégie de rétrécissement des données de forte dimension afin d'améliorer la prédiction d'un sous-modèle. Nous démontrons que, dans de nombreux cas, notre stratégie est plus efficace que les méthodes actuelles.

Isobel Loutit Invited Address
Allocution Isobel-Loutit

Chair/Président: Farouk Nathoo

Organizer/Responsable: Farouk Nathoo

Room/Salle: IIC 2001

Date: Monday June 3 / lundi 3 juin

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-17:00]

Johanna G. Nešlehová (McGill University)

Can Bayes Spaces Help to Detect Anomalous Risk Element Concentrations in Agricultural Soils?

Les espaces de Bayes peuvent-ils aider à détecter des concentrations anormalement élevées d'éléments dangereux dans les sols agricoles ?

Many countries have established geochemical mapping programs aimed at assessing soil contamination through hazardous chemical elements. This typically leads to large data sets of high quality whose analysis should make it possible to identify soils that violate legislative limits for safe food production. However, sophisticated data mining tools are needed to go beyond a superficial detection of extreme concentration anomalies while taking into account the natural variability of soil composition. In this talk, I will show how a distributional analog of the Hoeffding-Sobol identity leads to an orthogonal decomposition of probability densities in Bayes spaces, and I will exploit this result to perform a functional data analysis of chemical element concentration after regional post-stratification. Using data gathered between 1990 and 2009, several suspected contamination patterns in Czech agricultural soils were confirmed with this approach.

De nombreux pays ont instauré des programmes de cartographie géochimique visant à évaluer la contamination des sols par des éléments chimiques dangereux. Il en résulte généralement de grands ensembles de données de haute qualité dont l'analyse devrait permettre d'identifier les sols dépassant les normes légales de salubrité agro-alimentaire. Des outils sophistiqués d'exploration de données sont toutefois nécessaires pour aller au-delà d'une simple détection superficielle des concentrations extrêmes tout en tenant compte de la variabilité naturelle dans la composition des sols. Dans cet exposé, je montrerai comment un analogue distributionnel de l'identité de Hoeffding-Sobol induit une décomposition orthogonale des densités de probabilité dans les espaces de Bayes et j'exploiterai ce résultat pour effectuer une analyse fonctionnelle de la concentration d'éléments chimiques après post-stratification régionale. L'emploi de cette approche sur des données recueillies entre 1990 et 2009 a permis de confirmer plusieurs soupçons quant aux schémas de contamination des terres agricoles tchèques.

Chair/Président: Peijun Sang

Organizer/Responsable: Peijun Sang

Room/Salle: A 1045

Date: Monday June 3 / lundi 3 juin

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-16:00]

Linglong Kong (University of Alberta) **Bei Jiang** (University of Alberta)

Gaussian Differential Privacy on Riemannian Manifolds

Confidentialité différentielle gaussienne appliquée à des variétés riemanniennes

We develop an advanced approach for extending Gaussian Differential Privacy (GDP) to general Riemannian manifolds. The concept of GDP stands out as a prominent privacy definition that strongly warrants extension to manifold settings, due to its central limit properties. By harnessing the power of the renowned Bishop-Gromov theorem in geometric analysis, we propose a Riemannian Gaussian distribution that integrates the Riemannian distance, allowing us to achieve GDP in Riemannian manifolds with bounded Ricci curvature. To the best of our knowledge, this work marks the first instance of extending the GDP framework to accommodate general Riemannian manifolds, encompassing curved spaces, and circumventing the reliance on tangent space summaries. We provide a simple algorithm to evaluate the privacy budget μ on any one-dimensional manifold and introduce a versatile Markov Chain Monte Carlo (MCMC)-based algorithm to calculate μ on any Riemannian manifold with constant curvature.

Nous développons une approche de pointe servant à élargir l'application de la confidentialité différentielle gaussienne (GDP) à des variétés riemanniennes générales. Le concept de GDP se distingue en tant que définition importante de la confidentialité qui justifie fortement son usage dans des variétés, en raison de ses propriétés de limite centrale. En exploitant la puissance du renommé théorème Bishop-Gromov en analyse géométrique, nous proposons une distribution gaussienne et riemannienne qui comprend la distance riemannienne, ce qui permet d'obtenir la GDP dans des variétés riemanniennes avec courbure Ricci limitée. Au mieux de nos connaissances, ce travail est le premier cas d'élargissement de la GDP afin de s'adapter aux variétés riemanniennes générales, tout en englobant les espaces courbés et en contournant la dépendance aux sommaires d'espace de tangente. Nous offrons un algorithme simple pour évaluer le budget de confidentialité μ dans toute variété unidimensionnelle et présentons un algorithme Monte Carlo par chaîne de Markov (MCMC) polyvalent afin de calculer μ dans toute variété riemannienne avec courbure constante.

[16:00-16:30]

Dehan Kong (University of Toronto)

LLOT: application of Laplacian Linear Optimal Transport in Spatial Transcriptome Reconstruction

Application du transport optimal linéaire laplacien (LLOT) à la reconstruction spatiale du transcriptome

Single-cell RNA sequencing (scRNA-seq) allows transcriptional profiling, and cell-type annotation of individual cells. However, sample preparation in typical scRNA-seq experiments often homogenizes the samples, thus spatial locations of individual cells are often lost. Although spatial transcriptomic techniques, such as in situ hybridization (ISH) or Slide-seq, can be

Le séquençage de l'ARN sur cellules uniques (scRNA-seq) permet d'établir des profils transcriptionnels et d'annoter le type de cellules individuelles. Cependant, la préparation des échantillons dans les expériences typiques de scRNA-seq homogénéise souvent les échantillons, ce qui fait que la localisation spatiale des cellules individuelles est souvent perdue. Bien que les techniques transcriptomiques spatiales, telles que l'hybridation in situ ou le

used to measure gene expression in specific locations in samples, it remains a challenge to measure or infer expression level for every gene at a single-cell resolution in every location in tissues. We describe Laplacian Linear Optimal Transport (LLOT), a biologically interpretable method to integrate single-cell and spatial transcriptomics data to reconstruct missing information at a whole-genome and single cell resolution. LLOT has two essential features. First, it can recognize differences between datasets and correct platform effects efficiently based on a linear map. Second, it can handle complex spatial structures of tissues. We benchmarked LLOT against several other methods on real datasets. The results showed that LLOT consistently outperformed others in predicting spatial expressions and locations for single cells.

Slide-seq, puissent être utilisées pour mesurer l'expression des gènes à des endroits spécifiques des échantillons, il reste difficile de mesurer ou de déduire le niveau d'expression de chaque gène à la résolution d'une seule cellule, à chaque endroit des tissus. Nous décrivons le transport optimal linéaire laplacien (LLOT), une méthode biologiquement interprétable qui permet d'intégrer les données transcriptomiques spatiales et à l'échelle de la cellule afin de reconstruire les informations manquantes à l'échelle du génome entier et à l'échelle de la cellule. Le LLOT présente deux caractéristiques essentielles. Premièrement, il permet de reconnaître les différences entre les ensembles de données et corriger efficacement les effets de plateforme sur la base d'une carte linéaire. Deuxièmement, il peut gérer les structures spatiales complexes des tissus. Nous comparons le LLOT à plusieurs autres méthodes sur des ensembles de données réels. Les résultats montrent que le LLOT surpasse systématiquement les autres méthodes dans la prédiction des expressions spatiales et des emplacements de cellules individuelles.

[16:30-17:00]

Gregory Rice (University of Waterloo) **Sebastian Kuhnert** (University of California at Davis) **Alexander Aue** (University of California at Davis) **Jeremy VanderDoes** (University of Waterloo)

An operator-level functional GARCH model

Un modèle GARCH fonctionnel au niveau de l'opérateur

Conditionally heteroskedastic processes are commonly described by GARCH models. Such models have been intensively analyzed in the univariate and multivariate settings, and more recently in the settings of high-dimensional and function-valued data. So far in the functional setting, GARCH models have only been extended to "pointwise" models akin to the multivariate diagonal GARCH. In this talk we consider functional GARCH models that describe the evolution of the entire conditional covariance operator of the data, which we term "operator-level fGARCH" models. We derive sufficient conditions for the existence of unique, strictly stationary solutions of operator-level fGARCH recursions, moment properties, and consistency properties of regularized Yule-Walker estimators for the infinite dimensional model parameters. We demonstrate in several simulation experiments and data analyses to high-frequency asset returns data the usefulness of this new class of models.

Les processus conditionnellement hétéroscédastiques sont couramment décrits par le modèle GARCH. De tels modèles ont été analysés de manière intensive dans les contextes univariés et multivariés, et plus récemment dans le contexte de données de grande dimension et à valeur fonctionnelle. Jusqu'à présent, dans le contexte fonctionnel, les modèles GARCH n'ont été étendu qu'aux modèles ≪ ponctuels ≫ semblables au GARCH diagonal multivarié. Dans cet exposé, nous considérons une extension des modèles GARCH fonctionnels pour modéliser l'évolution de l'ensemble de l'opérateur de covariance conditionnelle des données, que nous appelons ≪ au niveau de l'opérateur ≫. Modèles fGARCH ". Nous dérivons des conditions suffisantes pour l'existence de solutions uniques et strictement stationnaires aux récursions fGARCH au niveau de l'opérateur, aux propriétés de moment et aux propriétés de cohérence des estimateurs Yule-Walker régularisés pour le paramètre. Nous démontrons dans plusieurs expériences de simulation et analyses de données que - la fréquence des retours d'actifs donne des données sur l'utilité de cette nouvelle classe de modèles.

Ensemble Learning for Developing Predictive Models
Apprentissage d'ensemble pour le développement de modèles prédictifs

Chair/Président: Rob Deardon

Organizer/Responsable: Jabed H Tomal

Room/Salle: A 1046

Date: Monday June 3 / lundi 3 juin

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-16:00]

Thomas M. Loughin (Simon Fraser University) **Jiahao Tian** (Simon Fraser University) **Hugh Chipman** (Acadia University)

MLCBART: Multilabel Classification with Bayesian Additive Regression Trees

MLCBART : Classification multi-étiquettes avec arbres de régression additifs bayésiens

Multilabel Classification (MLC) deals with the simultaneous binary classification of multiple labels, which is challenging because correlations among labels may exist even after accounting for effects of predictor variables. We present a Bayesian additive regression tree (BART) framework to solve the problem. Our adaptation, MLCBART, assumes that labels arise from thresholding an underlying numeric scale, where a multivariate normal model allows explicit estimation of the correlation structure among labels. This enables the discovery of complicated relationships and improves MLC predictive performance. Our method not only enables uncertainty quantification for each predicted label, but our MCMC draws produce an estimated conditional probability distribution of label combinations for any predictor values. Simulations demonstrate the effectiveness of the proposed model by comparing its performance with a set of models, including the oracle model with the correct functional form.

La classification multi-étiquettes (MLC) traite de la classification binaire simultanée de plusieurs étiquettes, ce qui est difficile car il peut y avoir des corrélations entre étiquettes même après avoir pris en compte les effets des variables prédictives. Nous présentons un arbre de régression additif bayésien (BART) pour résoudre ce problème. Notre adaptation, MLCBART, suppose que les étiquettes proviennent du seuillage d'une échelle numérique sous-jacente, où un modèle normal multivarié permet une estimation explicite de la structure de corrélation entre les étiquettes. Cela permet de découvrir des relations complexes et d'améliorer les performances prédictives de la MLC. Notre méthode permet non seulement de quantifier l'incertitude pour chaque étiquette prédite, mais nos tirages MCMC produisent une distribution de probabilité conditionnelle estimée des combinaisons d'étiquettes pour toutes les valeurs des prédicteurs. Nous démontrons par des simulations l'efficacité du modèle proposé en comparant ses performances avec un ensemble de modèles, y compris le modèle de l'oracle avec la forme fonctionnelle correcte.

[16:00-16:30]

Geoff Pleiss (The University of British Columbia)

Ensembles in the Age of Overparameterization: Promises and Pathologies

Les méthodes d'ensemble à l'ère du surparamétrage : promesses et pathologies

Ensemble methods have historically used either high-bias base learners (e.g. through boosting) or high-variance base learners (e.g. through bagging). Today, it is common to ensemble extremely overparameterized base learners like neural networks, which cannot be understood through the classical bias-variance tradeoff. This talk will cover surprising and counter-

Historiquement, les méthodes d'ensemble ont utilisé soit des apprenants de base à fort biais (par exemple par boosting), soit des apprenants de base à forte variance (par exemple par bagging). Aujourd'hui, il est courant d'assembler des apprenants de base extrêmement surparamétrés comme les réseaux de neurones, qui ne peuvent pas être compris par le compromis classique biais-variance. Cet exposé traitera des phénomènes surpre-

intuitive phenomena that emerge in this overparameterized regime. While ensembles improve the predictive performance of overparameterized base learners in a simple and cost-effective manner, their predictive capabilities and robustness are often outperformed by single (but larger) models. Furthermore, discouraging diversity amongst component models often improves the ensemble's predictive performance, counter to classic intuitions underpinning bagging and feature subsetting techniques. I will connect these empirical findings to the modern "double descent" interpretation of the bias-variance tradeoff, and I will conclude with implications for uncertainty quantification, robustness, and decision making.

nants et contre-intuitifs qui émergent dans ce régime de surparamétrage. Alors que les méthodes d'ensembles améliorent la performance prédictive des apprenants de base surparamétrés d'une manière simple et rentable, leurs capacités prédictives et leur robustesse sont souvent surpassées par des modèles uniques (mais plus grands). En outre, le fait de décourager la diversité parmi les modèles composant un ensemble améliore souvent sa performance prédictive, ce qui va à l'encontre des intuitions classiques qui sous-tendent les techniques de bagging et de feature subsetting. Je relierai ces résultats empiriques à l'interprétation moderne de la « double descente » du compromis biais-variance, et je conclurai sur les implications pour la quantification de l'incertitude, la robustesse et la prise de décision.

[16:30-17:00]

Jabed H Tomal (Thompson Rivers University)

Ensemble Learning Based on Subsets of Variables

Apprentissage d'ensemble basé sur des sous-ensembles de variables

An ensemble is an aggregated collection of models which as a whole can be considered a model. It has been confirmed by many authors that an ensemble can substantially improve the performance of a non-ensemble method. Breiman (1996) proposed an ensemble by using bootstrap aggregation (bagging) of classification and regression trees. Freund and Schapire (1997) proposed another ensemble (boosting) by incrementally building more powerful models, each focusing more on the data missed by the previous models. Breiman (2001) proposed random forests which uses, on top of bagging, a random subset of variables at each node of the trees in the ensemble. Cannings and Samworth (2017) proposed an ensemble using random subsets of variables. Tomal et al. (2015, 2016, 2017, 2021, 2022, 2023) proposed ensembles based on diverse subsets of predictors and showed how such subsets can be uncovered in large data. This talk aims to present some standard predictive ensembles along with the ensemble of subsets.

Un ensemble est un agrégat de modèles qui, dans son entièreté, peut être considéré comme un modèle. De nombreux auteurs confirment qu'un ensemble peut nettement améliorer la performance d'une méthode non-ensemble. Breiman (1996) a proposé un ensemble en utilisant une agrégation bootstrap (bagging) d'arbres de classification et de régression. Freund et Schapire (1997) ont proposé un autre ensemble (boosting) en construisant graduellement des modèles plus puissants, chacun davantage axé sur les données manquées dans les modèles précédents. Breiman (2001) proposait des forêts aléatoires utilisant, outre le bagging, un sous-ensemble aléatoire de variables à chaque nœud des arbres de l'ensemble. Cannings et Samworth (2017) ont proposé un ensemble utilisant des sous-ensembles aléatoires de variables. Quant à Tomal et al. (2015, 2016, 2017, 2021, 2022, 2023), en plus de proposer des ensembles basés sur divers sous-ensembles de prédicteurs, ils ont aussi montré que ces sous-ensembles peuvent être découverts dans les grands ensembles de données. Notre allocution vise à présenter certains ensembles prédictifs standards, de même que l'ensemble des sous-ensembles.

Chair/Président: Zelalem Firisa Negeri

Room/Salle: C 4036

Date: Monday June 3 / lundi 3 juin

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-15:45]

Jianan Peng (Acadia University)

Simultaneous Identifications of the Minimum Effective Dose in Each of Several Groups by the DR Method

Identifications simultanées de la dose minimale efficace dans chacun de plusieurs groupes par la méthode DR

Identification of the minimum effective dose (MED), which is the lowest dose level with an effect that exceeds that of the zero dose, is very important in drug development. For the MED problem, the probability of declaring an ineffective dose to be effective should be controlled. MED issue has been well studied in one-way anova. Jan et al. (Simultaneous Identifications of the Minimum Effective Dose in Each of Several Groups. Journal of Statistical Computation and Simulation. Vol 77, 149-161.) considered simultaneous identifications of the MED in two-way anova dose-response studies by extending step-down closed testing for the one-way anova. In this talk, we generalize Hsu and Berger (1999)'s DR method to two-way anova dose-response studies. A real data is used to illustrate the new method. Simulation studies are conducted to compare error rates for Jan et al.'s methods and the new method.

L'identification de la dose minimale efficace (DME), à savoir le niveau de dose le plus bas dont l'effet dépasse celui de la dose zéro, est très importante dans le développement des médicaments. Pour le problème de la DME, la probabilité de déclarer une dose inefficace comme étant efficace doit être contrôlée. La question de la DME a été bien étudiée dans le cadre de l'anova à un facteur. Jan et al. (Simultaneous Identifications of the Minimum Effective Dose in Each of Several Groups. Journal of Statistical Computation and Simulation. Vol 77, 149-161.) ont étudié les identifications simultanées de la dose minimale efficace dans les études dose-réponse par anova à deux facteurs en étendant les tests fermés par étapes de l'anova à un facteur. Dans cet exposé, nous généralisons la méthode DR de Hsu et Berger (1999) aux études dose-réponse par anova à deux facteurs. Nous utilisons des données réelles pour illustrer la nouvelle méthode. Nous menons des études de simulation pour comparer les taux d'erreur des méthodes de Jan et al. et de la nouvelle méthode.

[15:45-16:00]

Christopher Gravel (University of Ottawa) **Muhammad Mullah** (University of Ottawa)

Evaluation of Phenomenological Models for the Short-term Forecasting of Daily COVID-19 Case Incidence in Canada in the Presence of Multiple Waves

Évaluation de modèles phénoménologiques pour la prévision à court terme de l'incidence des cas de la COVID-19 au Canada en présence de vagues multiples

Phenomenological models were used for short-term forecasting of daily COVID-19 case incidence throughout the pandemic representing an important public health metric. These models typically rely on cumulative data to depict overall growth over time resulting in the smoothing of short-term fluctuations and rendering rapid changes in disease dynamics difficult to capture. In the presence of multiple waves, rapid adaptation to changes in trajectories around inflection points is an im-

Les modèles phénoménologiques ont été utilisés pour prévoir à court terme l'incidence quotidienne des cas de la COVID-19 tout au long de la pandémie, représentant ainsi une mesure importante de la santé publique. Ces modèles s'appuient typiquement sur des données cumulatives pour décrire une croissance globale avec le temps, résultant en un lissage des fluctuations à court terme et rendant difficiles à saisir les changements rapides dans la dynamique de la maladie. En présence de vagues multiples, l'adaptation rapide aux changements dans les trajectoires autour

portant characteristic of these models to enable early response to a case resurgence or possible new variant. We compared five classes of phenomenological models in the presence of multiple pandemic waves: generalized logistic, polynomial, autoregressive, sub-epidemic and a semi parametric mixed model. We explored their properties and relative performance for short-term forecasting by fitting these models to COVID-19 case data across various time intervals selected between March 10, 2020 to December 24, 2021.

des points d'inflexion est une caractéristique importante de ces modèles pour enclencher une réponse précoce à une résurgence de cas ou à un possible nouveau variant. Nous avons comparé cinq classes de modèles phénoménologiques en présence de vagues pandémiques multiples : le modèle logistique généralisé, polynomial, autorégressif, sous-épidémique et mixte semiparamétrique. Nous avons étudié les propriétés et la performance relative de ces modèles pour la prévision à court terme, en les ajustant aux données sur les cas de la COVID-19 à divers intervalles de temps choisis entre le 10 mars 2020 et le 24 décembre 2021.

[16:00-16:15]

Wei Liu (York University) **Dongwei Wei** (York University)

Semiparametric Nonlinear Mixed-Effects Models with Covariate Measurement Errors and Change Points, with Application to AIDS Studies

Modèles semi-paramétriques à effets mixtes non linéaires avec erreurs de mesure de la covariable et points de rupture, appliqués à des études sur le sida

Semiparametric nonlinear mixed-effects models are very flexible in modeling complex longitudinal data. Covariates are often introduced in the models to partially explain inter-individual variations. In practice, statistical analyses may become complicated due to measurement errors and missing data in covariates as well as change points on response trajectories. We consider semiparametric nonlinear mixed-effects models which incorporate measurement errors and missing data in time-varying covariates and change points. We propose an approximate joint model likelihood estimation method for the model parameters by using Monte Carlo Expectation-Maximization algorithm. The proposed method is illustrated in a real dataset. A simulation study is conducted for the method comparison and evaluation.

Les modèles semi-paramétriques à effets mixtes non linéaires sont très polyvalents dans la modélisation de données longitudinales complexes. Les covariables sont souvent introduites dans les modèles pour expliquer en partie les variations entre individus. En pratique, les analyses statistiques peuvent se complexifier en raison d'erreurs de mesure et de données manquantes dans les covariables ainsi que des points de rupture dans les trajectoires de réponse. Nous évaluons des modèles semi paramétriques à effets mixtes non linéaires comportant des erreurs de mesure, des données manquantes dans les covariables à temps variable et des points de rupture. Nous proposons une méthode d'estimation de la vraisemblance d'un modèle conjoint approximatif pour les paramètres de modèle au moyen d'un algorithme espérance-maximisation de Monte Carlo. La méthode proposée est illustrée dans un ensemble de données réelles. Nous menons une étude en simulations pour évaluer et comparer la méthode.

[16:15-16:30]

Samuel Perreault (University of Toronto) **Gracia Y. Dong** (University of Toronto) **Alex Stringer** (University of Waterloo) **Hwashin Shin** (Health Canada) **Patrick Brown** (University of Toronto)

Case-Crossover Designs and Overdispersion with Application in Air Pollution Epidemiology

Analyses cas-croisé et surdispersion avec application en épidémiologie de la pollution atmosphérique

Over the last three decades, case-crossover designs have found many applications in health sciences, especially in air pollution epidemiology. They are typically used, in combination with partial likelihood techniques, to define a conditional logistic model for the responses, usually health outcomes, conditional on the exposures. Despite the fact that conditional logistic models have been shown equivalent, in typical air pollution epidemiology setups, to specific instances of the well-known Poisson

Au cours des trois dernières décennies, les analyses de type cas-croisé ont trouvé de nombreuses applications en sciences de la santé, notamment en épidémiologie de la pollution atmosphérique. Elles sont généralement utilisées, en combinaison avec des techniques de vraisemblance partielle, pour définir un modèle logistique pour les réponses, souvent des résultats de santé, conditionnellement à l'exposition environnementale. Malgré que, dans la cadre conventionnel en épidémiologie de la pollution atmosphérique, les modèles logistiques conditionnels sont en fait

time series model, it is often claimed that they cannot allow for overdispersion. This paper clarifies the relationship between case-crossover designs, the models that ensue from their use, and overdispersion. As illustrative example, we investigate the association between air pollution and morbidity in Toronto, Canada.

des instances spécifiques du modèle classique de séries temporelles de Poisson, il est souvent mentionné qu'ils ne permettent pas l'inclusion de surdispersion. Cet article clarifie la relation entre les analyses de type cas-croisé, les modèles qui découlent de leur utilisation, et la surdispersion. À titre d'exemple illustratif, nous étudions l'association entre la pollution atmosphérique et la morbidité à Toronto (Canada).

Chair/Président: Daniel J. McDonald

Room/Salle: C 3053

Date: Monday June 3 / lundi 3 juin

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-15:45]

Martin St-Pierre (Statistics Canada) **Samer Farfour** (Statistics Canada)

The Sampling and Weighting Methodology of the 2021 Census Undercoverage Study

La méthodologie d'échantillonnage et de pondération de l'Étude sur le sous-dénombrement du recensement de 2021

Statistics Canada's 2021 Census of Population in Canada provides high-quality population counts for a diverse set of characteristics. However, the census is not free of coverage errors. Among other things, the Census Undercoverage Study (CUS) aims to produce estimates of the number of people who are part of the census target population, but who have not been enumerated by the census. The 2021 CUS published its estimates in the fall of 2023. The presentation will focus on methodological improvements to the CUS sampling strategy and weighting steps, including adjustment for CUS nonresponse and calibration steps. Some key results will be presented, as well as an overview of the challenges encountered during the current cycle and those to come for the 2026 Census.

Le Recensement de la population du Canada effectué en 2021 par Statistique Canada fournit des comptes de population de grande qualité et pour un ensemble varié de caractéristiques. Toutefois, Le recensement n'est pas exempt d'erreurs de couverture. Entre autres, l'Étude sur le sous-dénombrement du recensement (ESoR) a pour but de produire l'estimation du nombre de personnes faisant partie de la population cible du recensement mais n'ayant pas été dénombrées par celui-ci. L'ESoR de 2021 a publié ses estimations à l'automne de 2023. La présentation portera sur les améliorations méthodologiques à la stratégie d'échantillonnage de l'ESoR, ainsi qu'aux étapes de pondération, entre autres l'ajustement pour la non-réponse à l'ESoR et les étapes de calages. Quelques résultats principaux seront présentés, ainsi qu'un aperçu des défis rencontrés lors du cycle qui se termine et de ceux à venir pour le Recensement de 2026.

[15:45-16:00]

Anne Mather (Statistics Canada) **Cilanne Boulet** (Statistics Canada) **Golshid Chatrchi** (Statistics Canada) **Andrew Brennan** (Statistics Canada)

Subsampling for Non-response Follow-up in Social Surveys: A Case Study

Sous-échantillonnage pour le suivi de la non-réponse dans les enquêtes sociales : Une étude de cas

With declining response rates, dealing with non-response has become more costly and time-intensive. One practical approach is to carry out non-response follow-up on only a portion of the sample, such as by limiting computer-assisted telephone interviewing (CATI) to a random subsample. This is a generalization of the classical approach described by Hansen and Hurwitz (1946). The Canadian Social Survey experimented with this approach and restricted CATI operations on two iterations of the survey. This presentation covers the adjustments that were applied to the weights of responding units to adjust for this subsampling. Moreover, it

Avec la baisse des taux de réponse, le suivi de la non-réponse est devenu plus coûteux et prend de plus en plus de temps. Une approche pratique consiste à n'effectuer ce suivi que sur une partie de l'échantillon, par exemple en limitant les interviews téléphoniques assistés par ordinateur (ITAO) à un sous-échantillon aléatoire. Il s'agit d'une généralisation de l'approche classique décrite par Hansen et Hurwitz (1946). L'Enquête sociale canadienne a expérimenté cette approche et a restreint ses opérations ITAO lors de deux itérations de l'enquête. Cette présentation décrit les ajustements qui ont été appliqués aux poids des unités répondantes pour tenir compte de ce sous-échantillonnage. En outre, d'autres estimateurs qui pourraient être utilisés sont présentés, avec leurs

presents alternate estimators that could be used, their theoretical properties with respect to biasedness, and the results of a simulation study comparing these estimators.

propriétés théoriques en ce qui concerne le biais et les résultats d'une étude de simulation comparant ces estimateurs.

[16:00-16:15]

Ariane Boivin (Université Laval) **Anne-Sophie Charest** (Université Laval)

Between Quality and Confidentiality: Generation of Robust Synthetic Data

Entre qualité et confidentialité : génération de données synthétiques robustes

It is often difficult, even sometimes impossible, to share denominalized data between organisations and researchers due to ethical constraints regarding participant confidentiality. Synthetic datasets could facilitate data sharing. However, many current methods, which use multiple imputation (MI) techniques for missing data, lower the analysis potential and the quality of the results. This project therefore aims to assess the confidentialy guarantees of a promising new data synthesis method. This method adds a data masking step to a multiple imputation technique to generate synthetic data based on the desired level of protection. In particular, attribute disclosure risks, which refer to the disclosure of certain attributes based on other, known ones, are tested. The feasability and quality of the results are tested on the Québec survey on cannabis 2022, provided by l'Institut de la statistique du Québec.

Il est souvent difficile, voire impossible, de partager des données dénominalisées entre organisations et chercheurs en raison de contraintes éthiques liées à la confidentialité des répondants. Les jeux de données synthétiques pourraient simplifier ce partage de données. Or, plusieurs méthodes actuelles, utilisant des concepts d'imputation de données manquantes, affectent le potentiel d'analyse et la qualité des résultats. Ce projet consiste donc à évaluer les garanties de confidentialité d'une nouvelle méthode prometteuse. Elle intègre un mécanisme de masquage à de l'imputation multiple pour adapter le modèle génératif au niveau de protection désiré. En particulier, les risques de divulgation d'attributs, c'est-à-dire la révélation de certaines valeurs d'attributs en fonction d'autres attributs connus, seront testés. La faisabilité et la qualité des résultats sera également testée sur les données de l'Enquête québécoise sur le cannabis, fournie par l'Institut de la statistique du Québec.

[16:15-16:30]

Alexander Imbrogno (Statistics Canada)

Using Non-Binary Gender to Calibrate Survey Weights for the Canadian Long-Form Census Sample

Utilisation du genre non-binaire pour le calage de l'échantillon du questionnaire détaillé du Recensement Canadien

In 2021, Canada became the first country to collect and publish data on gender in a national census giving Canadians the option to choose male, female, or non-binary. Due to their small sizes, non-binary population totals were excluded from the 2021 long-form sample calibration, due to the risk of increasing the variance of estimates. This talk presents an alternative long-form calibration procedure. Artificial sub-provincial non-binary totals are introduced as a tool to decompose a provincial calibration to non-binary totals back into independent sub-provincial problems, maintaining the computational efficiencies of the usual long-form calibration. An algebraic expression for the artificial totals under the chi-squared distance is derived. Simulation results are presented demonstrating the benefits of non-binary calibration on data quality for non-binary domains.

En 2021, le Canada est devenu le premier pays à recueillir et publier des données sur le genre lors d'un recensement national permettant ainsi aux Canadiens de s'identifier comme homme, femme ou non-binaire. Étant donné leur faible taille, les totaux de population non-binaire ont été exclus du calage pour l'échantillon du questionnaire détaillé de 2021 afin d'éviter le risque d'accroitre la variance des estimations. On présente une méthode alternative de calage qui introduit des totaux non-binaires infra-provinciaux artificiels afin de permettre la décomposition d'un calage provincial en des problèmes infra-provinciaux indépendants de manière à préserver les efficacités computationnelles du calage traditionnel. Une forme algébrique des totaux artificiels sous la distance du khi carré est dérivée. Des résultats de simulations sont présentés afin de démontrer les bénéfices du calage sur la qualité des données pour les domaines de personnes non-binaires.

[16:30-16:45]

Oluwagbohunmi Adetunji Awosoga (University of Lethbridge) **Nse Odunaiya** (University of Ibadan, College of Medicine, Nigeria) **Adesola Odole** (University of Ibadan, College of Medicine, Nigeria) **Olufemi Oyewole** (Olabisi Onabanjo University, Teaching Hospital, Nigeria) **Mercy Adegoke** (University of Ibadan, College of Medicine, Nigeria) **Chiedozie Alumona** (University of Lethbridge) **Ogochukwu Onyeso** (University of Lethbridge) **Abiodun Adeoye** (University of Ibadan, College of Medicine, Nigeria) **Happiness Aweto** (University of Lagos, Lagos University Teaching Hospital, Nigeria)

Cardiovascular Disease Perception and Risk among Community-Dwelling Adults in Southwest Nigeria

Perception et risque de maladie cardiovasculaire chez les adultes vivant en communauté dans le sud-ouest du Nigeria

Nigeria is reported to have a high cardiovascular disease-associated mortality rate, highlighting the need to reduce its prevalence. The first step to reducing cardiovascular disease (CVD) prevalence and mortality rate is to identify those with poor perception and high risk. Thus, this study determined the CVD perception and risk and explored their sociodemographic variations and predictors. The cross-sectional study involved 1,493 community-dwelling adults across Oyo, Lagos, and Ogun states in Southwest Nigeria. Participants' perceptions and risk were obtained using the Perception of Risk of Heart Disease Scale and Non-Laboratory-Based Interheart Risk Score, respectively. Data were analyzed using mean, standard deviation, independent sample t-test, one-way ANOVA, and multiple linear regression. Rural dwellers and Ogun state residents had significantly lower perception scores while males and Lagos state residents had significantly higher risk scores than their counterparts. Having a secondary education or below, living in rural areas, and living in Oyo and Ogun states significantly predicted having a low perception score. The overall perception and risk scores were high and low, respectively. However, rural dwellers and Ogun State residents have poorer perceptions, while males and Lagos State residents have a higher risk score.

Le Nigeria présente un taux de mortalité élevé lié aux maladies cardiovasculaires (MCV), ce qui souligne la nécessité de réduire la prévalence de ces maladies. La première étape consiste à identifier les personnes qui ont une mauvaise perception de ces maladies et qui courent un risque élevé. Cette étude a permis de déterminer la perception et le risque de MCV et d'en explorer les variations sociodémographiques et les prédicteurs. L'étude transversale a porté sur 1 493 adultes vivant en communauté dans les États d'Oyo, de Lagos et d'Ogun, dans le sud-ouest du Nigeria. Les perceptions et les risques des participants ont été obtenus à l'aide des échelles Perception of Risk of Heart Disease Scale et Non-Laboratory-Based Interheart Risk Score, respectivement. Les données ont été analysées à l'aide de la moyenne, de l'écart-type, du test-t pour échantillon indépendant, de l'ANOVA à un facteur et de la régression linéaire multiple. Les habitants des zones rurales et de l'État d'Ogun avaient des scores de perception significativement plus faibles, tandis que les hommes et les habitants de l'État de Lagos avaient des scores de risque significativement plus élevés que leurs homologues. Le fait d'avoir un niveau d'éducation secondaire ou inférieur, de vivre dans des zones rurales et de vivre dans les États d'Oyo et d'Ogun permettait de prédire de manière significative un faible score de perception. Les scores globaux de perception et de risque étaient respectivement élevés et faibles. Toutefois, les habitants des zones rurales et de l'État d'Ogun ont une moins bonne perception, tandis que les hommes et les habitants de l'État de Lagos ont un score de risque plus élevé.

[16:45-17:00]

Mohammed Sanda (Social Security and National Insurance Trust)

Revitalizing Social Security Systems through Innovative Survey Methods: A Case Study of SSNIT's Transformational Journey

Étude de cas sur le parcours transformationnel du SSNIT : modernisation des systèmes de sécurité sociale par des méthodes d'enquête novatrices

As the global landscape of social security undergoes significant transformations, this paper delves into the critical role of advanced survey methodologies in guiding the evolution of these systems. Focusing on the Social Security and National Insurance Trust (SSNIT), Ghana as a case study, the paper boarded on an extensive exploration of the applications, challenges, and outcomes of employing innovative survey methods to enhance the

Alors que le paysage mondial de la sécurité sociale connaît d'importantes transformations, nous examinons le rôle essentiel des méthodes d'enquête modernes dans l'orientation de l'évolution de ces systèmes. Nous examinerons en détail les applications, les défis et les résultats de l'utilisation de méthodes d'enquête novatrices visant à améliorer l'efficacité, l'inclusivité et l'adaptabilité des cadres de sécurité sociale en nous concentrant sur l'étude de cas du Social Security and National Insurance Trust (SSNIT), au

efficiency, inclusivity, and adaptability of social security frameworks.

Ghana.

Chair/Président: Yanglei Song

Room/Salle: ED 2018B

Date: Monday June 3 / lundi 3 juin

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-15:45]

Qing Wang (Ca' Foscari University of Venice) **Roberto Casarin** (Ca' Foscari University of Venice) **Radu Craiu** (University of Toronto)

Markov Switching Tensor Regression

Régression tensorielle à commutation de Markov

A Markov Switching tensor regression model is proposed, which allows for model instability and accounts for multi-dimensional array data. Regarding model instability, the parameters are assumed to be time-varying and are driven by latent processes to address structural breaks in the data. Regarding high dimensionality, the Soft PARAFAC strategy is followed to achieve dimensionality reduction while preserving the structural information between the covariates. Modified multi-way shrinkage prior is further imposed to address over-parametrization issues. An efficient MCMC algorithm that adopts random scan Gibbs within a back-fitting strategy is developed to achieve better scalability of the posterior approximation. The performance of the MCMC algorithm is demonstrated using synthetic datasets in simulation studies. Real-world applications are used to test the proposed model against the benchmark Lasso regression, where the model delivers superior performance.

Un modèle de régression tensorielle à commutation de Markov est proposé, qui permet au modèle d'être instable et prend en compte les données de tableaux multidimensionnels. En ce qui concerne l'instabilité du modèle, les paramètres sont supposés varier dans le temps et sont régis par des processus latents afin de tenir compte des ruptures structurelles dans les données. En ce qui concerne la haute dimensionnalité, la stratégie Soft PARAFAC est suivie pour réduire la dimensionnalité tout en préservant l'information structurelle entre les covariables. Un a priori modifié de rétrécissement à vois multiples est également imposé pour traiter les problèmes de surparamétrisation. Un algorithme MCMC efficace qui adopte le balayage aléatoire de Gibbs dans le cadre d'une stratégie back-fitting est développé pour obtenir une meilleure évolutivité de l'approximation a posteriori. Les performances de l'algorithme MCMC sont démontrées à l'aide d'ensembles de données synthétiques dans des études de simulation. Des applications réelles sont utilisées pour tester le modèle proposé par rapport à la régression Lasso de référence, où le modèle fournit des performances supérieures.

[15:45-16:00]

Hedayat Fathi (Université Laval) **Marzia A. Cremona** (Université Laval) **Federico Severino** (Université Laval)

Selection of Functional Predictors and Smooth Coefficient Estimation for Scalar-on-function Regression Models

Sélection de prédicteurs fonctionnels et estimation lisse des coefficients pour les modèles de régression scalaire-sur-fonction

Despite their importance in obtaining reliable and interpretable models, variable selection techniques are still underdeveloped in the framework of scalar-on-function regression models – in which several functional variables are employed to predict a scalar response. The available functional variable selection methods are generally based on a grouping strategy. This approach lacks

Malgré leur importance pour obtenir des modèles fiables, les techniques de sélection de variables sont encore peu explorées pour des modèles de «régression scalaire-sur-fonction» où plusieurs variables fonctionnelles sont utilisées pour prédire une réponse scalaire. Les approches disponibles sont généralement basées sur les «méthodes de regroupement». Ces méthodes ne garantissent la pertinence de la sélection et ne sont pas efficaces quand le

theoretical guarantees on the selection of the relevant predictors and is not effective when a large number of variables are present. We propose a methodology for selecting relevant functional predictors, while simultaneously providing accurate smooth (or, more in general, regular) estimates of the functional coefficients. We suppose that the functional predictors lie in a real separable Hilbert space, while the functional coefficients belong to a specific subspace of this Hilbert space. Such subspace can be a Reproducing Kernel Hilbert Space (RKHS). Coefficient estimates are obtained by an adaptive penalized least square algorithm which employs functional subgradients to efficiently solve the minimization problem and allows us to establish asymptotic properties of our estimators. In particular, we prove that our method satisfies the functional oracle property. Performance in terms of variable selection and accuracy of coefficient estimation is assessed through simulations.

nombre de variables est grand. Nous introduisons une nouvelle méthodologie pour sélectionner des prédicteurs fonctionnels pertinents et, en même temps, fournir des estimations des coefficients fonctionnels à la fois précises et lisses. Nous supposons que les prédicteurs fonctionnels appartiennent à un espace de Hilbert réel séparable "H", et les coefficients fonctionnels appartiennent à un sous-espace spécifique de H, qui est un Espace de Hilbert à Noyau Reproduisant (EHNR). Les sélections de variables et les estimations des coefficients sont obtenues par un algorithme adaptatif de moindres carrés pénalisés. Nous utilisons "sous-gradients fonctionnels" pour résoudre le problème de minimisation et nous démontrons que notre méthode satisfait à la propriété oracle fonctionnelle. Finalement, la performance de la méthode en termes de sélection de variables et de précision de l'estimation est évaluée par des études de simulations.

[16:00-16:15]

Jonathan Jalbert (Polytechnique Montreal) **Auguste Paoli** (Polytechnique Montréal)

Goodness-of-Fit Tests for Multivariate Scaling Models of IDF Curves

Tests d'adéquation pour les modèles de scaling des courbes IDF

The Canadian Standards Association requires the use of precipitation Intensity-Duration-Frequency (IDF) curves to adequately design infrastructures exposed to extreme precipitation. These curves represent return levels for periods ranging from 2 to 100 years and accumulation durations from 5 minutes to 24 hours. Currently, Environment and Climate Change Canada independently estimates the return levels composing the IDF curves for each precipitation duration. However, due to the scarcity of available data, there is often large estimation variance. To mitigate this uncertainty, one approach is to leverage the functional relationship between precipitation accumulation across durations using a multivariate scaling model. Several scaling models exist in the literature but finding a suitable one can be challenging. The proposed talk aims to present a formal goodness-of-fit test to specifically develop to assess the adequacy of the multivariate scaling model to precipitation IDF data

L'Association canadienne de normalisation exige l'utilisation de courbes d'intensité-durée-fréquence (IDF) des précipitations pour concevoir des infrastructures exposées aux précipitations extrêmes. Ces courbes représentent les intensités pour des périodes de retour allant de 2 à 100 ans et des durées d'accumulation de 5 minutes à 24 heures. Actuellement, Environnement et Changement climatique Canada estime indépendamment les niveaux de retour des courbes IDF pour chaque durée de précipitation. Cependant, en raison de la rareté des données disponibles, il y a souvent une grande variance d'estimation. Pour atténuer cette incertitude, une approche consiste à exploiter la relation fonctionnelle entre l'accumulation des précipitations sur différentes durées à l'aide d'un modèle de scaling. Identifier un modèle de scaling adapté peut être difficile. La présentation proposée vise à décrire un test d'adéquation formel spécifiquement conçu pour évaluer la pertinence des modèles de scaling.

[16:15-16:30]

Ziang Zhang (University of Toronto) **Patrick Brown** (University of Toronto) **Jamie Stafford** (University of Toronto)

Unveiling Quasi-Periodic Patterns with Seasonal Gaussian Processes

Dévoilement de tendances quasi périodiques à l'aide de processus saisonniers gaussiens

Quasi-periodicity refers to a pattern in a function where

La quasi-périodicité fait référence à une tendance dans une fonc-

it appears periodic but has evolving amplitudes over time. This is often the case in practical settings such as the modeling of case counts of infectious disease. In this work, we consider a class of Gaussian processes, called seasonal Gaussian Processes (sGP), for inference of such quasi-periodic behavior in flexible Bayesian hierarchical model. To facilitate the computation for large datasets with irregular spacing, we develop a continuous finite dimensional approximation for sGP using the seasonal B-splines constructed by damping B-splines with sinusoidal functions. The accuracy of the proposed approximation is supported with both rigorous statistical theories and extensive simulation studies. We also provide a unified and interpretable way to define priors for sGP, based on the notion of predictive standard deviation. Finally, we illustrate the practical utility of the proposed method through various real data examples.

tion qui semble périodique, mais dont les amplitudes évoluent au fil du temps. C'est souvent le cas dans des situations pratiques comme la modélisation du dénombrement de cas de maladie infectieuse. Dans le cadre de ce travail, nous évaluons une classe de processus gaussiens, appelée processus gaussien saisonnier (PGS), afin d'inférer ce comportement quasi périodique dans un modèle hiérarchique bayésien polyvalent. Pour faciliter le calcul en présence de grands ensembles de données ayant des espacements irréguliers, nous développons une approximation dimensionnelle finie et continue pour les PGS à l'aide de B-splines saisonnières construites en amortissant les B-splines avec des fonctions sinusoïdales. La précision de l'approximation proposée est soutenue par des théories statistiques rigoureuses et des études en simulation approfondies. Nous fournissons aussi une façon unifiée et interprétable de définir les a priori pour les PGS, selon la notion de déviation prédictive standard. En conclusion, nous illustrons l'utilité en pratique de la méthode proposée par l'entremise de plusieurs exemples avec des données réelles.

[16:30-16:45]

Marcus Hlady (University of Manitoba) **Yuliya V. Martsynyuk** (University of Manitoba)

Estimators of Reliability Ratio, Their Asymptotic Properties and Use for Inference in Linear Structural Errors-in-Variables Models

Estimateurs du ratio de fiabilité, propriétés asymptotiques et utilisation pour l'inférence dans les modèles structurels linéaires avec erreur sur les variables

The reliability ratio plays an important role in inference in linear structural errors-in-variables models with univariate observations. Moreover, some inference requires an assumed knowledge of this ratio. We revisit known estimators of the reliability ratio, including the ones relying on prior reliability studies, and propose some new ones. We establish consistency and asymptotic normality for our proposed estimators, as well as for the least squares estimators of the slope and intercept adjusted with our estimators of the reliability ratio, assuming most general moment conditions on the explanatory variables and measurement errors.

Le ratio de fiabilité joue un rôle important dans l'inférence des modèles structurels linéaires avec erreur sur les variables pour des observations univariées. De plus, certaines inférences nécessitent une connaissance supposée de ce ratio. Nous réexaminons les estimateurs connus du ratio de fiabilité, y compris ceux qui s'appuient sur des études de fiabilité antérieures, et nous en proposons de nouveaux. Nous établissons la cohérence et la normalité asymptotique pour les estimateurs proposés, ainsi que pour les estimateurs des moindres carrés de la pente et de l'ordonnée à l'origine ajustés avec nos estimateurs du ratio de fiabilité, en supposant les conditions de moment les plus générales sur les variables explicatives et les erreurs de mesure.

Chair/Président: William Ruth

Room/Salle: C 2033

Date: Monday June 3 / lundi 3 juin

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-15:45]

Glen McGee (University of Waterloo) **Lan Wen** (University of Waterloo)

Estimating Average Causal Effects with Incomplete Exposure and Confounders

Estimation des effets causaux moyens en cas d'exposition incomplète et de facteurs de confusion

Standard methods for estimating average causal effects require complete observations of the exposure and confounders. In observational studies, however, missing data are ubiquitous. We propose methods for estimating average causal effects when exposures and potential confounders may be missing. We consider missingness at random (MAR) and missingness not at random (MNAR) assumptions. Under each setting, we show that the average causal effects are non-parametrically identified and propose targeted maximum likelihood estimators that are semi-parametric efficient and doubly robust, allowing misspecification of either (i) the outcome models or (ii) the exposure and missingness models. The proposed methods are suitable for binary (or any other) outcome types, and we apply them to a motivating study of the effect of opioid usage on all-cause mortality in the National Health and Nutrition Examination Survey (NHANES) data.

Les méthodes standard d'estimation des effets causaux moyens nécessitent des observations complètes de l'exposition et des facteurs de confusion. Cependant, dans les études d'observation, les données manquantes sont omniprésentes. Nous proposons des méthodes d'estimation des effets causaux moyens lorsque les expositions et les facteurs de confusion potentiels peuvent être manquants. Nous étudions les hypothèses de données manquantes au hasard et de données manquantes non au hasard. Dans chaque cas, nous démontrons que les effets causaux moyens sont identifiés de manière non paramétrique et nous proposons des estimateurs ciblés du maximum de vraisemblance qui sont semi-paramétriques efficaces et doublement robustes, permettant une erreur de spécification soit i) des modèles de résultats, soit ii) des modèles d'exposition et de données manquantes. Nous proposons des méthodes adaptées aux types de résultats binaires (ou autres) et les appliquons à une étude motivante des effets de la consommation d'opioïdes sur la mortalité toutes causes confondues à partir des données de l'enquête menée par le National Center for Health Statistics des États-Unis.

[15:45-16:00]

Yuliang Shi (University of Waterloo) **Yeying Zhu** (University of Waterloo) **Joel A. Dubin** (University of Waterloo)

Causal Inference on Missing Exposure via Robust Estimation

Inférence causale sur l'exposition manquante au moyen d'une estimation robuste

How to deal with missing data in observational studies is a common concern for causal inference. When the covariates are missing at random (MAR), multiple approaches have been provided to help solve the issue. However, if the exposure is MAR, few approaches are available and careful adjustments on both missingness and confounding issues are required. In this article, a new inverse probability weighting (IPW) estimator based on weighted estimating equations (WEE) is

Le traitement des données manquantes dans les études d'observation pose souvent problème pour l'inférence causale. Ainsi, de nombreuses approches ont été proposées pour résoudre ce problème lorsque les covariables sont manquantes de façon aléatoire. Cependant, il existe peu d'approches si l'exposition est manquante de façon aléatoire et il est nécessaire de procéder à des ajustements précis sur les questions de données manquantes et de facteurs de confusion. Dans cette présentation, nous proposons un nouvel estimateur de pondération par probabilité in-

proposed to incorporate weights from both the missingness and PS models, which can reduce the joint effect of extreme weights in finite samples. Additionally, we develop a triple robust (TR) estimator via WEE to further protect against the misspecification of the model. Based on the simulation studies, WEE also outperform others approaches in terms of bias and standard error. Finally, an application study is conducted to identify the causal effect of the presence of cardiovascular disease on mortality for COVID-19 patients.

verse créé à partir d'équations estimantes pondérées afin d'intégrer les pondérations des modèles de données manquantes et des modèles de score de propension, ce qui permet de réduire l'effet conjoint des pondérations extrêmes dans les échantillons finis. De plus, nous créons un estimateur triplement robuste à l'aide d'équations d'estimantes pondérées pour mieux éviter les erreurs de spécification du modèle. Il ressort d'études de simulation que les équations d'estimation pondérées sont également plus efficaces que d'autres approches en ce qui concerne le biais et l'erreur-type. Enfin, nous présentons une étude d'application permettant de déterminer l'effet causal de la présence d'une maladie cardiovasculaire sur la mortalité des patients atteints de la COVID-19.

[16:00-16:15]

Xiaoya Wang (University of Waterloo) **Richard J. Cook** (University of Waterloo) **Yeying Zhu** (University of Waterloo)

Two-stage Regression for Causal Inference Involving Semi-continuous Exposures and Two-dimensional Propensity Scores

Régression en deux étapes pour l'inférence causale impliquant des expositions semi-continues et des scores de propension bidimensionnels

Methods for causal inference with binary treatment have recently been extended to deal with continuous exposures. In many settings however, including our motivating study on the effects of prenatal alcohol exposure on child cognition, the exposure distribution is semi-continuous with a mass at zero with a sub-density characterizing exposure levels among exposed individuals. We develop a two-stage approach for modeling the causal effects of a semi-continuous exposure. In the first stage, the causal effect of the level of exposure is assessed among exposed individuals via a propensity score regression adjustment. In the second stage, the causal effect of the binary exposure is assessed via inverse probability weighted (IPW) and augmented inverse probability weighted (AIPW). We derive the large sample properties of the resulting estimators and construct joint confidence intervals for the causal effects. Simulation studies confirm good finite sample performance of the proposed estimators.

Les méthodes d'inférence causale avec traitement binaire ont récemment été étendues aux expositions continues. Cependant, dans de nombreux cas, y compris dans notre étude de motivation sur les effets de l'exposition prénatale à l'alcool sur la cognition de l'enfant, la distribution de l'exposition est semi-continue avec une masse à zéro et une sous-densité caractérisant les niveaux d'exposition parmi les individus exposés. Nous développons une approche en deux étapes pour modéliser les effets causaux d'une exposition semi-continue. Dans un premier temps, nous évaluons l'effet causal du niveau d'exposition sur les individus exposés par le biais d'un ajustement par régression du score de propension. Dans un deuxième temps, nous évaluons l'effet causal de l'exposition binaire au moyen de la pondération des probabilités inverses (IPW) et de la pondération des probabilités inverses augmentée (AIPW). Nous déduisons les propriétés de grand échantillon des estimateurs résultants et construisons des intervalles de confiance conjoints pour les effets causaux. Des études de simulation confirment la bonne performance des estimateurs proposés sur des échantillons finis.

[16:15-16:30]

Henan Xu (University of Waterloo) **Yeying Zhu** (University of Waterloo)

Functional Mediation Analysis with Zero-inflated Count Data

Analyse fonctionnelle de la médiation à partir de données de dénombrement avec excès de zéros

Mediation analysis is crucial for understanding how treatments exert effects on outcomes via a mediator. Zero-inflated count outcomes and time-varying mediators are prevalent in fields such as biomedicine, economics, and social sciences. We extend existing methodologies by integrating a functional mediator in

L'analyse de médiation est essentielle pour comprendre la manière dont les traitements exercent des effets sur les résultats par l'intermédiaire d'un médiateur. Les résultats de dénombrement avec excès de zéros et les médiateurs variables dans le temps sont courants dans des domaines, tels que la biomédecine, l'économie et les sciences sociales. Nous étendons les méthodologies ac-

the context of zero-inflated count outcomes. The potential outcomes framework is employed to define the mediation effects of interest in this context and provide the theoretical underpinning for our approach, including conditions for effect identification. To address both the time-varying nature of the mediator and the zero-inflated count outcomes, functional linear and nonlinear models are implemented. Estimation and inference on the mediation effects are performed by a quasi-Bayesian Monte Carlo approximation method based on the mediation formula. Simulation studies validate our approach, demonstrating its capability to reliably estimate mediation effects in this context.

tuelles en intégrant un médiateur fonctionnel dans le contexte des résultats de dénombrement avec excès de zéros. Nous utilisons le cadre des résultats potentiels pour définir les effets de médiation pertinents dans ce contexte et fournir les fondements théoriques de notre approche, ainsi que les conditions d'identification de l'effet. Pour tenir compte de la nature variable dans le temps du médiateur et des résultats de dénombrement avec excès de zéros, des modèles fonctionnels linéaires et non linéaires sont mis en œuvre. Nous effectuons une estimation et déterminons l'inférence sur les effets de médiation à l'aide d'une méthode quasi-bayesienne d'approximation de Monte Carlo reposant sur la formule de médiation. Des études de simulation valident notre approche, démontrant sa capacité à estimer de manière fiable les effets de médiation dans ce contexte.

[16:30-16:45]

Sumeet Kalia (University of Manitoba) **Olli Saarela** (University of Toronto) **Michelle Greiver** (University of Toronto) **Frank Sullivan** (University of St. Andrews)

Continuous-time Causal Inference With Marked Point Process Weights: An Example on Sodium-Glucose Co-Transporters 2 Inhibitor Medications and Urinary Tract Infection

Inférence causale en temps continu avec des poids de processus ponctuels marqués : exemple des médicaments inhibiteurs du cotransporteur sodium-glucose de type 2 contre l'infection urinaire

The phenomena of treatment-confounder exist as mediating factors that predict the subsequent treatment in time-to-recurrent analysis. Conventional models produce misleading statistical inference of causal effects in the presence of time-dependent covariates due to conditioning on the causal pathways. Marginal structural models can be applied to quantify the causal treatment effect, estimated using longitudinal weights which mimic the randomization procedure by balancing the covariate distributions across the treatment groups. We formulated a continuous-time marginal structural model to access the effect of cumulative exposure of Sodium-Glucose co-Transporters 2 Inhibitor (SGLT-2i) medications on time-to-recurrent outcome of urinary tract infection (UTI). Our results supported the earlier findings in which the recurrent episodes of UTI did not increase when patients were prescribed low dose or high dose of SGLT-2i medications.

Les phénomènes de traitement-confusion existent comme facteurs médiateurs qui prédisent le traitement ultérieur dans l'analyse du temps de récurrence. Les modèles conventionnels produisent une inférence statistique erronée des effets causaux en présence de covariables dépendant du temps, en raison du conditionnement sur les liens de causalité. Il est possible d'appliquer des modèles structurels marginaux pour quantifier l'effet causal du traitement, estimé à l'aide de pondérations longitudinales qui imitent la procédure de randomisation par l'équilibrage des distributions des covariables entre les groupes de traitement. Nous avons créé un modèle structurel marginal en temps continu pour mesurer l'effet de l'exposition cumulative aux médicaments inhibiteurs du cotransporteur sodium-glucose de type 2 (SGLT-2i) sur le temps de récurrence de l'infection urinaire. Nos résultats confirment les conclusions antérieures selon lesquelles les épisodes récurrents d'infection urinaire n'augmentent pas lorsque les patients se voient prescrire une dose faible ou élevée de médicaments SGLT-2i.

Chair/Président: Qihuang Zhang

Room/Salle: C 3033

Date: Monday June 3 / lundi 3 juin

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-15:45]

Yu Shi (University of Toronto Dalla Lana School of Public Health)

Unsupervised Deep Domain Adaptation for Predicting Patient-Specific Cancer Dependency Maps

Adaptation profonde non supervisée par domaine pour prédire les cartes de dépendance du cancer propres aux patients

Cancer dependency maps are pivotal for identifying genes crucial to cancer cell proliferation, laying the groundwork for targeted treatment strategies. Despite the preservation of core biological processes, significant distribution discrepancies between cancer cell line (CCL) models and patient-derived data challenge the direct application of CCL findings to clinical practices. To address this, we introduce a machine learning algorithm utilizing deep unsupervised domain adaptation with CORAL loss, statistically align feature distributions between distinct data domains. Trained on labeled CCL data and validated against unseen CCL and unlabeled patient data from The Cancer Genome Atlas (TCGA), our model predicts patient dependency maps with better accuracy than compared baselines. This unsupervised approach not only precisely predicts cancer dependency for patient-derived tumors but also informs the promising advancement of out-of-distribution generalization in therapeutic interventions.

Les cartes de dépendance du cancer sont essentielles pour identifier les gènes qui jouent un rôle crucial dans la prolifération des cellules cancéreuses, jetant ainsi les bases de stratégies de traitement ciblées. Malgré la préservation des processus biologiques fondamentaux, les écarts de distribution importants entre les modèles de lignées cellulaires cancéreuses et les données provenant des patients compliquent l'application directe des résultats des lignées cellulaires cancéreuses aux pratiques cliniques. Pour remédier à ce problème, nous proposons un algorithme d'apprentissage automatique utilisant l'adaptation profonde non supervisée par domaine avec une perte d'alignement de corrélation (CORAL) afin de faire correspondre statistiquement les distributions de caractéristiques entre des champs de données distincts. Notre modèle, entraîné sur des données de lignées cellulaires cancéreuses étiquetées et validé à partir de données de patients non étiquetés et de l'Atlas du génome du cancer, prédit les cartes de dépendance des patients avec une précision supérieure à celle des données de référence comparées. Cette approche non supervisée permet non seulement de prédire avec précision la dépendance au cancer pour les tumeurs des patients, mais aussi d'exploiter les avancées prometteuses de la généralisation en dehors de la distribution dans les interventions thérapeutiques.

[15:45-16:00]

Muditha L. Bodawatte Gedara (University of Manitoba) **Lisa M. Lix** (Department of Community Health Sciences, University of Manitoba) **Ridwan Sanusi** (Smart Mobility and Logistics, King Fahd University of Petroleum & Minerals) **Tolulope Sajobi** (Department of Community Health Sciences, University of Calgary)

Comparison of Cross-validation Methods for Tree-based Item-Focused Models to Detect Differential Item Functioning in Patient-reported Outcome Measures

Comparaison de méthodes de validation croisée pour des modèles d'arbres axés sur les items (IFT) pour la détection du fonctionnement différentiel d'un item (DIF) dans les mesures de résultats rapportés par les patients

The item-focussed tree (IFT) model, a model-based recursive partitioning, combines a classification tree

Le modèle d'arbre basé sur les items (IFT), un partitionnement récursif basé sur un modèle, combine un arbre de classification

with logistic regression to detect differential item functioning (DIF). DIF is a measurement bias that occurs when patients with the same health status do not interpret patient-reported outcome measures (PROMs) items similarly. Our study purpose was to compare the generalizability of DIF analyses for the IFT model in k-fold and holdout cross-validation (CV) methods using real-world clinical and simulated data. Real-world data included 247 patients. Simulation parameters were sample size, number of items, and DIF effect sizes in three items induced by five binary variables. In clinical data 5-fold CV identified DIF for five of seven items, while both holdout and 2-fold CV methods identified two DIF items, but on different variables. In simulations, holdout CV (0.24) had lower misclassification error rates than k-fold CV (0.33), suggesting its suitability for large samples (¿100).

avec une régression logistique pour la détection du fonctionnement différentiel d'un item (DIF). Celui-ci est un biais de mesure qui se produit quand des patients dont l'état de santé est le même n'interprètent pas les items dans l'outil de mesures des résultats rapportés par les patients (PROM) de la même manière. Le but de notre étude était de comparer la généralisabilité des analyses du fonctionnement différentiel d'un item pour le modèle IFT dans les méthodes de validation croisée à k plis et avec échantillon test (holdout) à l'aide de données cliniques réelles et de données simulées. Les données réelles portaient sur 247 patients. Les paramètres de la simulation étaient la taille de l'échantillon, le nombre d'items et les tailles des effets du DIF induits par cinq variables binaires dans trois items. Dans les données cliniques, la validation croisée à 5 plis a identifié le DIF pour cinq des sept éléments, tandis que les méthodes avec échantillon test et 2 plis ont identifié deux éléments du DIF, mais pour différentes variables. Dans les simulations, la validation croisée échantillon test (0,24) avait un taux d'erreur de classification plus faible que la validation croisée k plis (0,33), ce qui suggère sa convenance à de grands échantillons (¿100).

[16:00-16:15]

Ziqian Zhuang (University of Toronto Dalla Lana School of Public Health) **Wei Xu** (University Health Network, Biostatistics)

Joint Modeling of Complex Multivariate Adverse Events in Clinical Trial Data

Modélisation conjointe des événements indésirables multivariés complexes dans les données d'essais cliniques

Adverse events (AE) are harmful outcomes during medical care. The severity and frequency of these events serve as study endpoints in clinical trials, crucial for evaluating treatment safety. Patients may encounter multiple adverse events concurrently, and the recorded data exhibit diverse structures due to varying durations and characteristics of both short-term and long-term AEs. Moreover, AE severity may fluctuate over time due to disease progression or treatment response. Most current analyses focus solely on a single AE, neglecting severity information and failing to distinguish adequately between short-term and long-term AEs. In response, we propose an efficient joint model to assess treatment effects on multiple AE occurrences. This model comprehensively considers AE severities and correlations while effectively addressing structural differences between short-term and long-term AEs. Through simulation studies, this method has demonstrated high accuracy in parameter estimation.

Les événements indésirables (EI) surviennent lors des soins médicaux. Leur gravité et fréquence sont des critères d'évaluation essentiels dans les essais cliniques pour la sécurité des traitements. Les patients peuvent avoir plusieurs EI simultanément, avec des données enregistrées de structures diverses en raison de la durée et des caractéristiques variables des EI à court et long terme. La gravité des EI peut fluctuer avec la progression de la maladie ou la réponse au traitement. La plupart des analyses se concentrent sur un seul EI, négligeant la gravité et sans distinguer les EI à court et long terme. Nous proposons un modèle conjoint pour évaluer les effets du traitement sur plusieurs EI. Ce modèle tient compte des gravités et corrélations des EI tout en traitant les différences structurelles entre les EI. Les études de simulation ont démontré une grande précision dans l'estimation des paramètres.

[16:15-16:30]

Aoqi Xie (University of Toronto) **Peijin Wang** (School of Medicine, Duke University) **Aya A. Mitani** (Dalla Lana School of Public Health, University of Toronto) **Madison Aitken** (Department of Psychology, York University; Centre for Addiction and

Mental Health) **Wendy Lou** (Dalla Lana School of Public Health, University of Toronto) **Clement Ma** (Dalla Lana School of Public Health, University of Toronto; Centre for Addiction and Mental Health)

A novel Sequential Multiple Assignment Randomized Trial (SMART) design with Internal Pilot Study and Unblinded Sample Size Re-estimation

Un nouveau concept d'essai randomisé séquentiel à évaluation multiple (SMART) avec une étude pilote interne et une réestimation de la taille d'échantillon non aveugle

Sequential Multiple Assignment Randomized Trials (SMART) provide a framework for testing dynamic treatment strategies, which are pre-specified rules for re-randomizing participants based on their responses. We extend the fixed SMART design to include an Internal Pilot Study for feasibility and unblinded Sample Size Re-estimation (SSR). We apply the one-stage exact test for feasibility, and Adjusted Effect Size (AES) and Cui, Hung, Wang (CHW) methods for SSR. We assess the statistical properties of our design in terms of expected re-estimated sample size, type I error rate and statistical power using simulations. Results show that both SSR methods perform well in controlling type I error inflation and keeping the power over 80% after appropriate adjustment. Both methods recover the power compared to the fixed design when the true effect size is smaller than planned, while the CHW method performs slightly better in terms of power and requires a smaller sample size than the AES method.

Les essais randomisés séquentiels à évaluation multiple (SMART) procurent un cadre pour tester les stratégies de traitement dynamiques, qui sont des règles préétablies pour randomiser à nouveau des participants en fonction de leurs réponses. Nous élargissons le concept SMART fixe pour y insérer une étude pilote interne pour la faisabilité et une réestimation de la taille d'échantillon (SSR) non aveugle. Nous appliquons le test de faisabilité à une étape, la taille d'effet ajustée (AES), puis les méthodes Cui, Hung et Wang (CHW) pour la SSR. Nous évaluons les propriétés statistiques de notre conception en termes de taille d'échantillon réestimée attendue, de taux d'erreur de type I et de puissance statistique à l'aide de simulations. Les résultats démontrent que les deux méthodes SSR fonctionnent bien pour le contrôle de l'inflation d'erreur de type I et pour conserver une puissance supérieure à 80 % après un ajustement adéquat. Les deux méthodes récupèrent la puissance contrairement à celle qui est fixe lorsque la vraie taille d'effet est plus petite que prévu, alors que la méthode CHW fonctionne un peu mieux en matière de puissance et nécessite une taille d'échantillon plus petite que pour la méthode AES.

[16:30-16:45]

Qirui (Dylan) Hou (University of Toronto) **Amy Liu** (Princess Margaret Cancer Centre, University Health Network) **Peter Szatmari** (Centre for Addiction and Mental Health) **Clement Ma** (Centre for Addiction and Mental Health)

A Novel Multi-Arm, Two-stage Basket Design

Nouveau modèle de panier à deux étapes et à bras multiples

Introduction: Chen et al. proposed a two-stage randomized basket design, which tests l=1 treatment vs control across various cancer indications, We expand this design to evaluate l¿1 treatments vs a common control. Methods: We extended Chen's Type I error rates and power calculations. Our design allocates samples evenly to each basket, with additional samples added strategically at the second stage. A basket is pruned only if none of its treatments are effective at a predefined alpha level during the interim. We incorporated the Dunnett test to allow for the pruning of underperforming treatments while pooling effective ones, adjusting the covariance matrix based on Dunnett's findings. Simulations were used to validate our calculation results. Results: Our validation results show that for l=1 treatment, our result of type I error is very similar to Chen's design. For l¿1, we derived a reasonable result using a generalized com-

Introduction : Chen et al. ont proposé un modèle de panier randomisé en deux étapes, qui teste le traitement l=1 contre un contrôle dans diverses indications de cancer. Nous élargissons ce modèle pour évaluer les traitements l¿1 contre un contrôle commun. Méthodes : Nous avons étendu les taux d'erreur de type I et les calculs de puissance de Chen. Notre plan répartit les échantillons de manière égale dans chaque panier, des échantillons supplémentaires étant ajoutés de manière stratégique lors de la deuxième étape. Un panier n'est élagué que si aucun de ses traitements n'est efficace à un niveau alpha prédéfini pendant la période intermédiaire. Nous avons incorporé le test de Dunnett pour permettre l'élagage des traitements sous-performants tout en regroupant les traitements efficaces, en ajustant la matrice de covariance sur la base des résultats de Dunnett. Nous utilisons des simulations pour valider les résultats de nos calculs. Résultats : Nos résultats de validation montrent que pour le traitement l=1, notre résultat de l'erreur de type I est très similaire à la conception de Chen.

putational equation with validated through simulation.

Pour l¿1, nous avons obtenu un résultat raisonnable en utilisant une équation de calcul généralisée validée par des simulations.

[16:45-17:00]

Sirikkathuge Ishanka Randini Fernando (McMaster University)

Time-aligned Latent Dirichlet Allocation for Longitudinal Microbiome Data

Répartition de Dirichlet latente alignée dans le temps pour les données longitudinales de microbiome

Microbial communities are dynamic, evolving through interactions within taxa and in response to environmental changes. Longitudinal studies are increasingly critical for unraveling these complex temporal dynamics, providing insights into microbial functionality and interdependencies. However, traditional clustering methods often fail to analyze microbial data due to their inherent sparsity, high dimensionality, and heterogeneity, along with a failure to recognize samples' potential in belonging to multiple clusters. In this context, the probabilistic Latent Dirichlet Allocation (LDA) topic model emerges as a superior alternative. Our study introduces an adaptation of time-aligned LDA designed specifically for longitudinal microbiome analyses. This novel framework focus on aligning microbial topics across sequential time points, thereby addressing the unique challenges of time-variant data. Further, it involves constructing credible intervals from posterior samples and quantitatively delineating the differences in topic proportions across experimental conditions facilitated by a linear mixed model. Applying this framework to gut microbial specimens from pregnant women participating in the Be Healthy in Pregnancy study, we observed enhanced sensitivity in identifying significant temporal dynamics of microbial communities during and after pregnancy compared to the standard LDA.

Les communautés bactériennes sont dynamiques, et évoluent par des interactions à l'intérieur de taxons et en réponse à des changements environnementaux. Les études longitudinales sont de plus en plus cruciales pour démêler ces dynamiques temporelles complexes, et ainsi fournir des informations sur la fonction bactérienne et les interdépendances dans le microbiome. Cependant, les méthodes de regroupement traditionnelles peinent à analyser les données bactériennes à cause de leur éparsité inhérente, de leur grande dimension et de leur hétérogénéité, ainsi qu'une incapacité à reconnaître la possibilité que des échantillons appartiennent à plusieurs groupes. Dans ce contexte, le modèle probabiliste de répartition de Dirichlet latent (LDA) représente une option supérieure. Notre étude présente une adaptation de LDA alignée dans le temps conçue spécifiquement pour l'analyse longitudinale du microbiome. Ce nouveau cadre de travail se concentre à aligner les schémas bactériens à travers des points séquentiels dans le temps, et résout donc les défis particuliers liés aux données variant dans le temps. De plus, ce cadre comprend la construction d'intervalles de crédibilité à partir d'échantillons a posteriori et la définition quantitative des différences dans les proportions de thèmes sous différentes conditions expérimentales facilitées par un modèle linéaire mixte. Nous avons appliqué cette méthode à des spécimens de bactéries intestinales provenant de femmes enceintes participant à l'étude ≪ Be Healthy in Pregnancy ≫ et avons observé une sensibilité accrue pour l'identification de dynamiques temporelles significatives dans des communautés bactériennes pendant et après la grossesse par rapport à la LDA standard.

CRM-SSC Prize in Statistics Invited Address
Allocution du recipiendaire du Prix CRM-SSC en statistique

Chair/Président: Erica E. M. Moodie

Organizer/Responsable: Erica E. M. Moodie

Room/Salle: IIC 2001

Date: Tuesday June 4 / mardi 4 juin

Time/Heure: 08:30-09:50

Abstract/Résumé

[08:30-09:50]

Alexandre Bouchard-Côté (University of British Columbia)

Computational Lebesgue integration

Intégration computationelle de Lebesgue

In many modern applications in science and engineering, we seek to reconstruct a complicated object x from noisy data y, for example, one may seek to reconstruct an evolutionary tree from sequencing data. In principle, Bayesian statistics provides a broad framework to approach such problems, by modelling knowns and unknowns as random variables X and Y. Since the notion of a posterior distribution, X—Y, is defined under very general conditions, Bayesian inference is in a sense universal for the purpose of data analysis. In contrast, other inferential setups often require, among other things, for x to be real-valued in order to use approximations such as those based on the central limit theorem. However, this generality hinges on being able to approximate expectation with respect to an arbitrary measure. Can we develop generic sampling methods in such an unstructured context? Surprisingly, practical generic methodologies are indeed possible. I will review the literature and describe some of our work in the area with a focus on recent developments based on non-reversible and regenerative MCMC. My group is also working on making these complex Monte Carlo methods easy to use: check out https://pigeons.run/dev/, a package allowing the user to leverage clusters of 1000s of nodes to speed-up difficult Monte Carlo problems without requiring knowledge of distributed algorithms.

Dans de nombreuses applications modernes en science et en ingénierie, nous cherchons à reconstruire un objet compliqué x à partir de données bruyantes y. Par exemple, on peut chercher à reconstruire un arbre évolutif à partir de données de séquençage. En principe, la statistique bayésienne fournit un cadre général pour aborder de tels problèmes, en modélisant les objets connus et inconnus comme des variables aléatoires X et Y. Comme la notion de distribution a posteriori, X—Y, est définie dans des conditions très générales, l'inférence bayésienne est en quelque sorte universelle aux fins de l'analyse des données. En revanche, d'autres systèmes d'inférence exigent souvent, entre autres, que x soit à valeur réelle avant de pouvoir utiliser des approximations telles que celles du théorème de la limite centrale. Cependant, cette généralité dépend de la possibilité d'approximer l'espérance par rapport à une mesure arbitraire. Peut-on développer des méthodes d'échantillonnage génériques dans un tel contexte non structuré ? De manière surprenante, des méthodologies génériques pratiques sont en effet possibles. Je passerai en revue la littérature et décrirai certains de nos travaux dans ce domaine, en mettant l'accent sur les développements récents basés sur la MCMC non réversible et régénérative. Mon groupe s'efforce également de rendre ces méthodes de Monte Carlo complexes faciles à utiliser : consultez https ://pigeons.run/dev/, un progiciel permettant à l'utilisateur d'exploiter des grappes de plusieurs milliers de nœuds pour accélérer des problèmes de Monte Carlo difficiles sans connaissances préalables en algorithmes distribués.

Chair/Président: You Liang

Organizer/Responsable: You Liang

Room/Salle: A 1045

Date: Tuesday June 4 / mardi 4 juin

Time/Heure: 10:20-11:50

Abstract/Résumé

[10:20-10:50]

Longhai Li (University of Saskatchewan)

Z-Residual Diagnostic Tool for Assessing Covariate Functional Form in Proportional Hazards Models with Shared Frailty

Outil diagnostique Z-résiduel pour l'évaluation de la forme fonctionnelle des covariables dans les modèles de risques proportionnels avec fragilité partagée

Survival analysis often involves modelling hazard functions while considering frailty to account for unobserved cluster-level factors in clustered survival data. Shared frailty models have gained popularity for this purpose, but assessing covariate functional form in these models presents unique challenges. Martingale and deviance residuals are commonly used for visually assessing covariate functional form against continuous covariates. Nevertheless, their subjective nature and lack of a reference distribution make it challenging to derive numerical statistical tests from these residuals. To address these limitations, we propose "Z-residuals", a novel diagnostic tool designed for shared frailty models, leveraging the concept of randomized survival probability, and introducing both graphical and numerical tests. To implement this approach, we develop an R package to compute Z-residuals for shared frailty models. Through extensive simulation studies, we demonstrate the high power of our derived numerical test for assessing the functional form of covariates. To validate the effectiveness of our method, we apply it to a real data application concerning the modelling of survival time for acute myeloid leukemia patients. Our Z-residual diagnosis results reveal the inadequacy of log-transformation of the covariate, highlighting the limitations of other diagnostic methods for effectively assessing covariate functional form in shared frailty models.

L'analyse de survie comporte souvent la modélisation de fonctions de risque, tout en considérant la fragilité pour tenir compte des facteurs non observés au niveau des grappes dans les données de survie en grappes. À cet égard, les modèles de fragilité partagée ont gagné en popularité, mais l'évaluation de la forme fonctionnelle des covariables dans ces modèles présentent des problèmes uniques. La martingale et le résidu de déviance sont couramment utilisés pour évaluer visuellement la forme fonctionnelle des covariables par rapport aux covariables continues. Cependant, en raison de la nature subjective des résidus de martingale et de déviance qui manquent aussi de distribution de référence, il est difficile d'en dériver des tests statistiques. Pour aborder ces limites, nous proposons les ≪ Z-résidus ≫, un nouvel outil de diagnostic conçu pour les modèles de fragilité partagée, tirant parti du concept de probabilité de survie aléatoire et introduisant des tests à la fois graphiques et numériques. Pour implémenter notre approche, nous développons un package R pour calculer les Z-résidus de modèles de fragilité partagée. Nos études de simulation approfondies montrent le grand pouvoir du test numérique dérivé pour vérifier la forme fonctionnelle des covariables. Aux fins de validation de l'efficacité de notre méthode, nous l'appliquons à une application de données réelles sur la modélisation de la durée de survie de patients atteints de leucémie myéloïde aiguë. Les résultats du diagnostic Z-résiduel montrent qu'un modèle avec transformation logarithmique des covariables, mettant ainsi en lumière les limites d'autres méthodes diagnostiques pour une évaluation efficace de la forme fonctionnelle des covariables dans les modèles de fragilité partagée.

[10:50-11:20]

Li Xing (University of Saskatchewan)

Concurrent Prediction of Multiple Survival Outcomes with a Refined Stacking Algorithm

Prédiction concurrente de plusieurs résultats de survie avec un algorithme d'empilage raffiné

Xing et al. (2019) developed prediction algorithms, termed multi-task prediction algorithms using revised stacking (MTPS), to conduct the concurrent prediction for multiple outcome variables with high-dimensional predictors integrated into the prediction process. Their algorithms employed the strategy of the stacking algorithm to construct a multi-task learner with the flexibility to handle a mixed type of continuous and binary outcomes. However, they are not applicable to survival data, where the analysis is typically challenged by the issues of censoring. Expanding their work to handle multiple survival outcomes, we develop a new concurrent prediction algorithm by utilizing the revised residual stacking framework, where the parametric accelerated failure time (AFT) model and Elastic Net AFT model are employed. Through simulation studies and data application, we demonstrate that the novel enhancement of MTPS for survival outcomes surpasses the performance of their single learners.

Xing et al. (2019) ont conçu des algorithmes de prédiction (algorithmes de prédiction multitâche avec empilage révisé, ou MTPS) afin de mener la prédiction concurrente de plusieurs variables de résultat avec des prédicteurs de grande dimension intégrés dans le processus de prédiction. Leurs algorithmes adoptent la stratégie de l'algorithme d'empilage pour construire un système d'apprentissage multitâche ayant la flexibilité pour traiter plusieurs types de résultats binaires et continus. Cependant, ils ne sont pas utilisables pour les données de survie, car l'analyse est habituellement remise en question par le problème de censure. Afin de permettre à leur travail de traiter plusieurs résultats de survie, nous développons un nouvel algorithme de prédiction concurrent en nous servant du cadre d'empilage résiduel révisé, dans lequel on emploie le modèle de temps d'échec accéléré (AFT) et le modèle AFT à "Elastic Net". Par l'entremise d'études en simulation et d'application à des données, nous démontrons que cette nouvelle amélioration au MTPS pour les résultats de survie dépasse la performance de leurs systèmes d'apprentissage simples.

Chair/Président: Yanqing Yi

Organizer/Responsable: Yanqing Yi

Room/Salle: C 2033

Date: Tuesday June 4 / mardi 4 juin

Time/Heure: 10:20-11:50

Abstract/Résumé

[10:20-10:50]

Grace Y. Yi (University of Western Ontario) **Yasin Khadem Charvadeh** (University of Western Ontario)

Accommodating Misclassification Effects on Optimizing Dynamic Treatment Regimes with Q-Learning

Accommodation des effets de l'erreur de classification pour l'optimisation des régimes de traitement dynamique avec Q-learning

Research on dynamic treatment regimes has sparked extensive interest. Many methods have been proposed in the literature, which, however, are vulnerable to the presence of misclassification in covariates. In particular, although Q-learning has received considerable attention, its applicability to data with misclassified covariates is unclear. In this talk, I will discuss how ignoring misclassification in binary covariates can impact the determination of optimal decision rules in randomized treatment settings, and demonstrate its deleterious effects on Q-learning through empirical studies. I will describe correction methods to address misclassification effects on Q-learning. Numerical studies reveal that misclassification in covariates induces non-negligible estimation bias and that the correction methods successfully ameliorate bias in parameter estimation.

La recherche sur les régimes de traitement dynamique a suscité un grand intérêt. Plusieurs méthodes ont été proposées dans la littérature, mais elles sont par contre vulnérables à la présence d'erreurs de classification dans les covariables. En fait, même si Q-learning a reçu beaucoup d'attention, il n'est pas clair si cette méthode est utilisable avec des données contenant des erreurs de classification des covariables. Lors de cet exposé, je parlerai des conséquences d'ignorer les erreurs de classification dans les covariables binaires sur la détermination de règles décisionnelles optimales dans le cadre de traitement randomisé. Puis je démontrerai ses effets nuisibles sur Q-learning à travers des études empiriques. Je décrirai des méthodes de correction pour régler les effets des erreurs de classification sur Q-learning. Des études numériques révèlent que les erreurs de classification dans les covariables génèrent un biais d'estimation non négligeable et que les méthodes de correction réussissent à surmonter le biais dans l'estimation du paramètre.

[10:50-11:20]

Xikui Wang (University of Manitoba)

Bayesian Adaptive Design of Phase I Clinical Trials

Conception adaptative bayésienne d'essais cliniques de phase I

As the first phase of experiment on human subjects, dose finding clinical trials is complex in ethics and methodology. We not only need to minimize the overall risk of toxicity and treat efficiently as many patients in the trial as possible, but also reliably identify the maximum tolerated dose which is crucial for the success of subsequent clinical trials. In this talk, we discuss the use of Bayesian designs that strike a good balance between the collective ethics and individual ethics.

Dans la première phase d'une expérimentation sur des sujets humains, les essais cliniques de recherche de dose sont complexes sur le plan de l'éthique et de la méthodologie. Nous devons non seulement minimiser le risque global de toxicité et traiter efficacement le plus grand nombre possible de patients dans l'essai, mais aussi identifier de manière fiable la dose maximale tolérée, ce qui est crucial pour le succès des essais cliniques ultérieurs. Dans cet exposé, nous discuterons de l'utilisation de modèles bayésiens qui établissent un bon équilibre entre l'éthique collective et l'éthique individuelle.

Chair/Président: Yingli Qin

Organizer/Responsable: Yingli Qin

Room/Salle: SN 2109

Date: Tuesday June 4 / mardi 4 juin

Time/Heure: 10:20-11:50

Abstract/Résumé

[10:20-10:50]

Liqun Diao (University of Waterloo) **Yechao Meng** (University of Prince Edward Island) **Chengguo Weng** (University of Waterloo)

Mortality Prediction via Age-Specific Band Selection

Prévision de la mortalité par sélection de bandes spécifiques à l'âge

The talk will introduce a novel mortality prediction framework called age-specific band selection. This innovative approach leverages information from nearby age groups to train prediction models for each age in a mortality table. Furthermore, this concept is expanded to incorporate data from multiple populations, of which three distinct methods will be discussed: a distance-based approach, an ensemble approach, and an ACF model-based approach. The talk will showcase the results of extensive empirical studies conducted using the Human Mortality Database. These results will demonstrate the superior prediction accuracy of the proposed methods compared to relevant benchmark models.

Cet exposé présentera un nouveau cadre de prévision de la mortalité nommé sélection de bandes spécifiques à l'âge. Cette approche innovatrice tire avantage des renseignements relatifs aux groupes d'âge environnants pour informer les modèles de prédiction pour chaque âge dans un tableau de mortalité. De plus, ce concept est élargi pour intégrer des données provenant de plusieurs populations, ce que nous ferons avec trois approches : une basée sur la distance, une méthode d'ensemble, et une utilisant un modèle basé sur l'ACF. Cet exposé présentera les résultats d'études empiriques approfondies menées sur les données de mortalité humaine de la Human Mortality Database. Ces résultats démontreront la fiabilité supérieure de prévision des méthodes proposées par rapport aux modèles de références.

[10:50-11:20]

Hong Li (University of Guleph) **David Landriault** (University of Waterloo) **Bin Li** (University of Waterloo) **Yuanyuan Zhang** (University of Waterloo)

Risk Aversion and Longevity Risk Transfers: Reinsurance vs. Capital Market Solutions

Aversion pour le risque et transfert du risque de longévité : réassurance et solutions du marché des capitaux

This paper develops an economic framework to analyze optimal longevity risk transfers in both the reinsurance and capital markets, focusing on the differing risk aversions of buyers and sellers of longevity risk transfer contracts. Utilizing a Stackelberg game framework, we compare static longevity swap contracts, offering long-term protection with constant hedge ratios and predetermined hedging costs, against dynamic contracts, providing short-term coverage with variable contract terms. With real-life mortality data, our numerical analysis reveals that static contracts are preferred in the

Cet article élabore un cadre économique afin d'analyser les transferts optimaux de risque de longévité pour la réassurance et les marchés financiers, en se concentrant les différentes aversions pour le risque des acheteurs et les vendeurs de contrats de transfert de risque de longévité. En nous servant du duopole de Stackelberg, nous comparons les contrats de swap à longévité statique, offrant une protection à long terme avec des ratios de couverture constants, contre des contrats dynamiques, qui procurent une couverture à court terme avec des modalités de contrats à capital variable. Au moyen de données réelles sur la mortalité, nos analyses numériques démontrent que les contrats statiques sont préférables

reinsurance market as they lead to larger welfare gains for both participating parties and more flexible conditions for market existence. Conversely, dynamic contracts are favored in the capital market due to sellers' higher risk aversion. Additionally, information asymmetry is incorporated in the form of ambiguity. While ambiguity reduces welfare gains for both parties and leads to more stringent conditions for market existence, it does not alter the contract preferences. Our analysis provides theoretical explanations for several key empirical observations in the current longevity risk transfer market and offers new insights into the development of the longevity-linked capital market.

dans le marché de la réassurance, car ils génèrent davantage de gains d'aide sociale pour les parties participantes et procurent des conditions flexibles pour l'existence du marché. Inversement, les contrats dynamiques sont plus avantageux dans le marché financier en raison de l'aversion pour le risque des vendeurs. De plus, nous intégrons l'asymétrie de l'information sous la forme d'ambiguïté. L'ambiguïté réduit les gains d'aide sociale pour les deux parties et restreint les conditions d'existence du marché, mais elle ne change pas les préférences relatives aux contrats. Notre analyse procure des explications théoriques concernant certaines observations empiriques cruciales dans le marché du transfert de risque de longévité et apporte de nouvelles perspectives dans le développement du marché financier concernant la longévité.

[11:20-11:50]

Yechao Meng (University of Prince Edward Island)

Mortality Prediction: a Parameter Transfer Approach

Prédiction de mortalité : une approche de transfert de paramètre

Borrowing information from populations with similar mortality patterns is a recognized strategy for the mortality prediction of a target population. This mirrors the concept of Transfer Learning, a popular and promising area in modern data mining and machine learning, which aims at improving the performance of target learners on target domains by transferring the knowledge contained in different but related source domains. This project focuses on applying transfer learning to actuarial applications of mortality predictions. We explore how data from other mortality datasets can be effectively integrated into the parameter transfer learning framework to improve mortality predictions for a target population. Our approach includes incorporating existing mortality prediction models into a regularization framework with closed-form solutions. Additionally, we develop an iterative updating algorithm for classic mortality models and penalty forms.

Emprunter des informations sur des populations ayant des tendances de mortalité similaires est une stratégie reconnue pour la prédiction de mortalité d'une population cible. Cela reflète le concept d'apprentissage par transfert, une notion prometteuse et populaire dans les domaines modernes d'exploration de données et d'apprentissage automatique, qui cherche à améliorer la performance de systèmes d'apprentissages cibles pour des domaines ciblés, en transférant le savoir contenu dans différents domaines source. Ce projet s'intéresse à appliquer l'apprentissage par transfert à des applications actuarielles de prédictions de mortalité. Nous explorons de quelle façon les données provenant d'autres ensembles de données sur la mortalité peuvent être intégrées dans le cadre d'apprentissage par transfert de paramètre dans le but d'améliorer les prédictions de mortalité pour une population cible. Notre approche comprend l'intégration des modèles de prédiction de mortalité actuels dans un cadre de régularisation avec solutions analytiques. De plus, nous développons un algorithme de mise à jour itérative pour les modèles de mortalité classiques et les formes de pénalité.

Teaching Introductory Statistics to Non-specialists
Enseigner l'introduction à la statistique à des non-spécialistes

Chair/Président: Léo Belzile

Organizer/Responsable: Léo Belzile

Room/Salle: A 1046

Date: Tuesday June 4 / mardi 4 juin

Time/Heure: 10:20-11:50

Abstract/Résumé

[10:20-10:50]

Tiffany A. Timbers (The University of British Columbia)

Reflections on Scaling an Introduction to Data Science

Réflexions sur la généralisation d'une introduction à la science des données

Five years ago the Statistics Department at University of British Columbia launched an introductory data science course targeted at first year students. This new course started out small (with just 59 students) and aimed to use evidence-based pedagogical practices (e.g., flipped classroom, pair-programming), open educational resources, and modern technology platforms (R and the tidyverse in Jupyter notebooks, autograding using nbgrader, and a JupyterHub integrated with Canvas for authentication). Since its instantiation in 2019, the course has scaled dramatically, serving over 2300 students at UBC this year, and is now offered in a choice of programming languages (either R or Python). Here we reflect upon and discuss what has worked and what hasn't. We consider the perspectives from the students and the instructional team.

Il y a cinq ans, le Département de statistique de l'Université de la Colombie-Britannique a lancé un cours d'introduction à la science des données destiné aux étudiants de première année. Ce nouveau cours a commencé à petite échelle (avec seulement 59 étudiants) et visait à utiliser des pratiques pédagogiques fondées sur des preuves (par exemple, classe inversée, programmation en binôme), des ressources éducatives ouvertes et des plateformes technologiques modernes (R et tidyverse dans les carnets Jupyter, autogradage à l'aide de nbgrader, et un JupyterHub intégré à Canvas pour l'authentification). Depuis sa création en 2019, le cours s'est considérablement développé, desservant plus de 2 300 étudiants à UBC cette année, et est maintenant offert dans un choix de langages de programmation (soit R ou Python). Ici, nous réfléchissons et discutons de ce qui a fonctionné et de ce qui n'a pas fonctionné. Nous prenons en compte les points de vue des étudiants et de l'équipe pédagogique.

[10:50-11:20]

Nathalie Moon (University of Toronto)

Principles and Practices in Teaching STA130: Introduction to Statistical Reasoning and Data Science at the University of Toronto - A Collaborative Partnership Approach

Principes et pratiques de l'enseignement de STA130 : Introduction au raisonnement statistique et à la science des données à l'Université de Toronto - Un partenariat collaboratif

At UofT, students explore a range of interests before choosing their program. Given this, our first-year course STA130 is designed to be accessible to students from all backgrounds. Each year, over 1,100 students enroll, and coordination of this course is a team effort. Instructors work with a large team of 25 teaching assistants, a mentorship coordinator and a team of peer mentors, the English Language Learning unit, and more

À l'Université de Toronto, les étudiants explorent un large éventail d'intérêts avant de choisir leur programme. C'est pourquoi notre cours de première année STA130 est conçu pour être accessible à tous. Avec plus de 1100 étudiants par année, la coordination de ce cours est un travail d'équipe. Nous travaillons avec une équipe de 25 assistants d'enseignement, un coordinateur de mentorat et une équipe de mentors, l'équipe ≪ English Language Learning ≫, et plus encore, pour assurer le succès des étudiants, avec une at-

to support students effectively, with a special focus on the large proportion of international students. In this session, I will share the principles that guide my thinking around the design of this course, including choice of topics, low-stakes writing tasks, experiential learning, implementation of automated tests for programming assignments, and more. I will also share actionable suggestions on how to effectively coordinate a large multi-section class, from lessons I've learned teaching this course since 2018.

tention particulière aux besoins des étudiants internationaux qui forment une grande proportion des étudiants de ce cours. Je partagerai les principes qui guident ma planification de cours : le choix des sujets, les tâches d'écriture à faible enjeu, l'apprentissage par l'expérience, l'utilisation de tests automatisés pour les devoirs de programmation, et plus encore. Je partagerai également des suggestions pratiques sur la façon de coordonner une grande classe à sections multiples.

[11:20-11:50]

Carolyn Augusta (University of Saskatchewan)

One Size Does Not Fit All

Tous n'entrent pas dans le même moule

In the Edwards School of Business at the University of Saskatchewan, there are currently two undergraduate-level statistics courses: Foundations of Business Statistics (a first-year course, COMM 104) and Statistics for Business Decisions (a second-year course, COMM 207, which students often take in their fourth year). In each course, students are expected to use Excel. Each section of each course is large, and student preparedness varies greatly within and between terms. Some students are not comfortable with rearranging an equation or using a computer, and have not taken a math course in more than 10 years; others have taken calculus and have done some programming. Additionally, post pandemic re-opening, there has been an increase observed in the frequency of accommodation requests. Methods for addressing these disparities and providing flexibility for students in large enrolment non-specialist courses with limited teaching assistant resources will be proposed and described.

À la Edwards School of Business à la University of Saskatchewan, il existe présentement deux cours de statistique de premier cycle : « Fondations des statistiques d'entreprise » (cours de première année, COMM 104) et « Les statistiques dans les décisions d'entreprise » (cours de deuxième année, COMM 207, que les étudiants prennent souvent à leur quatrième année). Dans chaque cours, les étudiants doivent utiliser Excel. Chaque section de chacun des cours est volumineuse, et le degré de préparation des étudiants varie beaucoup au sein d'une même ainsi que d'une session à une autre. Certains ne sont pas à l'aise pour réarranger une équation ou utiliser un ordinateur, et n'ont pas suivis de cours de mathématique depuis plus de 10 ans, tandis que d'autres ont suivis des cours de calcul différentiel et intégral, et ont de l'expérience de programmation. De plus, depuis la fin de la pandémie, nous avons observé une augmentation des fréquences de demandes d'accommodation. Nous suggérerons et décrirons des méthodes afin d'aborder ces disparités et de fournir une flexibilité pour les étudiants dans des cours non spécialisés à inscription élevée avec des ressources d'assistance à l'enseignement limitées.

Addressing Practical Challenges in Longitudinal Causal Inference
Relever les défis pratiques de l'inférence causale longitudinale

Chair/Président: Mireille Schnitzer

Organizer/Responsable: Arthur Chatton, Mireille Schnitzer

Room/Salle: ED 2018A

Date: Tuesday June 4 / mardi 4 juin

Time/Heure: 10:20-11:50

Abstract/Résumé

[10:20-10:50]

Mohammad Ehsanul Karim (The University of British Columbia) **Lucy Mosquera** (University of British Columbia) **Md Belal Hossain** (University of British Columbia)

Properties of inverse probability of adherence weighted estimator of the per-protocol effect for sustained treatment strategies under different data-generating mechanisms and adherence patterns

Propriétés de l'estimateur de pondération par probabilité inverse d'adhésion de l'effet per-protocole pour les stratégies de traitement soutenu dans le cadre de différents mécanismes de génération de données et de modèles d'adhésion

Inverse Probability Weighted per-protocol (IPW-PP) estimators are gaining traction for analyzing medication adherence in pragmatic trial data, but their behavior under various data-generating mechanisms (DGMs) needs more study. We examined their performance in pragmatic randomized controlled trials, comparing them to naive and baseline-adjusted estimators across DGMs that mimic real-world trials with two ongoing treatment strategies. These DGMs considered the influence of baseline variables on future factors, adherence patterns between trial arms, and the sporadic recording of adherence and confounders. Adjusting for baseline confounders typically yields unbiased results, but the IPW-PP estimator may be better when essential variables are unmeasured. High non-adherence can adversely affect IPW-PP estimates, especially with confounding influenced by past adherence. We applied these estimators to a case study from the Lipid Research Clinics Trial, considering non-adherence.

Les estimateurs de pondération par probabilité inverse d'adhésion de l'effet per-protocole (IPW-PP) sont de plus en plus utilisés pour l'analyse de l'adhésion aux médicaments dans les essais pragmatiques, mais leur comportement sous différents mécanismes de génération des données (DGMs) requiert plus d'études. Nous avons évalué leur performance dans des essais contrôlés randomisés pragmatiques, les comparant à des estimateurs naïfs et ajustés selon les variables de base, dans des DGMs simulant des essais réels avec deux stratégies de traitement. Ces DGMs considéraient l'influence des variables de base, les modèles d'adhésion entre les bras d'essai, et l'enregistrement sporadique de l'adhésion et des confondants. L'ajustement pour les confondants de base donne généralement des résultats non biaisés, mais l'IPW-PP peut être meilleur si des variables clés ne sont pas mesurées. Une forte non-adhésion peut nuire aux estimations IPW-PP, surtout si la confusion est influencée par l'adhésion passée. Nous avons utilisé ces estimateurs sur une étude de cas, en considérant la non-adhésion.

[10:50-11:20]

Arthur Chatton (Université de Montréal) **Robert W. Platt** (McGill University) **Michael Schomaker** (Ludwig-Maximilans-Universität München, Germany) **Miguel-Angel Luque-Fernandez** (University of Granada, Spain) **Mireille Schnitzer** (Université de Montréal)

A Diagnostic Tool for Sequential Positivity Violations in Longitudinal Causal Inference

Un outil de diagnostic pour les violations de positivité séquentielle dans l'inférence causale longitudinale

In longitudinal settings, the positivity assumption – necessary for considering an association as causal – gen-

Dans le contexte longitudinal, l'hypothèse de positivité – requise pour considérer une association causale – se généralise à ce que

eralizes to the requirement that all individuals must be able to follow all relevant exposure trajectories over all time points. Thus, the diagnosis of positivity violations is complicated because positivity must be checked across all time points and generally relies on a large number of models, unlikely all correctly specified. In this talk, we propose an extension of the Positivity Regression Trees (PoRT) algorithm to longitudinal settings by considering the whole treatment regimen to identify the covariate values or patterns yielding a lack of positivity (i.e., with null or nearly null probabilities of some exposure level). We demonstrate the potential of this approach by reanalyzing a study investigating the effect of HIV antiretroviral therapy among children in South Africa.

tous les individus puissent suivre toutes les trajectoires plausibles d'exposition au cours du temps. Diagnostiquer des violations de cette hypothèse est compliqué car la positivité doit être vérifiée à tous les points dans le temps et se fonde généralement sur de nombreux modèles, pouvant difficilement être tous correctement spécifiés. Pour cette présentation, nous proposons une extension de l'algorithme PoRT (Positivity Regression Trees) au contexte longitudinal en considérant l'entièreté de la stratégie d'exposition afin d'identifier les motifs de covariables impliquant un problème de positivité (c-à-d avec une probabilité nulle, ou presque, de recevoir une certaine valeur de l'exposition). Nous démontrons le potentiel de cette approche en réanalysant une étude évaluant l'effet d'un traitement anti-VIH chez des enfants en Afrique du Sud.

[11:20-11:50]

Eleanor M. Pullenayegum (Hospital for Sick Children)

Causal Inference With Longitudinal Data Subject to Irregular Assessment Times

Inférence causale avec des données longitudinales soumises à des temps d'évaluation irréguliers

Data collected in the context of usual care present a rich source of longitudinal data for research, but often require analyses that simultaneously enable causal inferences from observational data while handling irregular and informative assessment times. An inverse-weighting approach to this was recently proposed, and handles the case where the assessment times are at random (ie, conditionally independent of the outcome process given the observed history). In this talk I will show how multiple outputation can be used to extend the inverse-weighting approach to handle a special case of assessment not at random, where assessment and outcome processes are conditionally independent given past observed covariates and random effects. The methods will be illustrated through a study of the causal effect of wheezing on time spent playing outdoors among children aged 2–9 years.

Les données recueillies dans le contexte de soin régulier représentent une source riche en données longitudinales pour la recherche, mais demandent souvent des analyses qui permettent simultanément des inférences causales à partir de données d'observation tout en traitant les temps d'évaluation informatifs et irréguliers. Une approche par pondération inverse a récemment été proposée en guise de solution, et elle réussit à traiter les cas où les temps d'évaluation sont aléatoires (p. ex. lorsqu'ils sont conditionnellement indépendants du processus de résultat étant donné l'historique observé). Dans le cadre de cet exposé, je démontrerai comment la sortie multiple peut servir à élargir l'approche de pondération inverse afin de traiter un cas particulier d'évaluation non aléatoire, dans lequel l'évaluation et les processus de résultat sont conditionnellement indépendants étant donné les covariables observées antérieurement et les effets aléatoires. Les méthodes seront illustrées par l'entremise d'une étude de l'effet causal de la respiration bruyante sur le temps passé à jouer dehors parmi des enfants âgés et 2 à 9 ans.

Chair/Président: Po Yang

Room/Salle: C 3033

Date: Tuesday June 4 / mardi 4 juin

Time/Heure: 10:20-11:50

Abstract/Résumé

[10:20-10:35]

Mark Reesor (Wilfrid Laurier Universtiy) **Mark Drmac Walid Mnif Arie Zeldenrijk**

Incorporating Climate Risk into Portfolio Credit Risk Models via Distortion

Intégrer le risque climatique dans des modèles de risque de portefeuille de crédit par distorsion

Regulatory requirements are evolving towards mandating financial institutions to estimate and report their climate-related financial risks. Climate risks are medium- to long-term in nature and are important risk factors for credit portfolios. Threshold models for portfolio credit risk specify account level models with both systematic and idiosyncratic effects. These aggregate to generate the portfolio loss distribution from which risk metrics are calculated. Distortion provides a method for re-weighting a probability distribution. The amount of deformation depends on the choice of distortion function and its parameter. Here, we propose distortion as a way of incorporating climate risk into existing credit risk models. Some properties of the distorted credit risk models are derived and explored. The relation between our proposed models and existing climate-adjusted portfolio credit risk models will be discussed.

Les exigences réglementaires évoluent vers l'obligation des institutions financières à estimer et rapporter leurs risques financiers liés au climat. Les risques climatiques sont à moyen et long terme par nature et représentent des facteurs de risque important pour des portefeuilles de crédit. Les modèles de seuil pour le risque d'un portefeuille de crédit spécifient les modèles de niveau de compte avec des effets systématiques et idiosyncrasiques. Ceux-ci s'agrègent pour générer la distribution de perte du portefeuille à partir de laquelle on calcule les mesures de risque. La distorsion procure une méthode pour re-pondérer une distribution de probabilité. La quantité de déformation varie selon le choix de la fonction de distorsion et de son paramètre. Nous proposons ici la distorsion en guise de moyen pour intégrer le risque climatique dans les modèles de risque de crédit actuels. Nous dérivons et explorons certaines propriétés des modèles de risque de crédit. Nous parlerons du lien entre nos modèles proposés et les modèles actuels de risque de crédit de portefeuille ajusté au climat.

[10:35-10:50]

Manal Teto (University of Ottawa)

Dynamic Programming Approach to Price a Panel of American Options

Approche de programmation dynamique pour fixer le prix d'un panel d'options américaines

American-style options are Financial Derivatives that offer the flexibility of early exercise opportunities. This feature significantly influences their valuation methods and poses the difficulty of solving a dynamic optimization problem to determine the optimal exercise strategy. This research proposes an efficient method for pricing a panel of American-style options integrating a Stochastic Dynamic Programming approach (SDP) to overcome computational challenges associated with estimation of Greeks and efficient calibration to market data. One significant advantage of the SDP is that its so-

Les options américaines sont des dérivés financiers qui offrent la possibilité d'être exercé prématurément. Cette caractéristique influence grandement leurs méthodes d'évaluation et rend difficile la résolution de problème d'optimisation dynamique cherchant à déterminer la stratégie d'exercice optimale. Cette recherche propose une méthode efficace pour fixer le prix d'un panel d'options américaines intégrant une approche de programmation dynamique stochastique (SDP) afin de surmonter les défis de calcul associés à l'estimation des lettres grecques et la calibration efficace aux données du marché. L'un des avantages significatifs de la SDP est que ses solutions génèrent des approximations

lution yields numerical approximations to option prices and sensitivities across the entire state-space partition. Through leveraging the homogeneity property, we efficiently value diverse options with varying moneyness and maturity levels. The methodology is versatile, applicable to general models capturing interest rate term structures, showcasing the substantial benefits of our SDP method.

numériques pour les prix des options et des sensibilités à travers toute la partition spatiotemporelle. En tirant avantage de la propriété d'homogénéité, nous estimons de façon efficace diverses options ayant des montants investis et des niveaux de maturité variables. La méthodologie est polyvalente et applicable à des modèles généraux qui capturent des structures de terme de taux d'intérêt, démontrant ainsi la supériorité substantielle de notre méthode SDP.

[10:50-11:05]

Esam Mahdi (Carleton University)

New Mixed Portmanteau Tests for Time Series Models

Nouveaux tests portmanteau mixtes pour les modèles de séries temporelles

Omnibus portmanteau tests for contrasting adequacy of time series models are proposed. The test statistics are based on combining the autocorrelation function of the conditional residuals, the autocorrelation function of the conditional squared residuals, and the cross-correlation function between these residuals and their squares. The maximum likelihood estimator is used to derive the asymptotic distribution of the proposed test statistics under a general class of time series models, including ARMA, GARCH, and other nonlinear structures. An extensive Monte Carlo simulation study shows that the proposed tests successfully control the type I error probability and tend to have more power than other competitor tests in many scenarios. Two applications to a set of weekly stock returns for 92 companies from the S&P 500 demonstrate the practical use of the proposed tests.

Nous proposons des tests portmanteau omnibus pour contraster l'adéquation des modèles de séries temporelles. Les statistiques de test sont basées sur la combinaison de la fonction d'autocorrélation des résidus conditionnels, de la fonction d'autocorrélation des résidus conditionnels au carré et de la fonction de corrélation croisée entre ces résidus et leurs carrés. Nous utilisons l'estimateur du maximum de vraisemblance pour dériver la distribution asymptotique des statistiques de test proposées dans une classe générale de modèles de séries temporelles, dont ARMA, GARCH et d'autres structures non linéaires. Nous montrons par une étude de simulation Monte Carlo approfondie que les tests proposés contrôlent avec succès la probabilité d'erreur de type I et tendent à plus de puissance que d'autres tests concurrents dans de nombreux scénarios. Nous démontrons par deux applications à un ensemble de rendements boursiers hebdomadaires pour 92 sociétés du S&P 500 l'utilisation pratique des tests proposés.

[11:05-11:20]

Haixu Wang (University of Calgary) **Jiguo Cao** (Simon Fraser University)

Nonlinear Prediction of Functional Time Series

Prédiction non linéaire de séries temporelles fonctionnelles

We propose a nonlinear prediction (NOP) method for functional time series. Conventional methods for functional time series are mainly based on functional principal component analysis or functional regression models. These approaches rely on the stationary or linear assumption of the functional time series. The NOP method employs a nonlinear mapping for functional data that can be directly applied to multivariate functions without any preprocessing step. It is a one-step model that constructs feature space with forecast information, hence it provides a better ground for predicting future trajectories. Compared to the conventional methods, the NOP method avoids calculating covariance functions and enables online estimation and pre-

Nous proposons une méthode de prédiction non linéaire (NOP) pour les séries temporelles fonctionnelles. La méthode NOP utilise une cartographie non linéaire pour les données fonctionnelles qui peut être directement appliquée aux fonctions multivariées sans aucune étape de prétraitement. Il s'agit d'un modèle en une étape qui construit un espace de caractéristiques avec des informations prévisionnelles, ce qui permet de mieux prédire les trajectoires futures. Par rapport aux méthodes conventionnelles, la méthode NOP évite de calculer les fonctions de covariance et permet une estimation et une prédiction en ligne. Trois applications réelles démontrent les avantages de la méthode NOP pour prédire la qualité de l'air, le prix de l'électricité et le taux de mortalité.

diction. Three real applications demonstrate the advantages of the NOP method model in predicting air quality, electricity price, and mortality rate.

[11:20-11:35]

Maciej Augustyniak (Université de Montréal) **Alexandru Badescu** (University of Calgary) **Jean-François Bégin** (Simon Fraser University) **Sarath Kumar Jayaraman** (University of Calgary)

On the Relation Between Discrete and Continuous-Time Affine Option Pricing Models

À propos de la relation entre les modèles d'évaluation des options affines à temps continu et ceux à temps discret

This article studies the weak convergence of discrete-time affine stochastic volatility models driven by both Gaussian and non-Gaussian innovations. Our results generalize the existing diffusion limits for affine GARCH models and provide new insights on their relationship with continuous-time stochastic volatility models. Notably, we show that the canonical affine volatility models popularized in discrete and continuous time are not the analog of one another from the point of view of weak convergence.

Cet article étudie la convergence faible des modèles de volatilité stochastique affine à temps discret inspirés d'innovations gaussiennes et autres. Nos résultats généralisent les limites de diffusion actuelles pour affiner les modèles GARCH et offrent de nouvelles perspectives sur leur relation avec les modèles de volatilité stochastiques à temps continu. Notamment, nous démontrons que les modèles de volatilité affine canoniques popularisés en temps discret et continu ne sont pas analogues entre eux du point de vue de la convergence faible .

Chair/Président: Brian Franczak

Room/Salle: C 4036

Date: Tuesday June 4 / mardi 4 juin

Time/Heure: 10:20-11:50

Abstract/Résumé

[10:20-10:35]

Mina Aminghafari (University of Calgary) **Saeid Hoseinipour Hoseinipour** (Amirkabir University of Technology) **Adel Mohammadpour** (Amirkabir University of Technology) **Mohamed Nadif** (Université Paris Cité)

Exponential Family and Latent Block Model for Co-clustering

Famille exponentielle et modèle de blocs latents pour le Co-clustering

Co-clustering is a technique that simultaneously clusters rows and columns of a matrix, enabling the discovery of local patterns and structures. In recent years, co-clustering models have led to the development of numerous algorithms, demonstrating their superiority over traditional clustering methods. This advantage is particularly notable in the analysis of high-dimensional data, such as document-word matrices, which are the primary interest of this work. This work employs Latent Block Models (LBMs), which are robust statistical models known for their flexibility, efficiency, and parsimony. LBMs have been suggested for different types of data. This work is integrating many LBMs into a unified framework. The exponential family is a perfect candidate for this task, capable of embedding various distributions. In addition, sparse models can be derived to cope with the high dimensionality of document-word matrices, which makes co-clustering results easier to interpret.

Le Co-clustering est une technique qui regroupe simultanément les lignes et les colonnes d'une matrice, ce qui permet de découvrir des structures locaux. Au cours des dernières années, les modèles de Co-clustering ont démontré leur supériorité par rapport aux méthodes traditionnelles de clustering. Cet avantage est particulièrement considérable dans l'analyse de données de très grande dimension, telles que les données document-terme, que nous considérons dans ce travail. Ce travail utilise les Modèles de Blocs Latents (MBL), qui sont des modèles statistiques robustes connus par leur flexibilité, leur efficacité et leur parcimonie. Ce travail intègre de nombreux MBL dans un cadre unifié. La famille exponentielle est un candidat parfait à ce propos, capable d'intégrer diverses distributions. De plus, des modèles clairsemés (sparse) peuvent être utilisés pour faire face à la très grande dimension des matrices document-terme, ce qui rend les résultats de Co-clustering plus faciles à interpréter.

[10:35-10:50]

Julie Carreau (Polytechnique Montreal)

A Spatially Adaptive Multi-resolution Generative Algorithm: Application to Simulating Flood Wave Propagation

Algorithme génératif multi-résolutions spatialement adaptatif : application à la simulation de la propagation des ondes de crue

We propose a statistical model suitable for large spatio-temporal data sets exhibiting complex patterns such as simulated by physics-based hydraulic models over high resolution 2D meshes. As long computation times limit their applicability, statistical models that may emulate them quickly are developed. Our model relies on an extension of multiresolution analysis for spatio-temporal

Nous proposons un modèle statistique adapté aux grands ensembles de données spatio-temporelles présentant des schémas complexes tels que ceux simulés par les modèles physiques hydrauliques sur des mailles 2D à haute résolution. Comme les temps de calcul longs limitent leur applicabilité, nous développons des modèles statistiques qui peuvent les émuler rapidement. Notre modèle repose sur une extension de l'analyse multi-résolutions

data exploiting the idea that dominant spatial features remain present through time. An interpretable non-parametric representation can be derived from the lifting scheme by combining a smoothed version of the data with details. A generative algorithm is built by introducing the information provided by a low resolution model, whose computation times are orders of magnitude smaller, yielding a downscaling model. Our model is applied to a dam break experiment using a synthetic urban configuration and to a field-scale test case simulating the propagation of a dike break flood wave into a Sacramento urban area.

pour les données spatio-temporelles, en exploitant l'idée que les caractéristiques spatiales dominantes restent présentes dans le temps. Une représentation non paramétrique interprétable peut être dérivée du schéma de lifting en combinant une version lissée des données avec des détails. Un algorithme génératif est construit en introduisant les informations fournies par un modèle à basse résolution, dont les temps de calcul sont inférieurs de plusieurs ordres de grandeur, ce qui donne un modèle de réduction d'échelle. Notre modèle est appliqué à une expérience de rupture de digue utilisant une configuration urbaine synthétique et à un cas test à l'échelle du terrain simulant la propagation d'une onde de crue de rupture de digue dans une zone urbaine de Sacramento.

[10:50-11:05]

Samuel Morrissette (University of Manitoba) **Saman Muthukumarana** (University of Manitoba) **Maxime Turgeon** (University of Manitoba)

Parsimonious Dirichlet Process Mixture Models for Clustering with Dissimilarities

Modèles de mélange de processus de Dirichlet parcimonieux pour le regroupement avec dissimilarités

Model-based clustering is a popular clustering technique used to discover interesting structures and patterns in data. However, when only distances between observations are available, model-based clustering cannot be used. Oh and Raftery (2007) proposed a method to overcome this obstacle by first estimating the object configuration of the distances through multi-dimensional scaling and then fitting a Bayesian mixture model. Morrissette et al. (2024) expanded on this method by proposing parsimonious models to accurately cluster distance data in high dimensions. However, the Bayesian fitting process can be computationally demanding when the number of candidate models is large. As a result, we propose a non-parametric Bayesian method for clustering distance data using the Dirichlet Process. This method greatly reduces the number of candidate models and, as a result, increases computational efficiency. We provide a software implementation in R and demonstrate the method's capability through simulations and a real-world application.

Le regroupement basé sur un modèle est une technique de regroupement populaire utilisée pour découvrir des structures et des modèles intéressants dans les données. Toutefois, lorsque seules les distances entre les observations sont disponibles, il n'est pas possible d'utiliser ce type de regroupement. Oh et Raftery (2007) ont proposé une méthode pour surmonter cet obstacle en estimant d'abord la configuration de l'objet des distances par une mise à l'échelle multidimensionnelle et en ajustant ensuite un modèle de mélange bayésien. Morrissette et al. (2024) ont développé cette méthode en proposant des modèles parcimonieux pour regrouper avec précision les données de distance en grande dimension. Cependant, le processus d'ajustement bayésien peut être très exigeant en termes de calcul lorsque le nombre de modèles candidats est élevé. C'est pourquoi nous proposons une méthode bayésienne non paramétrique pour le regroupement des données de distance qui utilise le processus de Dirichlet. Cette méthode réduit considérablement le nombre de modèles candidats et, par conséquent, augmente l'efficacité des calculs. Nous fournissons une implémentation logicielle en R et démontrons la capacité de la méthode à travers des simulations et une application réelle.

[11:05-11:20]

Yi-Shu Lin (CHES, Sickkids) **Linke Li** (The Hospital for Sick Children) **Anna Heath** (The Hospital for Sick Children) **James O'Mahony** (University College Dublin)

Using Machine Learning Algorithms to Identify Relevant Strategies for Simulation in Cost-Effectiveness Analysis of Screening

Utilisation d'algorithmes d'apprentissage machine (ML) pour identifier des stratégies adéquates de simulation dans l'analyse de la rentabilité (CEA) du dépistage

The omission of relevant strategies is a recognized problem in the cost-effectiveness analyses (CEA) of screening, which prevents policymakers from best allocating

L'omission de stratégies adéquates dans les analyses de la rentabilité du dépistage est un problème connu qui empêche les décideurs de faire la meilleure répartition possible des ressources

healthcare resources. Computational expense is often the reason for such an omission as simulating all the possible screening strategies is not practical. We developed a method using machine learning (ML) to efficiently identify cost-effective screening strategies and to determine how much data are required to build these ML models. We used an existing CEA of cervical cancer screening to generate the net monetary benefit (NMB) of screening strategies. We employed Extreme Gradient Boosting to build the relationship between NMB and variables used to define strategies. The accuracy was low when relying solely on the predicted NMB of our model but the accuracy improved when more relevant strategies were chosen. Thus, ML models can identify optimal screening strategies at a low computational cost.

en soins de santé. La dépense computationnelle est souvent la raison d'une telle omission, car la simulation de toutes les stratégies possibles de dépistage n'est pas pratique. Nous avons développé une méthode faisant appel à l'apprentissage machine pour identifier efficacement des stratégies de dépistage économiques et déterminer la quantité nécessaire de données pour construire ces modèles ML. Nous avons utilisé une analyse CEA existante sur le dépistage du cancer du col de l'utérus pour générer le bénéfice monétaire net (NMB) des stratégies de dépistage. Nous avons employé l'augmentation extrême du gradient (EGB) pour établir la relation entre le NMB et les variables utilisées pour définir les stratégies. L'exactitude obtenue était faible en nous basant uniquement sur le NMB prédit de notre modèle, mais elle s'est améliorée lorsque des stratégies plus adéquates ont été choisies. Par conséquent, les modèles ML peuvent identifier des stratégies optimales de dépistage à faible coût de computation.

[11:20-11:35]

Devan G. Becker (Wilfrid Laurier University)

Defining SARS-CoV-2 Lineages with Temporally Consistent Mutation Clusters in Wastewater Samples

Définition des lignées SARS-CoV-2 avec des grappes de mutations cohérentes dans le temps dans les échantillons d'eaux usées

SARS-CoV-2 lineages are defined according to placement in a phylogenetic tree, but approximated by a list of mutations based on sequences collected from clinical sampling. Wastewater lineage abundance is generally estimated under the assumption that the mutation frequency is approximately equal to the sum of the abundances of the lineages to which it belongs. By leveraging many samples collected over time, it is possible to estimate the abundance of lineages as well as the definitions of those lineages. This is accomplished by a novel k-means-like algorithm, where the lineage abundance is estimated according to a time-varying coefficient model and then mutations are assigned to lineages are assigned according to distance of the observed mutation frequencies to the estimated lineage abundance. An important aspect of lineage assignment is that a mutation can be assigned to more than one lineage, in which case their abundances are additive to estimate the frequency.

Les lignées du SARS-CoV-2 sont définies en fonction de leur position dans un arbre phylogénétique, mais approximées par une liste de mutations basée sur des séquences collectées à partir d'échantillonnages cliniques. L'abondance des lignées dans les eaux usées est généralement estimée en partant de l'hypothèse que la fréquence des mutations est approximativement égale à la somme des abondances des lignées auxquelles elle appartient. En tirant parti de nombreux échantillons collectés au fil du temps, il est possible d'estimer l'abondance des lignées ainsi que les définitions de ces lignées. Pour ce faire, on utilise un nouvel algorithme de type k-moyenne, dans lequel l'abondance des lignées est estimée selon un modèle de coefficient variable dans le temps, puis les mutations sont attribuées aux lignées en fonction de la distance entre les fréquences de mutation observées et l'abondance estimée des lignées. Un aspect important de l'attribution des lignées est qu'une mutation peut être attribuée à plus d'une lignée, auquel cas leurs abondances s'additionnent pour estimer la fréquence.

[11:35-11:50]

Elif Fidan Acar (University of Manitoba) **Martin Lysy** (University of Waterloo)

Automated Statistical Methods for High-Throughput Phenotyping Experiments

Méthodes statistiques automatisées pour expériences de phénotypage à haut débit

Many health applications produce ever-increasing quantities of biological data. As such applications often rely on automated pipelines for data analysis, an impor-

Dans le domaine de la santé, de nombreuses applications produisent des quantités toujours croissantes de données biologiques. Comme ces applications s'appuient souvent sur des pipelines au-

tant statistical challenge is to evaluate and refine these pipelines as more and more data are acquired. This challenge is exemplified by the high-throughput phenotyping experiments conducted by the International Mouse Phenotyping Consortium, where multiple phenotype measurements are obtained for a small set of gene-edited mice and a large set of controls acquired continually over time. In order to increase the power of detecting the gene effect, model selection is a fundamental component of the automated pipeline. However, the effect of post-selection inference in this setting is not well understood. In this talk, I present ongoing work on evaluating and improving the IMPC statistical pipeline along this line of inquiry.

tomatisés pour l'analyse des données, un défi statistique important consiste à évaluer et à affiner ces pipelines au fur et à mesure de la production d'un nombre croissant de données. Ce défi est illustré par les expériences de phénotypage à haut débit menées par l'International Mouse Phenotyping Consortium (IMPC), où de multiples mesures de phénotypes sont obtenues pour un petit ensemble de souris génétiquement modifiées et un grand ensemble de contrôles acquis continuellement au fil du temps. Afin d'augmenter la puissance de détection de l'effet du gène, la sélection du modèle est un élément fondamental du pipeline automatisé. Cependant, l'effet de l'inférence post-sélection dans ce contexte n'est pas bien compris. Dans cet exposé, je présenterai les travaux en cours sur l'évaluation et l'amélioration du pipeline statistique de l'IMPC dans cette optique.

Chair/Président: Bruno N. Rémillard

Organizer/Responsable: Bruno N. Rémillard

Room/Salle: A 1043

Date: Tuesday June 4 / mardi 4 juin

Time/Heure: 10:20-11:50

Abstract/Résumé

[10:20-10:50]

Bouchra Nasri (Université de Montréal)

Tests of Serial Dependence for Multivariate Time Series with Arbitrary Distributions

Tests de dépendance sérielle pour des séries chronologiques multivariées de distribution arbitraire

One studies tests of serial independence using a fixed number of consecutive observations from a stationary time series, first in the univariate case, and then in the multivariate case, where even high-dimensional vectors can be used. The common distribution function is not assumed to be continuous, and the test statistics are based on the multilinear copula process. A bootstrap procedure based on multipliers is also proposed and shown to be valid. Tests based on Spearman's rho and Kendall's tau statistics are also considered, extending the results known for the case of continuous distributions. Contiguity results are obtained for some specific models and sufficient conditions for consistency of test statistics are stated, as well as a data-driven procedure to select . Also, numerical experiments are performed to assess the finite sample level and power of the proposed tests. A case study using a time series of Arctic sea ice extent images is used to illustrate the usefulness of the proposed methodologies. The R package MixedIndTests (Nasri et al., 2022) includes all the methodologies proposed in this article.

On étudie les tests d'indépendance sérielle en utilisant un nombre fixe d'observations consécutives d'une série temporelle stationnaire, d'abord dans le cas univarié, puis dans le cas multivarié, où l'on peut même utiliser des vecteurs à grande dimension. La fonction de distribution commune n'est pas supposée continue et les statistiques de test sont basées sur le processus de copule multilinéaire. Une procédure bootstrap basée sur les multiplicateurs est également proposée et s'avère valide. Les tests basés sur les statistiques rho de Spearman et tau de Kendall sont également pris en compte, étendant les résultats connus pour le cas des distributions continues. Des résultats de contiguïté sont obtenus pour certains modèles spécifiques et des conditions suffisantes pour la cohérence des statistiques de test sont énoncées, ainsi qu'une procédure basée sur les données pour sélectionner . Des expériences numériques sont également réalisées pour évaluer le niveau d'échantillon fini et la puissance des tests proposés. Une étude de cas utilisant une série temporelle d'images de l'étendue de la glace de mer dans l'Arctique est utilisée pour illustrer l'utilité des méthodologies proposées. Le package R MixedIndTests (Nasri et al., 2022) comprend toutes les méthodologies proposées dans cet article.

[10:50-11:20]

Mohamedou Ould Haye (Carleton University) **Anne Philippe** (Nantes University)

Inference for Discrete Randomized Linear Processes

Inférence pour processus linéaires aléatoires discrets

Irregularly observed time series occur in many fields such as astronomy, finance, environmental, and biomedical sciences. Statistical tools available to handle unevenly time series are essentially developed for short range dependence. In this paper we investigate large

On rencontre des séries temporelles observées de manière irrégulière dans de nombreux domaines tels que l'astronomie, la finance, l'environnement et les sciences biomédicales. Les outils statistiques permettant de traiter de telles séries sont essentiellement développés pour la dépendance à court terme. Dans cet article,

samples properties of such processes in a broader context of dependence by considering them as a regular time series process observed at random time points using a renewal process. We establish a sharp difference in the asymptotic behaviour of the self normalized sample mean of the observed process depending on the renewal process. If the later has finite moment (such as Poisson process), then the limiting distribution is a normal one while if it has a heavy tail distribution then the limit is a so-called Variance Mixing Normal (VMN) one and we give full description of this VMN as the product of two independent random variables, one is normal and the other is a function of a stable variable.

nous étudions les propriétés des grands échantillons de ces processus dans un contexte plus large de dépendance en les considérant comme des séries temporelles régulières observées à des moments aléatoires, avec un processus de renouvellement. Nous établissons une différence nette dans le comportement asymptotique de la moyenne d'échantillon auto-normalisée du processus observé, en fonction du processus de renouvellement. Si ce dernier a un moment fini (comme le processus de Poisson), la distribution limite est une distribution normale, tandis que s'il a une distribution à queue lourde, la limite est ce que l'on appelle une Variance Mixing Normal (VMN). Nous donnons une description complète de cette VMN en tant que produit de deux variables aléatoires indépendantes, dont l'une est normale et l'autre est une fonction d'une variable stable.

[11:20-11:50]

Masoud M. Nasari (Bank of Canada) **Mohamedou Ould Haye** (Carleton University)

A New Inferential Framework

Un nouveau cadre d'inférence

Central limit theorems for dependent data are reliant on structural models, such as various types of time series models like short and long memory linear processes, subordinated Gaussian, GARCH, etc. Different assumptions about the data and their auto-dependence structures can lead to different limiting distributions, such as normal or more complex ones like Rosenblatt and Hermit. However, verifying these structural assumptions empirically is challenging. Additionally, computing the percentiles of some of these limiting distributions is non-trivial. In this talk we introduce a new inferential framework that combines the information contained in a data set with a set of exogenous randomizing agents generated from a stable distribution, such as Cauchy. The framework is valid for a vast class of dependent data under very minimal conditions. It also easily accommodates replication of the randomization to ensures robustness

Le théorème central limite pour les données dépendantes se base sur des modèles structurels, comme plusieurs types de modèles de séries temporelles (processus linéaire à mémoire longue ou courte, gaussien subordonné, GARCH, etc.). Différentes hypothèses concernant les données et leurs structures d'autodépendance peuvent mener à différentes distributions de restriction ; des régulières ou des complexes comme celles de Rosenblatt et Hermit. Cependant, il est difficile de vérifier empiriquement ces hypothèses structurelles. De plus, le calcul des pourcentages est majeur pour certaines de ces distributions de restriction. Dans le cadre de cette présentation, nous présentons un nouveau cadre d'inférence qui combine l'information contenue dans un ensemble de données avec un ensemble d'agents randomisant exogènes généré à partir d'une distribution stable, comme celle de Cauchy. Ce cadre est valide pour un grand nombre de classes de données dépendantes selon des conditions très minimes. Il s'adapte aussi facilement à la réplication de la randomisation pour assurer une robustesse.

Statistical Modeling of Complex Medical Research Data
Modélisation statistique de données complexes de recherche médicale

Chair/Président: Neil A. Spencer

Organizer/Responsable: Tessema Astatkie

Room/Salle: A 1049

Date: Tuesday June 4 / mardi 4 juin

Time/Heure: 10:20-11:50

Abstract/Résumé

[10:20-10:50]

Demissie Alemayehu (Columbia University)

An Enriched Approach to Combining High-dimensional Genomic and Low-dimensional Phenotypic Data

Une approche enrichie pour combiner des données génomiques en grande dimension et des données phénotypiques en petite dimension

We propose an approach for combining and analyzing high-dimensional genomic and low-dimensional phenotypic data. The approach leverages a scheme of weights applied to the variables instead of observations and, hence, permits incorporation of the information provided by the low dimensional data source. It can also be incorporated into commonly used downstream techniques, such as random forest or penalized regression. Finally, simulated lupus studies involving genetic and clinical data are used to illustrate the overall idea and show that the proposed enriched penalized method can select significant genetic variables while keeping several important clinical variables in the final model.

Nous proposons une approche permettant de combiner et d'analyser des données génomiques en grande dimension et des données phénotypiques en petite dimension. Cette approche s'appuie sur un système de poids appliqués aux variables plutôt qu'aux observations et permet donc d'intégrer les informations fournies par la source de données à petite dimension. Elle peut également être incorporée dans des techniques d'aval couramment utilisées, telles que la forêt aléatoire ou la régression pénalisée. Enfin, nous utilisons des études simulées sur le lupus impliquant des données génétiques et cliniques pour illustrer l'idée générale et montrer que la méthode pénalisée enrichie proposée permet de sélectionner des variables génétiques significatives tout en conservant plusieurs variables cliniques importantes dans le modèle final.

[10:50-11:20]

Birol Emir (Columbia Univ)

A Flexible Alternative to Standard Modeling Techniques for Extrapolated Mean Survival Times Needed for Cost-Effectiveness Analyses

Solution de rechange polyvalente aux techniques de modélisation standard pour l'extrapolation des durées moyennes de survie nécessaires aux analyses coût-efficacité

This presentation explores the suitability of finite mixture models compared to common survival models in fitting heterogeneous data for estimating mean survival times crucial for cost-effectiveness analysis. Utilizing publicly available data, we digitized nonproprietary overall survival (OS) and progression-free survival (PFS) curves, fitting various regression models. Notably, the 3-Weibull and 2-Weibull mixture models demonstrated superior performance in PFS analysis, surpassing others by over 40 AIC points. For OS,

Dans cette présentation, nous examinons la capacité des modèles de mélanges finis à s'adapter aux données hétérogènes pour estimer les durées moyennes de survie, essentielles pour l'analyse coût-efficacité, puis nous comparons ces modèles aux modèles de survie classiques. Nous avons numérisé des courbes de survie globale et de survie sans progression non exclusives à partir de données publiques et nous avons ajusté différents modèles de régression. Il est à noter que les modèles de mélange de Weibull à deux et à trois composantes ont donné de meilleurs résultats dans l'analyse de la survie sans progression, surpassant

all models exhibited comparable AIC values, with 3-Weibull and 2-Weibull mixture models providing estimates closest to Kaplan-Meier mean. Censored PFS analysis yielded results akin to uncensored PFS. Extrapolating mean OS, all models produced estimates within 10% of Kaplan-Meier mean survival time. This underscores the flexibility and advantages of finite mixture models in handling heterogeneous data for survival time estimation in cost-effectiveness analysis.

les autres modèles de plus de 40 points selon le critère d'information d'Akaike. Pour la survie globale, tous les modèles présentaient des valeurs du critère d'information d'Akaike comparables, les modèles de mélange de Weibull à deux et à trois composantes fournissant les estimations les plus proches de la moyenne de Kaplan-Meier. L'analyse de la survie sans progression tronquée a donné des résultats similaires à ce type d'analyse non tronquée. Par extrapolation de la survie globale moyenne, tous les modèles ont produit des estimations se situant à moins de 10 % de la durée de survie moyenne de Kaplan-Meier. Ces résultats mettent en évidence la polyvalence et les avantages des modèles de mélange fini dans le traitement de données hétérogènes pour l'estimation de la durée de survie dans l'analyse coût-efficacité.

[11:20-11:50]

Javier Cabrera (Rutgers University)

Differential Projection Pursuit Methods and Their Applications to Differential Experiments

Méthodes de poursuite de la projection différentielle et ses applications aux expériences différentielles

The novel concept of differential projection pursuit, and its applications to the analysis of large datasets, are introduced. Projection pursuit has been applied for many years as a standard methodology for analyzing multivariate data. But in the applications of projection pursuit in the experimental setting, there are two issues of importance, which are the large number of observations and the differential nature of most experiments. We will introduce a new index, similar to the Natural Hermite index, that is suitable for measuring differences between 2 or more distributions. This index is also suitable for datasets with small and large numbers of observations, like in flow cytometry. We will also present a differential projection pursuit analysis of a large flow cytometry dataset with a treatment sample and a control sample. The algorithm will search for optimal projections and display clusters of treated cells in regions where there are few control cells and apply a rotation.

Le nouveau concept de poursuite de projection différentielle et ses applications à l'analyse de grands ensembles de données sont présentés. La poursuite de projection est appliquée depuis de nombreuses années comme méthodologie standard pour analyser des données multivariées. Mais dans les applications de la recherche par projection dans le cadre expérimental, deux problèmes importants se posent : le grand nombre d'observations et la nature différentielle de la plupart des expériences. Nous introduirons un nouvel indice, similaire à l'indice Natural Hermite, qui convient pour mesurer les différences entre 2 ou plusieurs distributions. Cet indice convient également aux ensembles de données comportant un petit et un grand nombre d'observations, comme en cytométrie en flux. Nous présenterons également une analyse de poursuite par projection différentielle d'un grand ensemble de données de cytométrie en flux avec un échantillon de traitement et un échantillon de contrôle. L'algorithme recherchera des projections optimales, affichera des groupes de cellules traitées dans des régions où il y a peu de cellules témoins et appliquera une rotation.

Chair/Président: Rob Deardon

Organizer/Responsable: Rob Deardon

Room/Salle: C 2045

Date: Tuesday June 4 / mardi 4 juin

Time/Heure: 10:20-11:50

Abstract/Résumé

[10:20-10:42]

Jee Yeon (Joanne) Kim (The Ohio State University) **Andrew B. Lawson** (Medical University of South Carolina)

A Novel Bayesian Spatio-temporal Surveillance Metric to Predict Emerging Infectious Disease High-risk Clusters

Nouvelle mesure bayésienne de surveillance spatio-temporelle pour prédire les nouveaux foyers de contagion à risque élevé

Identification of high-risk disease clusters has been one of the top goals for infectious disease public health surveillance. The proposed metric consists of three components: the area's own risk profile, temporal risk trend, and spatial neighborhood influence. We also introduce a weighting scheme to balance these three components, which accommodates the characteristics of the infectious disease outbreak, and spatial disease trends. Thorough simulation studies were conducted to identify the optimal weighting scheme and evaluate the performance of the proposed cluster prediction surveillance metric. Results indicate that the area's own risk and the neighborhood influence play an important role to make a highly sensitive metric, and the risk trend term is important for the specificity and the accuracy of prediction. The proposed cluster prediction metric was applied to the COVID-19 case data of South Carolina from March 12th, 2020, and the subsequent 30 weeks of the data.

La surveillance de la santé publique des maladies infectieuses vise principalement à identifier les foyers de contagion à risque élevé. La mesure proposée comprend trois éléments : le profil de risque propre à la zone, la tendance temporelle du risque et l'influence spatiale du voisinage. Pour équilibrer ces trois composantes, nous introduisons également un système de pondération qui tient compte des caractéristiques du foyer de contagion et des tendances spatiales de la maladie. Nous avons réalisé des études de simulation approfondies afin d'identifier le schéma de pondération optimal et d'évaluer les résultats de la mesure de surveillance proposée pour la prédiction des grappes. Les résultats indiquent que le risque de la zone et l'influence du voisinage jouent un rôle important dans l'élaboration d'une mesure très sensible, et que le terme de tendance du risque est important pour la spécificité et la précision de la prédiction. La mesure de prédiction des grappes proposée est appliquée aux données de cas de COVID-19 de la Caroline du Sud depuis le 12 mars 2020, ainsi qu'aux données des 30 semaines suivantes.

[10:42-11:05]

Madeline Ward (University of Calgary) **Rob Deardon** (University of Calgary) **Lorna Deeth** (University of Guelph) **Caitlin Ward** (University of Minnesota)

Accounting for Behavioural Changes in Epidemic Models

Prendre en compte les changements comportementaux dans les modèles épidémiques

People will often respond to changes in epidemic trajectories by adjusting their behaviours - either by taking additional protective measures when case counts are high, or relaxing those behaviours when case counts are low. Additionally, the degree to which people change their behaviours may depend on the amount of time passed in the epidemic, or on whether the cases are

Les gens vont souvent réagir aux changements de trajectoires d'une épidémie en ajustant leur comportement, soit en adoptant une mesure de protection supplémentaire quand les risques sont élevés, ou en réduisant les mesures de sécurité lorsque le risque est faible. De plus, l'envergure du changement de comportement peut varier en fonction du temps passé dans une épidémie, ou de si les cas d'infection sont en croissance ou en décroissance. Nous

currently increasing or decreasing. We will present a framework for modelling the effect of these dynamic behavioural changes in epidemic models fitted within a Bayesian framework. The application of these dynamic behavioural change models will be illustrated by comparing the behaviour changes across different locations throughout the COVID-19 pandemic.

présenterons un cadre pour modéliser l'effet de ces changements comportementaux dynamiques dans des modèles épidémiques ajustés dans un cadre bayésien. L'application de ces modèles de changement comportemental dynamique sera illustrée en comparant les changements comportementaux à différents emplacements durant la pandémie de la COVID-19.

[11:05-11:27]

Joanna Elizabeth Mills Flemming (Dalhousie University)

Exploring Encounter Processes: From Cell-Cell Interactions to Interspecies Dynamics

Exploration des processus de rencontre : des interactions entre cellules aux dynamiques inter-espèces

This talk centres on events unfolding in two dimensions, delving into formal definitions of encounters and interactions. The significance of these definitions to epidemiological and population dynamics models is explored. State-of-the-art modelling techniques for processes related to encounters are examined. Carefully selected case studies are used to demonstrate fitting such models to real data, explore statistical tools for inference and prediction, and inspire future research endeavours.

Cet exposé se concentre sur les événements qui se déroulent en deux dimensions et approfondit les définitions formelles des rencontres et des interactions. Nous explorons l'importance de ces définitions pour les modèles épidémiologiques et de dynamique des populations et examinons les techniques de modélisation les plus récentes pour les processus liés aux rencontres. Nous nous appuyons sur des études de cas soigneusement sélectionnées pour démontrer l'adaptation de ces modèles à des données réelles, explorer les outils statistiques pour l'inférence et la prédiction, et inspirer les futurs efforts de recherche.

[11:27-11:50]

Laura L.E. Cowen (University of Victoria)

From Ecology to Epidemiology

De l'écologie à l'épidémiologie

Ecological researchers are used to answering the question "how many" for conservation or management purposes. This question might be equally important in epidemiological and public health applications and we can make use of ecological models, modified for their particular application. I will discuss how we can enumerate COVID-19 cases, the homeless population of Victoria, and people who use injection drugs. For public health authorities, knowing "how many" is a starting point for addressing many public health concerns.

Les chercheurs en écologie répondent souvent de la question "combien de" à des fins de conservation ou de gestion. Cette question est également importante dans les applications épidémiologiques et de santé publique et on peut donc développer des modèles écologiques modifiés pour ces applications particulières. Je discuterai de la manière dont nous pouvons énumérer les cas de COVID-19, la population des sans-abri de Victoria et les personnes qui consomment des drogues injectables. Pour les autorités de santé publique, savoir "combien de" est le point de départ pour faire face aux problèmes de santé publique.

Chair/Président: Tolu Sajobi

Room/Salle: C 3053

Date: Tuesday June 4 / mardi 4 juin

Time/Heure: 10:20-11:50

Abstract/Résumé

[10:20-10:35]

Yixuan Li (McGill) **Ariane Marelli** (McGill Adult Unit for Congenital Heart Disease (MAUDE Unit), McGill University of Health Centre) **Yi Yang** (McGill) **Yue Li** (McGill)

MixEHR-SurG: a Joint Proportional Hazard and Guided Topic Model for Inferring Mortality-Associated Topics from Electronic Health Records

MixEHR-SurG : un modèle conjoint à risques proportionnels et à sujets prédéfinis pour inférer des sujets liés à la mortalité à partir de dossiers médicaux électroniques

To improve survival analysis using EHR data, we aim to develop a supervised topic model called MixEHR-SurG to simultaneously integrate heterogeneous EHR data and model survival hazard. Our technical contributions are three-folds: Integrating EHR topic inference with Cox-PH model; Inferring patient-specific topic hyperparameters using the PheCode concepts such that each topic can be identified with exactly one PheCode-associated phenotype; Multi-modal survival topic inference. We evaluated MixEHR-SurG using a simulated dataset and two real-world EHR datasets. MixEHR-SurG achieved a superior dynamic AUROC for mortality prediction, with a mean AUROC score of 0.89 in the simulation dataset and a mean AUROC of 0.645 on the CHD dataset. MixEHR-SurG associates severe cardiac conditions with high mortality risk among the CHD patients after the first heart failure hospitalization and critical brain injuries with increased mortality among the MIMIC-III patients after their ICU discharge.

Afin d'améliorer l'analyse de la survie à l'aide de données de dossiers médicaux électroniques (EHR), nous visons à développer un modèle à sujet supervisé appelé MixEHR-SurG pour intégrer des données EHR hétérogènes et modéliser le risque proportionnel de façon simultanée. Nos apports techniques comportent trois volets : l'intégration d'une inférence de sujets à partir des EHR avec le modèle des risques proportionnels de Cox ; l'inférence d'hyperparamètres de sujets spécifiques au patient en utilisant les phecodes de façon à ce que chaque sujet puisse être identifié exactement avec un phénotype associé à un phecode ; une inférence multimodale de sujets en survie. Nous avons évalué le MixEHR-SurG en utilisant un ensemble de données simulées et deux ensembles de données EHR du monde réel. Le modèle MixEHR-SurG a obtenu une aire sous la courbe ROC (AUROC) dynamique supérieure pour la prédiction de la mortalité, avec un score AUROC moyen de 0,89 pour l'ensemble de données simulées et de 0,645 pour l'ensemble de données sur la maladie cardiaque congénitale (CHD). Le MixEHR-SurG associe les problèmes cardiaques graves à un risque élevé de mortalité chez les patients CHD après la première hospitalisation pour insuffisance cardiaque et des lésions cérébrales critiques et une mortalité accrue parmi les patients dans la base de données MIMIC-III après leur congé de l'unité des soins intensifs.

[10:35-10:50]

Laura Bumbulis (University of Waterloo) **Richard J. Cook** (University of Waterloo)

Testing Process Reliability under a Limit of Detection: Issues of Robustness and Efficiency

Test de fiabilité de processus selon une limite de détection : problèmes de robustesse et d'efficacité

In biotechnology processes are said to be reliable if they produce samples satisfying regulatory benchmarks.

En biotechnologie, les processus sont considérés comme fiables s'ils produisent des échantillons satisfaisant les références réglementaires.

Through laboratory studies it may be necessary to show, for example, that levels of an analyte rarely (e.g. in less than 5% of samples) exceed a tolerance threshold. This can be challenging when measurement systems feature a lower limit of detection rendering some observations left-censored. In this talk we discuss the implications of detection limits for statistical inference in reliability studies, including their impact on large and finite sample properties of parameter estimates; power of tests for reliability and goodness of fit; and sensitivity of results to model misspecification. To balance efficiency and power with robustness, we investigate the use of smoothing and more flexible highly parameterized models (e.g. piecewise-constant hazard-based models). We conclude with some recommendations on the design of future reliability studies.

Par l'entremise d'étude en laboratoire, il serait nécessaire de montrer, entre autres, que les niveaux d'un analyte vont rarement (moins de 5 % des échantillons) dépasser un seuil de tolérance. Cela peut poser problème lorsque les systèmes de mesure possèdent une limite de détection inférieure, ce qui rend certaines observations censurées à gauche. Dans le cadre de cette présentation, nous aborderons les effets des limites de détection de l'inférence statistique dans les études de fiabilité, y compris leur impact sur les propriétés asymptotiques et non asymptotiques des estimateurs des paramètres, la puissance de tests pour la fiabilité et l'adéquation, et la sensibilité des résultats relatifs aux erreurs de spécification du modèle. Afin d'équilibrer l'efficacité et la puissance avec la robustesse, nous enquêtons sur l'utilisation du lissage et de modèles flexibles grandement paramétrés (p. ex. les modèles de risque constants par morceaux). Nous concluons avec quelques recommandations concernant la conception d'études de fiabilité à venir.

[10:50-11:05]

Xianwei Li (University of Waterloo) **Richard J. Cook** (University of Waterloo) **Liqun Diao** (University of Waterloo)

Prediction for Illness-death Processes under Intermittent Observation

Prévision pour un processus maladie-décès en cas d'observation intermittente

Illness-death models are commonly used to study chronic diseases characterized by multiple stages, where subjects are at a non-negligible risk of death. Examples of such diseases include cancer, HIV, and Alzheimer's disease. When the disease status can only be determined by periodic assessments, the exact entry times to states are unknown. We aim to use the multistate data under intermittent observation along with high-dimension covariates, to jointly predict disease progression and death at a particular time horizon. We formulate an illness-death model and a penalized likelihood in settings where the disease processes are under intermittent observation and death is subject to right censoring. An innovative expectation-maximization (EM) algorithm is then developed which can flexibly incorporate different penalty functions and allows one to exploit existing packages. The method will be illustrated in the context of a biomedical study to jointly predict a non-fatal event and death.

Les modèles maladie-décès sont couramment utilisés pour étudier les maladies chroniques caractérisées par plusieurs stades, où les sujets courent un risque non négligeable de mourir. Le cancer, le VIH et la maladie d'Alzheimer en sont des exemples. Lorsque l'état de la maladie ne peut être déterminé que par des évaluations périodiques, les temps d'entrée exacts dans chaque état sont inconnus. Notre objectif est d'utiliser les données multi-états tirées de l'observation intermittente, ainsi que des covariables de grande dimension, pour prédire conjointement la progression de la maladie et le décès à un horizon temporel donné. Nous formulons un modèle maladie-décès et une vraisemblance pénalisée dans des contextes où les processus pathologiques font l'objet d'une observation intermittente et où le décès est soumis à une censure à droite. Nous développons ensuite un algorithme innovant de maximisation de l'espérance, qui permet d'intégrer de manière flexible différentes fonctions de pénalité et d'exploiter les librairies existantes. Nous illustrons la méthode dans le contexte d'une étude biomédicale visant à prédire conjointement un événement non fatal et le décès.

[11:05-11:20]

Wenling Zhang (University of Waterloo) **Cecilia A. Cotton** (University of Waterloo) **Lan Wen** (University of Waterloo)

Targeted Maximum Likelihood and Other Robust Estimators for Recurrent Causal Events

Estimation ciblée du maximum de vraisemblance et autres estimateurs robustes pour événements causaux récurrents

In clinical studies, understanding the causal impacts of treatment on recurrent disease episodes like nonfatal

Dans les études cliniques, il est vital de bien comprendre l'impact causal du traitement sur les épisodes récurrents de la mala-

strokes or heart attacks is vital for improving patients' quality of life. Compared to other existing methodologies, targeted maximum likelihood estimation (TMLE) offers distinct advantages, such as double robustness, fewer data assumptions, and the flexibility to incorporate multiple algorithms into model fitting. This presentation explores various robust methods, including TMLE, to estimate the average causal effect on recurrent event outcomes with censoring. The theoretical insights are validated through simulations. To demonstrate our method, we examine the causal effect of intensive versus standard blood pressure lowering therapy on acute kidney injury recurrences, with data from the Systolic Blood Pressure Intervention Trial (SPRINT).

die, tels que les accidents vasculaires cérébraux non mortels ou les crises cardiaques, pour améliorer la qualité de vie des patients. Comparée à d'autres méthodologies existantes, l'estimation ciblée du maximum de vraisemblance (TMLE) offre des avantages distincts, tels qu'une double robustesse, moins d'hypothèses sur les données et la possibilité d'incorporer divers algorithmes dans l'ajustement du modèle. Cette présentation explore diverses méthodes robustes, y compris la TMLE, pour estimer l'effet causal moyen sur les résultats des événements récurrents avec censure. Nous validons ces idées théoriques par des simulations. Pour démontrer notre méthode, nous examinons l'effet causal d'une thérapie intensive par rapport à une thérapie standard d'abaissement de la pression artérielle sur les récidives de lésions rénales aiguës, avec des données provenant de l'essai d'intervention sur la pression artérielle systolique (SPRINT).

[11:20-11:35]

Connie Stewart (University of New Brunswick) **Tyler Rideout** (University of New Brunswick Saint John) **Matthew Stephenson** (Quantics)

Prey Selection for Fatty Acid Signature Analysis Using the Akaike Information Criterion

Sélection de proie pour l'analyse des signatures des acides gras en utilisant le critère d'information d'Akaike

Estimation of the diet compositions of marine predators through the method of fatty acid signature analysis allows valuable insights into the trophic structures of marine ecosystems by comparing the fatty acid signatures of predators with a library of potential prey. These prey libraries consist of the fatty acid signatures of individuals grouped by species, with the number of species limited by the number of dietary fatty acids under analysis. We propose a novel application of the Akaike information criterion to identify the correct set of species within a wider set of potential prey. The estimation of true zeroes through the removal of species from the prey library results in reduced variability and greater accuracy in the estimation of non-zero proportions. Outcomes from simulation studies as well as the analysis of real-life grey seal data will be used to explore this method's performance.

L'estimation des compositions des régimes alimentaires des prédateurs marins par la méthode de l'analyse des signatures des acides gras donne les indications précieuses sur les structures trophiques des écosystèmes en comparant les signatures des acides gras des prédateurs avec une catalogue de proies potentielles. Ces catalogues sont constitués de signatures des acides gras de proies individuelles regroupées par espèces, avec le nombre d'espèces limité par le nombre d'acides gras diététiques sous analyse. Nous proposons une application nouvelle du critère d'information d'Akaike pour identifier le bon ensemble d'espèces d'un ensemble plus large de proies potentielles. L'estimation des vrais zéros en supprimant des espèces du catalogue de proies se traduit par une variabilité réduite et une plus grande exactitude dans l'estimation des proportions non nuls. Les résultats des études de simulation et l'analyse des données réelles sur les phoques gris seront utilisé pour explorer cette méthode.

Chair/Président: Kevin McGregor

Organizer/Responsable: Kevin McGregor

Room/Salle: A 2071

Date: Tuesday June 4 / mardi 4 juin

Time/Heure: 10:20-11:50

Abstract/Résumé

[10:20-10:50]

Alexander Shestopaloff (Memorial University of Newfoundland) **Mihai Cucuringu** (University of Oxford) **Yichi Zhang** (University of Oxford) **Stefan Zohren** (University of Oxford)

Robust Detection of Lead-Lag Relationships in Lagged Multi-Factor Models

Détection robuste de relations lead-lag dans les modèles multifacteurs décalés

In multivariate time series systems, key insights can be obtained by discovering lead-lag relationships inherent in the data, which refer to the dependence between two time series shifted in time relative to one another, and which can be leveraged for the purposes of control, forecasting or clustering. We develop a clustering-driven methodology for robust detection of lead-lag relationships in lagged multi-factor models. Within our framework, the envisioned pipeline takes as input a set of time series, and creates an enlarged universe of subsequence time series via a sliding window approach. This is followed by an application of clustering techniques to group the subsequence time series. Lead-lag estimates across clusters are robustly aggregated to enhance the identification of lead-lag relationships in the original universe. We demonstrate that our method is not only able to robustly detect lead-lag relationships in financial markets, but in an environmental data set.

Dans les systèmes de séries chronologiques multivariées, des informations clés peuvent être obtenues en découvrant les relations lead-lag inhérentes aux données, qui font référence à la dépendance entre deux séries chronologiques décalées dans le temps l'une par rapport à l'autre, et qui peuvent être exploitées à des fins de contrôle, prévision ou regroupement. Nous développons une méthodologie basée sur le regroupement pour une détection robuste des relations lead-lag dans les modèles multifactoriels décalés. Dans notre cadre, le pipeline envisagé prend en entrée un ensemble de séries temporelles et crée un univers élargi de séries temporelles de sous-séquences, via une approche de fenêtre glissante. Ceci est ensuite suivi par l'application de diverses techniques de regroupement de séries temporelles de sous-séquences. Les estimations lead-lag entre les regroupements sont agrégées de manière robuste pour améliorer l'identification de celles-ci dans l'univers d'origine. Nous démontrons que notre méthode est non seulement capable de détecter de manière robuste les relations lead-lag sur les marchés financiers, mais également lorsqu'elle est appliquée à un ensemble de données environnementales.

[10:50-11:20]

James H. McVittie (University of Regina)

Survival Analysis Methodologies for Wildlife Studies

Méthodologie d'analyse de survie dans les études sur la faune

In a wildlife study, where a group of animals are monitored from some prespecified age to the event of death, the observed data consists of a set of failure/censoring times as well as other measured covariates. Based on the ages of the animals upon entry into the study, the

Dans une étude sur la faune, lorsqu'un groupe d'animaux est surveillé à partir d'un certain âge prédéterminé jusqu'au moment de la mort, les données observées consistent en un ensemble de temps de défaillance/de censure, de même que d'autres covariables mesurées. En fonction de l'âge des animaux au mo-

observed cohort comprises partially observed durations that are possibly unbiased or left-truncated. In this talk, we will present some methodological strategies for simultaneously including multiple types of partially observed duration data into a single modelling procedure. We will discuss some recent applications of these procedures in modelling African lion and baboon mortality as well as some open statistical problems for future research.

ment de leur intégration à l'étude, la cohorte observée comprend des durées partiellement observées qui sont possiblement non biaisées ou tronquées à gauche. Nous présentons des stratégies méthodologiques pour l'inclusion simultanée de types multiples de données de durée partiellement observées en une seule procédure de modélisation. Nous discutons également de certaines applications récentes de cette procédure dans la modélisation de la mortalité du lion et du babouin d'Afrique ainsi que de certains problèmes statistiques ouverts pour la recherche subséquente.

On Some Multivariate Distributions and Recent Advances in Robust Inference
Des distributions multivariées et avancées récentes en inférence robuste

Chair/Président: Mai Ghannam

Organizer/Responsable: Mai Ghannam

Room/Salle: A 2065

Date: Tuesday June 4 / mardi 4 juin

Time/Heure: 10:20-11:50

Abstract/Résumé

[10:20-10:50]

Sévérien Nkurunziza (University of Windsor) **Mai Ghannam** (University of Ottawa)

Some Recent Identities in Tensor Elliptically Contoured Distributions and Their Applications

Quelques identités récentes pour des distributions elliptiques tensorielles et leurs applications

In this presentation, we consider inference problem concerning the matrix or tensor parameter of a random matrix (or a random tensor) that follows some elliptically contoured distributions under possible uncertain constraints. Under such statistical model, we discuss some recent properties of multivariate distributions which are useful in shrinkage estimation methods. More precisely, we also present some identities as well as sharp inequalities which are useful in deriving the risk dominance of some matrix or tensor estimators.

Dans cet exposé, nous considérons un problème d'inférence concernant le paramètre matriciel (ou tensoriel) d'une matrice aléatoire (ou d'un tenseur aléatoire) qui suit une distribution elliptique. En particulier, on considère le scénario où le paramètre en question est susceptible de satisfaire certaines restrictions. Eu égard à ce modèle statistique, nous présentons certaines propriétés récentes des distributions multivariées qui sont utiles dans les méthodes d'estimation à rétrécissement. Plus précisément, nous présentons des identités remarquables ainsi que des inégalités utiles pour établir le risque et l'optimalité de certains estimateurs matriciels ou tensoriels.

[10:50-11:20]

Serge B. Provost (The University of Western Ontario)

Identities Stemming from Matrix-variate Density Functions

Identités découlant de fonctions de densité des variables matricielles

Certain identities derived from the normalizing constants of selected matrix-variate densities will be utilized to determine the moments and moment-generating functions of said distributions. Subsequently, numerous particular cases of interest will be identified. It will also be explained that a lesser-known Jacobian of matrix transformations can singularly simplify the derivation of the Wishart density function. [Invited talk; the session organized by Mai Ghannam is entitled 'On some multivariate distributions and recent advances in robust inference']

Nous utiliserons certaines identités dérivées des constantes de normalisation des densités de variables matricielles sélectionnées pour déterminer les moments et les fonctions génératrices de moments de ces distributions. Par la suite, nous identifierons de nombreux cas particuliers d'intérêt. Nous expliquerons également qu'un jacobien des transformations matricielles moins connu peut singulièrement simplifier la dérivation de la fonction de densité de Wishart.

[11:20-11:50]

Éric P. Marchand (Université de Sherbrooke)

The search for efficient predictive densities for multivariate data

On Some Multivariate Distributions and Recent Advances in Robust Inference
Des distributions multivariées et avancées récentes en inférence robuste

La recherche de densités prédictives efficientes pour des données multivariées

This talk will address the estimation of predictive densities, Bayesian or otherwise, and their efficiency as measured by frequentist risk. For Kullback-Leibler, $\alpha-$divergence and integrated L_1 losses, we review several recent findings that bring into play improvements by scale expansion, as well as duality relationships with point estimation and point prediction problems. A range of models are studied and include multivariate normal with both known and unknown covariance structure, scale mixture of normals, mean mixture of normals including skew-normal distributions, Gamma, as well as models with restrictions on the parameter space.

Cet exposé abordera l'estimation de densités prédictives, bayésiennes ou non, et leur efficacité telle que mesurée par le risque fréquentiste. Pour les coûts Kullback-Leibler, $\alpha-$divergence et L_1 intégrée, nous passons en revue des résultats récents mettant en évidence des améliorations sur des densités-cible, dont la meilleure densité équivariante, en procédant par expansion d'échelle et des relations de dualité avec des problèmes d'estimation et de prédiction ponctuelle. Plusieurs modèles sont étudiés dont des lois normales multivariées avec une structure de covariance connue et inconnue, des mélanges de lois normales, des lois Gamma et des modèles avec des restrictions sur l'espace des paramètres.

Chair/Président: Joseph Beyene

Room/Salle: ED 2018B

Date: Tuesday June 4 / mardi 4 juin

Time/Heure: 10:20-11:50

Abstract/Résumé

[10:20-10:35]

Thierry Chekouo (University of Minnesota) **Samuel Babatunde Samuel Babatunde** (University of Calgary) **Samuel Babatunde** (University of Calgary)

A Bayesian Variable Selection for Semicontinuous Response data: Application to cardiovascular disease

Un modèle de sélection bayésien pour des données de réponses semi-continues : application aux maladies cardiovasculaires

Bayesian variable selection methods have been extensively developed for regression analyses, but have not been adequately developed for semicontinuous outcomes. We propose a Bayesian two-part model for variable selection where one part of the model estimates the probability of zero responses while the other part estimates positive responses. We link the two parts through two variable selection indicators by using either the same indicators between the two models or imposing a Markov Random Field prior to the two indicators that favors the selection of common features. We assessed the performance and importance of the proposed methods on both simulated and coronary artery disease data.

Les méthodes bayésiennes de sélection de variables ont été largement développées pour les analyses de régression, mais n'ont pas été suffisamment développées pour les réponses semi-continues. Nous proposons un modèle bayésien en deux parties pour la sélection de variables, dans lequel une partie du modèle estime la probabilité de zero réponses tandis que l'autre partie estime les réponses positives. Nous lions les deux parties via deux indicateurs de sélection de variables en utilisant soit les mêmes indicateurs entre les deux modèles, soit en imposant un champ aléatoire de Markov a priori sur les deux indicateurs qui favorise la sélection de caractéristiques communes. Nous avons évalué les performances et l'importance des méthodes proposées sur des données simulées et des données de maladies coronariennes.

[10:35-10:50]

Larry Dong (University of Toronto Dalla Lana School of Public Health) **Eleanor M. Pullenayegum** (The Hospital for Sick Children) **Olli Saarela** (University of Toronto)

On Bayesian Joint Modelling with Irregularly Observed Data to Estimate Optimal Treatment Regimes

Au sujet des modèles bayésiens conjoints avec des données irrégulières pour l'estimation des régimes de traitement optimaux

Optimal dynamic treatment regimes (DTR) consist of cascading decision rules aimed at determining the sequence of treatments tailored to patients, maximizing a long-term outcome. While conventional DTR estimation uses longitudinal data, there is little work on devising methods that use irregularly observed data to infer optimal DTRs. In this work, we first extend the target trial framework - a paradigm to estimate specified statistical estimands under hypothetical scenarios using observational data - to the DTR context; this extension

Les régimes de traitement dynamiques optimaux (DTR) sont des règles de décision visant à déterminer la séquence de traitements adaptés aux patients, maximisant leurs résultats à long terme. Alors que l'estimation conventionnelle des DTRs utilise des données longitudinales, il y a peu de travail sur les méthodes utilisant des données observées de manière irrégulière dans le contexte de l'estimation des DTRs optimaux. Dans le cadre de notre projet, nous abordons d'abord les essais ciblés, un paradigme permettant d'estimer des paramètres causals spécifiés dans des scénarios hypothétiques tout utilisant des données observa-

allows treatment regimes to be defined with intervenable visit times. We then propose 1) an adapted version of G-computation marginalizing over random effects and 2) a Bayesian joint model to handle correlated random effects between the outcome, visit and treatment processes. We show via simulation studies that failure to account for the observational treatment and visit processes produces bias in the estimation of regime rewards.

tionelles, au contexte des DTRs ; cette extension nous permet de définir des régimes de traitement avec des temps de visite intervenables. Nous proposons ensuite 1) une version adaptée du ≪G-computation≫ qui se base sur la loi marginale des effets aléatoires et 2) un modèle bayésien conjoint pour les effets aléatoires corrélés entre les processus de résultat, de visite et de traitement. Nous montrons à travers des études de simulation que, en ignorant la corrélation due aux processus de traitement et de visite, nous obtenons un biais dans l'estimation des valeurs de régime.

[10:50-11:05]

Yushu Zou (University of Toronto Dalla Lana School of Public Health) **Aya A. Mitani** (Dalla Lana School of Public Health, University of Toronto) **Olli Saarela** (Dalla Lana School of Public Health, University of Toronto) **Kuan Liu** (Dalla Lana School of Public Health, University of Toronto; Institute of Health Policy, Management, and Evaluation, University of Toronto)

A Bayesian Sensitivity Analysis Approach for Unmeasured Confounding in Longitudinal Data

Une approche d'analyse de sensibilité bayésienne pour les variables confondantes non-mesurées pour des données longitudinales

Causal estimation relies on the untestable assumption of no unmeasured confounding to ensure the causal parameter of interest is identifiable. Sensitivity analysis (SA) quantifies unmeasured confounding's impact on causal estimates which has been proposed in the literature. Among SA methods, the latent confounder approach is favored for its intuitive interpretation via the use of sensitivity parameters to specify the relationship between the observed and unobserved variables. However, this approach has not been adapted to longitudinal data. We developed a parametric Bayesian Sensitivity Analysis approach to quantify the impact of time-varying unmeasured confounding with time-varying treatment. We conducted simulation studies to examine the performance of our approach and applied it to a multi-centre pediatric disease registry.

L'estimation causale repose sur l'hypothèse non vérifiable de l'absence de variables confondantes non mesurées pour garantir que le paramètre causal d'intérêt soit identifiable. L'analyse de sensibilité (SA) quantifie l'impact de la confusion non mesurée sur les estimations causales. Parmi les méthodes de SA, l'approche de la variable confondante latente est souvent préférée pour son interprétation intuitive à travers l'utilisation de paramètres de sensibilité pour spécifier la relation entre les variables observées et non observées. Cependant, cette approche n'a pas encore été adaptée pour l'analyse causale avec des données longitudinales. Nous avons développé une approche bayésienne de SA pour quantifier l'impact de la confusion non mesurée variant dans le temps avec des traitements variant également dans le temps. Nous avons mené des études de simulation pour examiner la performance de notre approche et l'avons appliquée à des données provenent d'un registre de maladies pédiatriques multi-centres.

[11:05-11:20]

Wen Teng (The Hospital for Sick Children) **Niall Ferguson** (University Health Network) **Ewan Goligher** (University Health Network) **Anna Heath** (The Hospital for Sick Children)

Bayesian Joint Modeling for Longitudinal Magnitude Data with Informative Dropout: an Application to Critical Care Data

Modélisation bayésienne conjointe pour données longitudinales d'amplitude avec abandon informatif : une application aux données de soins intensifs

Biomedical studies often focus on the magnitudes of data, especially when algebraic signs are irrelevant or lost. Analyzing such data, particularly in repeated measures studies, requires models with random effects to account for subject-level heterogeneity and variability and enhance parameter estimation precision. However, existing regression methods lack incorporation of random

Les études biomédicales se concentrent souvent sur l'ampleur des données, en particulier lorsque les signes algébriques ne sont pas pertinents ou sont perdus. L'analyse de ces données, en particulier dans les études à mesures répétées, nécessite des modèles à effets aléatoires pour tenir compte de l'hétérogénéité et de la variabilité au niveau du sujet et améliorer la précision de l'estimation des paramètres. Cependant, les méthodes de régression existantes

effects for magnitude outcomes. Our work fills this gap with Bayesian regression models tailored for magnitude data, integrating random effects. Additionally, we extend the method to handle multiple causes of informative dropout using joint modeling strategies as dropout is commonly encountered in repeated measures studies. Numerical simulations, mirroring our motivating study, validate our approach, showing accurate estimation and bias mitigation for missing data. We use these models to study data from motivating study, investigating how sex impacts magnitude changes in diaphragm thickness for ICU patients.

n'intègrent pas d'effets aléatoires pour les résultats d'ampleur. Notre travail comble cette lacune avec des modèles de régression bayésienne adaptés aux données d'amplitude et qui intègrent des effets aléatoires. En outre, nous étendons la méthode pour traiter les causes multiples d'abandon informatif en utilisant des stratégies de modélisation conjointe, car l'abandon est fréquent dans les études à mesures répétées. Des simulations numériques, reflétant notre étude de motivation, valident notre approche, montrant une estimation précise et une atténuation du biais pour les données manquantes. Nous utilisons ces modèles pour étudier les données de l'étude de motivation, à savoir l'impact du sexe sur les changements d'épaisseur du diaphragme chez les patients des unités de soins intensifs.

[11:20-11:35]

Michelle F. Miranda (University of Victoria)

A CANDECOMP/PARAFAC basis for fast Bayesian Estimation of Multi-Subject fMRI

Une base CANDECOMP/PARAFAC pour une estimation bayésienne rapide d'une IRMf à multisujet

Task-evoked fMRI studies, such as the Human Connectome Project (HCP), are a powerful tool for exploring how brain activity is influenced by cognitive tasks like memory retention, decision-making, and language processing. A fast Bayesian function-on-scalar model is proposed for estimating population-level activation maps linked to a working memory task. The model is based on the CANDECOMP/PARAFAC tensor decomposition of coefficient maps obtained for each subject. This decomposition effectively yields a tensor basis capable of extracting both common and subject-specific features from the coefficient maps. These subject-level features, in turn, are modeled as a function of covariates of interest. The dimensionality reduction achieved with the tensor basis allows for a fast MCMC estimation of population-level activation maps. This model is applied to one hundred unrelated subjects from the HCP dataset, yielding significant insights into brain signatures associated with working memory.

Les études d'IRMf provoquées par la tâche, telles que l'≪ Human Connectome Project ≫ (HCP), sont un outil puissant pour étudier comment l'activité cérébrale est influencée par des tâches cognitives telles que la rétention de la mémoire, la prise de décision et le traitement du langage. Un modèle à fonctions scalaires bayésien rapide est proposé pour estimer les cartes d'activation au niveau de la population reliée à un test de mémorisation immédiate. Le modèle est basé sur la décomposition du tenseur CANDECOMP/PARAFAC des cartes de coefficient obtenues pour chaque sujet. Cette décomposition génère efficacement une base de tenseur capable d'extraire les caractéristiques communes et spécifiques au sujet à partir des cartes de coefficient. Ces caractéristiques au niveau du sujet sont ensuite modélisées en une fonction de covariables pertinentes. La réduction de dimension obtenue grâce à la base de tenseur permet une estimation MCMC rapide de cartes d'activation au niveau de la population. Ce modèle est appliqué à cent sujets sans lien à partir des données HCP, et procure des perspectives significatives sur les signatures cérébrales associées à la mémoire opérationnelle.

[11:35-11:50]

Lara Maleyeff (McGill University) **Shirin Golchi** (McGill University) **Erica Moodie** (McGill University)

An Adaptive Enrichment Design using Bayesian Model Averaging for the Identification of Tailoring Variables

Plan d'enrichissement adaptatif utilisant la moyenne des modèles bayésiens pour l'identification des variables d'adaptation

As with many chronic conditions, selecting the optimal treatment to patients with rheumatoid arthritis is challenging. The current trial-and-error approach, in which patients cycle through one of the many treatment options available until remission, leads to months or years

Comme pour de nombreuses maladies chroniques, il est difficile de choisir le traitement optimal pour les patients atteints de polyarthrite rhumatoïde. L'approche actuelle par essais et erreurs, dans laquelle les patients passent par l'une des nombreuses options thérapeutiques disponibles jusqu'à la rémission, conduit à

of suboptimal disease control, considerable loss to patient well-being, and a burden on healthcare systems. Precision medicine, in which patient's characteristics inform treatment decisions, requires new approaches in clinical trial design and analysis. Existing designs generally focus on biomarkers which are categorized into pre-specified subgroups, such as sex, and utilize a single model specification. Motivated by a trial studying available treatments in rheumatoid arthritis, we create a novel adaptive enrichment design using splines and Bayesian model averaging to flexibly identify the region of a multi-dimensional biomarker space where treatment is effective, while also allowing for early stopping.

des mois ou des années de contrôle sous-optimal de la maladie, à une perte considérable de bien-être pour le patient et à un fardeau pour les systèmes de soins de santé. La médecine de précision, dans le cadre de laquelle les caractéristiques du patient influencent les décisions thérapeutiques, nécessite de nouvelles approches en matière de conception et d'analyse des essais cliniques. Les modèles actuels se concentrent souvent sur les biomarqueurs qui sont classés dans des sous-groupes prédéfinis (sexe) et utilisent une spécification de modèle unique. Motivés par une étude sur les traitements disponibles pour la polyarthrite rhumatoïde, nous créons un nouveau modèle d'enrichissement adaptatif utilisant des splines et une moyenne de modèle bayésienne pour facilement identifier la zone d'un espace multidimensionnel de biomarqueurs où le traitement est efficace tout en permettant l'arrêt précoce du traitement.

Chair/Président: Zeny Feng

Organizer/Responsable: Zeny Feng

Room/Salle: ED 2018A

Date: Tuesday June 4 / mardi 4 juin

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-14:00]

Jiahua Chen (The University of British Columbia)

Moment Estimator and the Optimal Minimax Convergence Rate

Estimateur de moment et taux de convergence optimal minimax

When a population is suspected to comprise several homogeneous subpopulations, each adequately modeled by a standard parametric distribution, the overall distribution can be described as a finite mixture. While finite mixture models are well-motivated and find broad applications, they present numerous technical challenges in developing valid and effective statistical inference procedures. A notable gap exists in our current understanding concerning the convergence rate for estimating the mixing distribution, particularly when the order of the finite mixture model is overspecified based on independent and identically distributed (iid) observations. The best attainable minimax rate for the subpopulation distribution with a single parameter, in cases where the order is overspecified by one, was initially believed to be $n^{-1/4}$ but has been revised to $n^{-1/6}$. In this presentation, we will elucidate some findings pertaining to the moment estimator and its minimax convergence rate.

Si l'on estime qu'une population est composée de plusieurs sous-populations homogènes, chacune étant modélisée de manière adéquate par une distribution paramétrique standard, la distribution globale peut être décrite comme un mélange fini. Bien que les modèles de mélanges finis soient bien motivés et trouvent de nombreuses applications, ils présentent de nombreux défis techniques dans l'élaboration de procédures d'inférence statistique valides et efficaces. En effet, il existe une lacune notable dans nos connaissances actuelles concernant le taux de convergence pour l'estimation de la distribution de mélange, notamment lorsque l'ordre du modèle de mélange fini est surspécifié selon des observations indépendantes et identiquement distribuées. Le meilleur taux minimax réalisable pour la distribution de sous-population avec un seul paramètre, dans les cas où l'ordre d'un paramètre est surspécifié, était initialement estimé à $n^{-1/4}$, mais a été révisé à $n^{-1/6}$. Dans cette présentation, nous expliquerons certains résultats relatifs à l'estimateur de moment et à son taux de convergence minimax.

[14:00-14:30]

Sanjeena Dang (Carleton University) **Andrea Payne** (Carleton University) **Anjali Silva** (University of Toronto) **Steven Rothstein** (University of Guelph) **Paul David McNicholas** (McMaster University)

A Parsimonious Family of Mixtures of Multivariate Poisson Log-Normal Factor Analyzers for Clustering Count Data

Une famille parcimonieuse de mélanges d'analyseurs de facteur log-normal de Poisson multivariés pour le regroupement de données de dénombrement

Multivariate count data are commonly encountered in bioinformatics. Although the Poisson distribution seems a natural fit for these count data, its multivariate extension is computationally expensive. Recently, mixtures of multivariate Poisson lognormal (MPLN) models

On rencontre fréquemment des données de dénombrement multivariées en bio-informatique. La distribution de Poisson peut sembler idéale pour ces données de dénombrement, mais son extension multivariée demande beaucoup de calcul. Récemment, des mélanges de modèles log-normal de Poisson multivariés

have been used to efficiently analyze these multivariate count measurements. In the MPLN model, the counts, conditional on the latent variable, are modelled using a Poisson distribution, and the latent variable comes from a multivariate Gaussian distribution. Due to this hierarchical structure, the MPLN model can account for over-dispersion as opposed to the traditional Poisson distribution and allows for correlation between the variables. The mixture of multivariate Poisson-log normal distributions for high dimensional data is extended by incorporating a factor analyzer structure in the latent space. A family of parsimonious mixtures of multivariate Poisson lognormal distributions are proposed by decomposing the covariance matrix and imposing constraints on these decompositions. The performance of the model is demonstrated using simulated and real datasets.

(MPLN) ont servi à analyser de façon efficace ces mesures de dénombrement multivariées. Dans le modèle MPLN, les dénombrements (conditionnels à la variable latente) sont modélisés au moyen d'une distribution de Poisson, tandis que la variable latente provient d'une distribution gaussienne multivariée. En raison de sa structure hiérarchique, le modèle MPLN peut tenir compte de la surdispersion contrairement à la distribution de Poisson traditionnelle et permet la corrélation entre les variables. Le mélange des distributions log-normal de Poisson multivariées pour des données de grande dimension est élargi en intégrant une structure d'analyseur de facteur dans l'espace latent. Nous proposons une famille de mélanges parcimonieux de distributions log-normal de Poisson multivariées en décomposant la matrice de covariance et en imposant des restrictions sur ces décompositions. La performance du modèle est démontrée au moyen d'ensembles de données réelles et simulées.

[14:30-15:00]

Pengfei Li (University of Waterloo) **Tao Yu** (National University of Singapore) **Jing Qin** (National Institutes of Health)

Maximum Binomial Likelihood for Multivariate Mixture Data

Vraisemblance binomiale maximale pour données de mélanges multivariées

Multivariate mixture data analysis presents numerous challenges and constitutes a vital area of interest in the fields of statistics and data science. In this talk, we focus on nonparametric estimation techniques for multivariate mixture data. Specifically, we assume a known number of subpopulations and propose a binomial likelihood method, along with an efficient numerical algorithm, to estimate the mixing proportions and cumulative distribution functions of these subpopulations without relying on parametric assumptions. Through extensive numerical experiments, we demonstrate three key advantages of our approach: (1) Our method eliminates the need for tuning parameters. (2) It does not require the assumption of continuous component density functions. (3) Our method consistently delivers stable performance. To illustrate the practical performance of our method, we include a real-data example.

L'analyse des mélanges de données multivariées présente de nombreux défis et constitue un domaine d'intérêt vital dans les domaines de la statistique et de la science des données. Dans cet exposé, nous nous concentrons sur les techniques d'estimation non paramétriques pour les mélanges de données multivariées. Plus précisément, nous supposons un nombre connu de sous-populations et proposons une méthode de vraisemblance binomiale, ainsi qu'un algorithme numérique efficace, pour estimer les proportions de mélange et les fonctions de distribution cumulative de ces sous-populations sans nous appuyer sur des hypothèses paramétriques. Grâce à des expériences numériques approfondies, nous démontrons trois avantages clés de notre approche : (1) Notre méthode élimine le besoin de paramètres de réglage. (2) Elle ne nécessite pas l'hypothèse de fonctions de densité de composants continues. (3) Elle offre des performances stables et constantes. Pour illustrer les performances pratiques de notre méthode, nous incluons un exemple de données réelles.

Chair/Président: Liqun Diao

Organizer/Responsable: Liqun Diao

Room/Salle: A 1045

Date: Tuesday June 4 / mardi 4 juin

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-14:00]

Joel A. Dubin (University of Waterloo) **Minzee Kim** (University of Waterloo) **Tatiana Krikella** (University of Waterloo)

Advances in Similarity-based Predictive Modeling Methods

Progrès dans les méthodes de modélisation prédictive par similarités

Earlier work has shown that similarity-based predictive models can improve upon predictive performance, as compared to using the entire training data to help build models, particular regarding model discrimination. My collaborators and I have some updated results to share, regarding similarity-based modeling for joint consideration of model calibration and discrimination, as well as for dynamic prediction models. Properties of our methods will be investigated in comprehensive simulation studies, and we will demonstrate the methods through separate analyses of a publicly available ICU database.

Des travaux précédents ont révélé que les modèles prédictifs par similarités peuvent améliorer les résultats prédictifs par rapport à l'utilisation de l'ensemble des données d'apprentissage pour créer les modèles, en particulier en ce qui concerne la discrimination des modèles. Mes collaborateurs et moi-même avons de nouveaux résultats à présenter concernant la modélisation par similarités pour la prise en compte conjointe de l'étalonnage et de la discrimination des modèles, ainsi que pour les modèles de prédiction dynamiques. Nous analyserons les propriétés de nos méthodes dans des études de simulation complètes, puis nous ferons une démonstration des méthodes à l'aide d'analyses distinctes d'une base de données publique d'unités de soins intensifs.

[14:00-14:30]

Ameer Dharamshi (University of Washington) **Anna Neufeld** (Fred Hutchinson Cancer Center) **Keshav Motwani** (University of Washington) **Lucy L. Gao** (University of British Columbia) **Daniela Witten** (University of Washington) **Jacob Bien** (University of Southern California)

Data Thinning with Applications in the Health Sciences

Affinage des données avec application en sciences de la santé

We propose data thinning, a new approach for splitting an observation from a known distributional family with unknown parameter(s) into multiple independent pieces that together can be recombined to recover the original. This proposal is very general, it can be applied to a wide range of distributions including the Gaussian, Poisson, negative binomial, multinomial, Wishart, and many others both inside the exponential family, and beyond. The independent pieces generated by data thinning can be used for various data analysis tasks including model selection, model validation, and inference. For instance, cross-validation via data thinning provides an attractive

Nous proposons un affinage des données, une nouvelle approche pour le fractionnement d'une observation d'une famille de distribution, avec un ou des paramètres inconnus, en morceaux indépendants qui peuvent être tous recombinés pour retrouver l'originale. Notre proposition est générale et peut s'appliquer à une vaste série de lois, y compris la loi gaussienne, de Poisson, binomiale négative, multinomiale, Wishart, et plusieurs autres, à la fois dans et au-delà de la famille exponentielle. Les morceaux indépendants générés par l'affinage des données peuvent servir à diverses analyses de données dont la sélection de modèle, la validation de modèle et l'inférence. Par exemple, la validation croisée au moyen de l'affinage des données fournit une solution

alternative to the traditional cross-validation via sample splitting, especially in settings such as unsupervised learning in which the latter is not applicable. The utility of data thinning will be illustrated with applications across the health sciences.

de rechange intéressante à la validation croisée traditionnelle par fractionnement de l'échantillon, en particulier dans un contexte comme celui de l'apprentissage non supervisé où cette dernière n'est pas applicable. Nous illustrons l'utilité de l'affinage des données avec des applications en sciences de la santé.

[14:30-15:00]

Jon Steingrimsson (Brown University)

Generalizability of Study Results

Généralisabilité des résultats d'études

Study results are often used and/or interpreted in the context of a target population that differs from the study population from which the data used to develop the model comes from (e.g., a different health-care system or a different geographic region). Generalizability, or the lack thereof, of study results is a well-known challenge and often limits the interpretability and usefulness of study findings including randomized trials used for the approval of treatments and prediction models used for risk stratification. In this talk, we will discuss methods for generalizing measures of model performance to a target population when outcome and covariate information are available from the study data and covariate but no outcome data are available on a sample from the target population.

Les résultats d'études sont souvent utilisés ou interprétés dans un contexte de population cible différente de la population étudiée de laquelle proviennent les données utilisées pour développer le modèle (par exemple, un système de soins de santé différent ou une région géographique différente). La généralisabilité des résultats d'études, ou le manque de généralisabilité, est une difficulté bien connue qui souvent limite l'interprétabilité et l'utilité des résultats, y compris des essais randomisés servant à l'approbation des traitements et aux modèles prédictifs utilisés pour la stratification du risque. Nous discutons aujourd'hui des méthodes pour généraliser les mesures de la performance du modèle à une population cible, lorsque les données et covariables de l'étude permettent de disposer du résultat et de l'information sur les covariables, mais sans que soient disponibles les données des résultats d'un échantillonnage de la population cible.

Statistical Learning and Decision Making in Biostatistics
Apprentissage statistique et prise de décision en biostatistique

Chair/Président: Xikui Wang

Organizer/Responsable: Xikui Wang

Room/Salle: A 1049

Date: Tuesday June 4 / mardi 4 juin

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-14:00]

Yanqing Yi (Memorial University of Newfoundland)

Stochastic Modeling and Optimal Adaptive Design of Clinical Trials

Modélisation stochastique et conception adaptative optimale d'essais cliniques

Response adaptive designs use the information collected during a trial to modify the randomization probabilities in order to allocate more patients to the potential better treatment. The designs have ethical advantages over the traditional 50-50 randomization designs, but introduces a dependency in data, which may result in a statistical power loss. In this talk, we will discuss the trade-offs between the ethical gain and the loss of statistical power to explore the optimal design of adaptive clinical trials. We formulate the randomization process in an adaptive clinical trial as a stochastic sequential decision problem. When the information of previous treatment allocations and associated responses are summarized with sufficient statistics for unknown parameters, the decision process of treatment randomization becomes a Markov process. It is proven that the average reward under the policy identified from the span-contractor operator converges almost surely to the optimal value.

Les plans adaptatifs à la réponse utilisent les données collectées au cours d'un essai clinique pour modifier les probabilités de randomisation afin que davantage de patients reçoivent le meilleur traitement possible. Ces plans présentent des avantages éthiques par rapport aux plans de randomisation 50-50 traditionnels, mais ils introduisent de la dépendance dans les données, qui peut causer une perte de puissance statistique. Dans cette présentation, nous discutons des compromis entre le gain éthique et la perte de puissance statistique afin d'explorer la conception optimale des essais cliniques adaptatifs. Nous formulons le processus de randomisation dans un essai clinique adaptatif comme un problème de décision séquentielle stochastique. Lorsque les données relatives aux traitements précédents et aux réponses associées sont résumées à l'aide de statistiques suffisantes pour les paramètres inconnus, le processus de décision de la randomisation du traitement devient un processus de Markov. Il est prouvé que la récompense moyenne pour la stratégie identifiée à partir de l'opérateur « span-contractor » converge presque sûrement vers la valeur optimale.

[14:00-14:30]

Wenqing He (University of Western Ontario) **Grace Y. Yi** (University of Western Ontario) **Raymond Carroll** (Taxas A & M University)

Feature Screening with Large Scale and High Dimensional Survival Data

Sélection de caractéristiques pour données de survie à grande échelle et en grande dimension

Data with a huge size present great challenges in modeling, inferences, and computation. We will present a screening method for large-sized survival data, where the sample size is large and the dimension of covariates is of non-polynomial order of the sample size. We rigorously establish theoretical results and conduct numerical studies to assess the performance of the pro-

Les données de grande taille posent de grands défis en matière de modélisation, d'inférence et de calcul. Nous présenterons une méthode de sélection pour les données de survie de grande taille, lorsque la taille de l'échantillon est importante et que la dimension des covariables est d'un ordre non polynomial de la taille de l'échantillon. Nous établissons rigoureusement des résultats théoriques et menons des études numériques pour évaluer les per-

posed method. The method capitalizes on the connections among useful regression settings and offers a computationally efficient screening procedure.

formances de la méthode proposée. La méthode exploite les liens entre les paramètres de régression utiles et offre une procédure de sélection efficace sur le plan informatique.

[14:30-15:00]

You Liang (Toronto Metropolitan University) **Aleksandar Popovic** (Toronto Metropolitan University) **Na Yu** (Toronto Metropolitan University) **Xun Zhou** (St. Michael's Hospital) **Keanu Uchida** (St. Michael's Hospital) **Tomasz Tkaczyk** (Rice University) **Neeru Gupta** (St. Michael's Hospital; University of Toronto) **Yeni Yucel** (St. Michael's Hospital; University of Toronto)

A Graph-based Semantic Segmentation Algorithm for Hyperspectral Fluorescence Microscopy Imaging Data

Algorithme de segmentation sémantique basé sur les graphes pour données d'imagerie de microscopie à fluorescence hyperspectrale

The development of image processing methods and algorithms of hyperspectral fluorescence microscopy imaging (HFMI) has facilitated the detection of various fluorescent contrast sources in hyperspectral data cubes. The demand for efficient image processing techniques for diverse types of HFMI data across biomedical applications is evident, especially concerning semantic segmentation to label HFMI data. We propose a novel graph-based algorithm for the semantic segmentation of HFMI. First, superpixels are generated to remove the image noise and create small homogenous regions. Second, the normalized graph cuts algorithm is used to perform an initial segmentation of the image. Moreover, linear unmixing adds extra abundance information and the normalized cuts algorithm is applied again. This proposed segmentation algorithm is promising for enhancing the comprehension and diagnosis of eye diseases, including Spaceflight-Associated Neuro-ocular Syndrome and Amyotrophic lateral sclerosis.

Le développement de méthodes de traitement d'images et d'algorithmes d'imagerie hyperspectrale par microscopie à fluorescence (HFMI) a facilité la détection de diverses sources de contraste fluorescentes dans les cubes de données hyperspectrales. La demande de techniques efficaces de traitement d'images pour divers types de données HFMI dans les applications biomédicales est évidente, en particulier en ce qui concerne la segmentation sémantique permettant d'étiqueter les données HFMI. Nous proposons un nouvel algorithme basé sur les graphes pour la segmentation sémantique des HFMI. Tout d'abord, des superpixels sont générés qui suppriment le bruit de l'image et créent de petites régions homogènes. Ensuite, l'algorithme des coupes de graphes normalisé permet une segmentation initiale de l'image. En outre, le démélange linéaire ajoute des informations supplémentaires sur l'abondance et l'algorithme des coupes normalisé est à nouveau appliqué. L'algorithme de segmentation proposé est prometteur pour améliorer la compréhension et le diagnostic de maladies oculaires, notamment le syndrome neuro-oculaire associé aux vols spatiaux et la sclérose latérale amyotrophique.

New Development in Functional Data Analysis
Nouveaux développements en analyse des données fonctionnelles

Chair/Président: Jiguo Cao

Organizer/Responsable: Jiguo Cao

Room/Salle: A 2071

Date: Tuesday June 4 / mardi 4 juin

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-14:00]

Edward Gunning (University of Pennsylvania) **Giles Hooker** (University of Pennsylvania)

A New Perspective on Principal Differential Analysis

Nouvelle perspective de l'analyse différentielle principale

One of the unique features of functional data analysis (FDA) is the ability to use derivatives (i.e., rates of change) in modelling. Ramsay (1996) initially proposed to use derivatives in FDA in an approach called Principal Differential Analysis (PDA). In PDA, a linear ordinary differential equation (ODE) is estimated from functional data and its solutions used as basis functions to represent the data. We present a new methodological contribution, where we move beyond using PDA as a data representation tool, and formulate it as a generative statistical model. This perspective has two main consequences – 1) a more complete characterization of the sources of variability in PDA models, but 2) parameter estimates that can be severely biased. We propose an iterative bias correction to improve parameter estimates, and demonstrate our approach on simulated data from linear and non-linear differential equation models and real data from human movement biomechanics.

Une des caractéristiques uniques de l'analyse de données fonctionnelles (FDA) est la capacité d'utiliser des dérivés (par ex. : les taux de change) dans la modélisation. Ramsay (1996) a proposé au départ d'utiliser les dérivés dans la FDA dans une approche appelée analyse différentielle principale (PDA). Dans cette dernière approche, une équation différentielle ordinaire linéaire (ODE) est estimée à partir de données fonctionnelles et ses solutions sont utilisées comme fonctions de base pour représenter les données. Nous présentons une nouvelle contribution méthodologique, dans laquelle nous allons au-delà de l'utilisation d'une PDA comme outil de représentation des données et la formulons comme un modèle statistique génératif. Cette perspective a deux conséquences principales : 1) une caractérisation plus complète des sources de variabilité dans les modèles de PDA, mais 2) des estimations de paramètres qui peuvent être très biaisées. Nous proposons une correction itérative du biais pour améliorer les estimations des paramètres et illustrons notre approche à l'aide de données simulées tirées de modèles d'équations différentielles linéaires et non linéaires, ainsi que de données réelles en biomécanique des mouvements du corps humain.

[14:00-14:30]

Luo Xiao (North Carolina State University) **Ruonan Li** (North Carolina State University)

Latent Factor Model for Multivariate Functional Data

Modèle à facteurs latents pour les données fonctionnelles multivariées

For multivariate functional data, a class of functional latent factor model (FLFM) is proposed, extending the traditional latent factor model for multivariate data. The proposed model uses unobserved stochastic processes to induce the dependence among the different functions, and thus, for a large number of functions, may pro-

Pour les données fonctionnelles multivariées, nous proposons une classe de modèle à facteurs latents fonctionnel, qui étend le modèle de facteur latent traditionnel pour les données multivariées. Le modèle proposé utilise des processus stochastiques non observés pour créer une dépendance entre les différentes fonctions et peut donc, pour un grand nombre de fonctions, fournir

vide a more parsimonious and interpretable characterization of the otherwise complex dependencies between the functions. Sufficient conditions are provided to establish the identifiability of the proposed model. We shall use an application to electroencephalography data to illustrate the unsupervised FLFM. Then we apply a special case of FLFM, multivariate functional mixed model (MFMM), to jointly model multiple longitudinal biomarkers and time to event data for an AD study. Finally, we discuss a few future extensions.

une caractérisation plus parcimonieuse et plus facile à interpréter des dépendances autrement complexes qui existent entre les fonctions. Des conditions suffisantes sont prévues pour établir l'identifiabilité du modèle proposé. Nous présenterons une application aux données d'électroencéphalographie pour illustrer le modèle à facteurs latents fonctionnel non supervisé. Nous appliquerons ensuite un cas particulier de modèle à facteurs latents fonctionnel, le modèle à facteurs latents fonctionnel mixte multivarié pour modéliser conjointement plusieurs biomarqueurs longitudinaux et des données sur le temps écoulé jusqu'à l'événement dans le cadre d'une étude sur la maladie d'Alzheimer. Enfin, nous discuterons de quelques extensions futures.

[14:30-15:00]

Tianyu Guan (Brock University) **Shifan Jia** (Simon Fraser University) **Haolun Shi** (Simon Fraser University)

Semiparametric Function-on-function Regression Models

Modèles de régression fonction-sur-fonction semi-paramétriques

We propose a semiparametric function-on-function regression model that predicts a functional response by both a nonparametric dynamic effect of a functional predictor and a parametric concurrent effect of another functional predictor. The dynamic effect is characterized by taking an integral of a time-dependent two-dimensional smooth surface and the concurrent effect is modeled through a time-varying coefficient. The model combines the flexibility of nonparametric modeling with the interpretability of the parametric concurrent effect. We approximate the smooth surface using tensor product basis expansions, and for the time-varying coefficient, we employ B-spline expansions. The expansion parameters for each effect are estimated iteratively to account for the mutual dependencies between these two estimated effects. We establish the asymptotic properties of our estimator. The numerical performance of the proposed method is illustrated by simulation studies and two real data applications.

Nous proposons un modèle de régression fonction-sur-fonction semi-paramétrique qui prédit une réponse fonctionnelle à l'aide d'un effet dynamique non-paramétrique d'un prédicteur fonctionnel et d'un effet concomitant paramétrique d'un autre prédicteur fonctionnel. L'effet dynamique est caractérisé par l'intégrale d'une surface lisse bidimensionnelle dépendante du temps et l'effet concomitant est modélisé par un coefficient variable dans le temps. Le modèle combine la souplesse de la modélisation non-paramétrique avec l'interprétabilité de l'effet concomitant paramétrique. Nous approximons la surface lisse à l'aide d'expansions de base de produits tensoriels et, pour le coefficient variable dans le temps, nous utilisons des expansions B-spline. Les paramètres d'expansion pour chaque effet sont estimés itérativement pour tenir compte des dépendances mutuelles entre ces deux effets estimés. Nous établissons les propriétés asymptotiques de notre estimateur. Nous illustrons la performance numérique de la méthode proposée à l'aide d'études de simulation et de deux applications à des données réelles.

Recent Advances in Capital Structure Models and Contingent Capital
Avancées récentes en modèles de structure du capital et capital contingent

Chair/Président: Mark Reesor

Organizer/Responsable: Mark Reesor

Room/Salle: A 2065

Date: Tuesday June 4 / mardi 4 juin

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-14:00]

Francois Michel Boire (University of Ottawa) **Mark Reesor** (Wilfrid Laurier University) **Hatem Ben-Ameur** (HEC Montréal) **Pascal François** (HEC Montréal) **Lars Stentoft** (University of Western Ontario)

A Dynamic Structural Model for Contingent Convertible Debt

Un modèle dynamique structurel pour des obligations convertibles à coupon

This paper provides a simple framework to analyze the design of contingent capital for depository institutions. We consider a debt portfolio composed of straight debt and contingent convertible (coco) bonds. Coco bonds are automatically written down or converted into equity upon meeting predetermined capitalization ratio thresholds, thus providing capital loss absorption mechanisms. The quality of the design of coco bonds (parameterized by write-down/conversion factors and trigger levels) is gauged with default probabilities at various horizons, shareholder's incentive for risk-shifting, and the present value of bankruptcy costs.

Cet article présente un cadre analytique simple pour la conception d'obligations convertibles. Nous considérons le bilan d'une institution de dépôt financée par une obligation à coupon et une obligation conditionnelle convertible (coco). Pour absorber les pertes de capital, les cocos sont automatiquement décotées ou converties en fonds propres lorsque la valeur des actifs de la firme atteint certains seuils de capitalisation prédéterminés. Nous mesurons les effets des dispositions du contrat coco (paramétrée par les facteurs de décote/conversion et les seuils de déclenchement) sur la solvabilité de la firme, la propension des actionnaires au transfert de risque, ainsi que la valeur actualisée des coûts de faillite.

[14:00-14:30]

Di Meng (Wilfrid Laurier University) **Adam Metzler** (Wilfrid Laurier University) **Mark Reesor** (Wilfrid Laurier University)

Capital Structural Models and Contingent Convertible Securities

Modèles structurels de capitaux et titres convertibles contingents

We implement a methodology to calibrate capital structural models for financial firms that have issued contingent convertible securities (CoCo). Typical studies involving capital structural model calibration focus on non-financial firms as they have lower leverage and no contingent convertible securities. From a theoretical perspective, we find that jumps in the asset-value process are necessary to obtain a satisfactory fit to market data. In practice, contingent capital conversion triggers are discretionary, and there is considerable uncertainty around when regulators are likely to enforce conversion. The market-implied conversion triggers we obtain indicate that the market expects regulators to enforce con-

Nous implantons une méthodologie afin de calibrer des modèles structurels de capitaux pour les entreprises financières ayant émis des titres convertibles contingents (CoCo). Les études habituelles concernant la calibration de modèle structurel de capital se concentrent sur des entreprises non financières, car elles ont un levier financier faible et ne possèdent pas de CoCo. Selon une perspective théorique, on découvre que les sauts dans le processus de la valeur de l'actif sont nécessaires pour obtenir un ajustement adéquat aux données de marché. En pratique, les déclenchements de conversion de capital contingent sont discrétionnaires, et une incertitude considérable est présente lorsque les régulateurs sont susceptibles d'appliquer une conversion. Les déclenchements de conversion sous-entendus par le marché que nous obtenons in-

version while the issuing firm is a going concern, as opposed to a gone concern. This fact would presumably be of interest to potential CoCo investors.

diquent que le marché s'attend à ce que les régulateurs appliquent une conversion lorsque l'entreprise émettrice est une entreprise en exploitation, au contraire d'une entreprise en déclin. Cette information pourrait intéresser les investisseurs potentiels en CoCo.

[14:30-15:00]

Joe Campolieti (Wilfrid Laurier University)

Last Hitting Times, Excursions and Meanderings of Solvable Diffusions

Derniers temps de passage, excursions et méandres des diffusions solubles

We briefly present new theorems for computing the distribution of the last hitting (passage) time to any given level, as well as its joint distribution with the process value, and its extrema within any finite time horizon. The general formulae link the last and first hitting times and are valid for any scalar time-homogeneous diffusion. By employing spectral expansions, we derive newly explicit formulae for commonly solvable processes. Moreover, we extend the formulae to several families of multiparameter (nonlinear local volatility) solvable models. We give some numerical applications of these models by pricing new types of step options involving the last hitting time. We also develop new general formulae for distributions that are central to excursions of a diffusion straddling any two levels. Our formulae involve only univariate integral expressions that are readily implementable for any solvable diffusion. We demonstrate the applicability of our formulae. Moreover, we develop new formulae for the distribution of any solvable meandering diffusion. As part of the analysis, we derive new formulae for the joint distribution of last hitting time, the meander process value and its terminal value. We conclude by pointing out new possible implementations of the last passage time within some recently developed credit risk models.

Nous présentons brièvement de nouveaux théorèmes permettant de calculer la distribution du dernier temps de passage à un niveau donné, ainsi que sa distribution conjointe avec la valeur du processus et ses extrema dans un horizon temporel fini. Les formules générales relient le dernier et le premier temps de passage et sont valables pour toute diffusion scalaire homogène dans le temps. Nous obtenons de nouvelles formules explicites pour les processus généralement solubles grâce à l'utilisation d'expansions spectrales. De plus, nous étendons les formules à plusieurs familles de modèles multiparamétriques solubles (volatilité locale non linéaire). Nous présentons quelques applications numériques de ces modèles en évaluant de nouveaux types d'options par étapes utilisant le dernier temps de passage. Nous proposons également de nouvelles formules générales pour les distributions qui sont au cœur des excursions d'une diffusion se situant entre deux niveaux. Nous démontrons l'applicabilité de nos formules, qui ne nécessitent que des expressions intégrales à une variable qui peuvent être facilement mises en œuvre pour toute diffusion soluble. De plus, nous concevons de nouvelles formules pour la distribution de n'importe quelle diffusion méandrique soluble. Dans le cadre de l'analyse, nous obtenons de nouvelles formules pour la distribution conjointe du dernier temps de passage, de la valeur du processus de méandres et de sa valeur finale. En conclusion, nous présentons de nouvelles possibilités de mise en œuvre du temps de dernier passage dans certains modèles de risque de crédit récemment élaborés.

Bridging the Gap: Navigating Collaborative Research with Non-Statisticians (Panel)
Combler le fossé : la recherche collaborative avec des non-statisticiens (table ronde)

Chair/Président: Reza Ramezan

Organizer/Responsable: Reza Ramezan

Room/Salle: SN 2109

Date: Tuesday June 4 / mardi 4 juin

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-15:00]

Daniel J. McDonald (University of British Columbia) **Tolulope Sajobi** (University of Calgary) **Andrew Irwin** (Dalhousie University) **Martin Lysy** (University of Waterloo) **Mireille Schnitzer** (Université de Montréal)

Bridging the Gap: Navigating Collaborative Research with Non-Statisticians

Combler le fossé : Naviguer la recherche collaborative avec des non-statisticien-ne-s

Collaborative research between statisticians and scientists from other disciplines has grown remarkably in recent decades. Although rewarding, interdisciplinary research can pose challenges in effectively communicating methodological and applied statistical research to non-statisticians. This panel brings together a group of seasoned statisticians with a proven track record of fruitful collaborative work with non-statisticians. They will share insights on addressing the intricacies of multidisciplinary research, including essential skills, publication strategies, and navigating diverse cultures across disciplines. Representing varied research domains, the panelists provide attendees with an opportunity to observe and grasp the nuances of collaborative work in different fields. This is an interactive and engaging session, including open discussion and a Q&A period, encouraging participants to exchange ideas.

La recherche collaborative entre statisticien-ne-s et scientifiques d'autres disciplines s'est considérablement développée au cours des dernières décennies. Bien qu'enrichissante, la recherche interdisciplinaire peut poser des défis de communication efficace de la recherche statistique méthodologique et appliquée aux non-statisticien-ne-s. Ce panel rassemble un groupe de statisticien-ne-s chevronné-e-s ayant fait leurs preuves avec un historique de collaborations multidisciplinaires fructueuses. Ils partageront leurs idées sur la manière d'aborder les subtilités de la recherche multidisciplinaire, y compris les compétences essentielles, les stratégies de publication et la navigation dans diverses cultures disciplinaires. Représentant des domaines de recherche variés, les panélistes offrent aux participants l'occasion d'observer et de saisir les nuances du travail collaboratif dans différents domaines. Il s'agit d'une session interactive et engageante, comprenant une discussion ouverte et une période de questions-réponses, au cours de laquelle les personnes participantes seront encouragées à échanger des idées.

Chair/Président: Marcos Escobar-Anel

Organizer/Responsable: Marcos Escobar-Anel

Room/Salle: C 2033

Date: Tuesday June 4 / mardi 4 juin

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-14:00]

Marcos Escobar-Anel (Western University)

Portfolio Optimization in Affine GARCH models

Optimisation de portefeuille dans les modèles affines GARCH

This talk reveals analytical solutions to dynamic portfolio allocation problems in a discrete time setting where the price process of the risky assets follows various affine GARCH models. The results include Gaussian, Inverse Gaussian and Levy cases of affine GARCH, as well as solutions within the frameworks of expected utility theory, with and without consumption, as well as mean-variance theory. The impact of heteroscedasticity on portfolio decision is exemplified on a concrete investment involving the S&P500.

Cette présentation fournit des solutions analytiques à des problèmes d'attribution de portefeuilles dynamiques dans un cadre de temps discret où le processus de prix des actifs risqués répond à divers modèles affines GARCH. Les résultats incluent les cas gaussiens, gaussiens inversés et de Levy du modèle affine GARCH, ainsi que des solutions dans le cadre de la théorie de l'utilité attendue, avec et sans consommation, et de la théorie de la moyenne-variance. L'impact de l'hétéroscédasticité sur les décisions de portefeuille est illustré par un investissement concret dans l'indice S&P500.

[14:00-14:30]

Bruno N. Rémillard (HEC Montréal) **Jean Vaillancourt** (HEC Montréal) **Pierre Laroche** (Banque Nationale du Canada)

Parrondo's Paradox and Financial Applications

Le paradoxe de Parrondo et applications financières

In this talk, I will start by giving an introduction to Parrondo's paradox, then I will present recent results on this topic, and finally I will talk about financial applications.

Dans cet exposé, je commencerai par une introduction du célèbre paradoxe de Parrondo, suivi d'une présentation de résultats récents sur le sujet, et je terminerai avec un exemple d'application en finance.

[14:30-15:00]

Anatoliy V. Swishchuk (University of Calgary)

Applications of Geometric Compound Hawkes Process in Finance

Applications du processus de Hawkes composé géométrique en finance

We introduce a new model for a stock price, namely, geometric compound Hawkes process, and show how this model can be applied to solving many problems in finance, including European and American option pricing (perpetual American options), and Merton portfolio optimization problem. This model is a generalization of some well-known models in finance, such as Cox-Ross-

Nous introduisons un nouveau modèle pour le prix d'une action, à savoir le processus de Hawkes composé géométrique, et montrons comment ce modèle peut être appliqué à la résolution de nombreux problèmes en finance, y compris l'évaluation des options européennes et américaines (options américaines perpétuelles), et le problème d'optimisation du portefeuille de Merton. Ce modèle est une généralisation de modèles bien connus en finance, tels que

Rubinstein model (1976) (geometric binomial process), Aase model (1988) (geometric compound Poisson process) and geometric Markov renewal model (2013). Numerical examples are presented as well.

le modèle de Cox-Ross-Rubinstein (1976) (processus binomial géométrique), le modèle d'Aase (1988) (processus de Poisson composé géométrique) et le modèle de renouvellement markovien géométrique (2013). Nous présentons également des exemples numériques.

Chair/Président: Lei Sun

Organizer/Responsable: Lei Sun

Room/Salle: C 2045

Date: Tuesday June 4 / mardi 4 juin

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-14:00]

Qihuang Zhang (McGill University) **Qicheng Zhao** (McGill University)

Bayesian Model for Disease-Specific Gene Detection in High-Dimensional Spatially Resolved Transcriptomics

Modèle bayésien pour la détection de gènes spécifiques à la maladie dans la transcriptomique à haute résolution spatiale

Identifying disease-indicative genes is critical for deciphering disease mechanisms and continues to attract significant interest. Spatial transcriptomics offers unprecedented insights for the detection of disease-specific genes by enabling within-tissue contrasts. However, this new technology poses challenges for conventional statistical models developed for RNA-seq, as these models often neglect the spatial organization of tissue spots. In this talk, we discuss a new Bayesian shrinkage model to characterize the relationship between high-dimensional gene expressions and the disease status of tissue spots, incorporating spatial correlation among these spots through autoregressive terms. Our model adopts a hierarchical structure to accommodate for the missing data within tissues and is further extended to facilitate the analysis of multiple correlated samples. To ensure the model's applicability to datasets of varying sizes, we carry out two computational frameworks for Bayesian parameter estimation, tailored to both small and large sample scenarios. Simulation studies are conducted to evaluate the performance of the proposed model, and we also apply our model to analyze the data arising from a HER2-positive breast cancer study.

L'identification des gènes indicatifs de la maladie est essentielle pour déchiffrer les mécanismes de la maladie et continue de susciter un grand intérêt. La transcriptomique spatiale offre des perspectives sans précédent pour la détection de gènes spécifiques à une maladie en permettant des contrastes intra-tissulaires. Cependant, cette nouvelle technologie pose des défis aux modèles statistiques conventionnels développés pour l'ARN-seq, car ces modèles négligent souvent l'organisation spatiale des taches tissulaires. Dans cet exposé, nous discutons d'un nouveau modèle bayésien de rétrécissement pour caractériser la relation entre les expressions génétiques à haute dimension et l'état pathologique des taches tissulaires, en incorporant la corrélation spatiale entre ces taches par le biais de termes autorégressifs. Notre modèle adopte une structure hiérarchique pour tenir compte des données manquantes dans les tissus et est étendu pour faciliter l'analyse d'échantillons multiples corrélés. Pour garantir l'applicabilité du modèle à des ensembles de données de différentes tailles, nous réalisons deux cadres de calcul pour l'estimation bayésienne des paramètres, adaptés aux scénarios de petits et de grands échantillons. Des études de simulation sont menées pour évaluer les performances du modèle proposé, et nous appliquons également notre modèle pour analyser les données réelles.

[14:00-14:30]

Lin Zhang (Simon Fraser University, Burnaby) **Lei Sun** (University of Toronto) **Andrew Paterson** (The Hospital for Sick Children)

Allele-frequency Estimation and Ancestry Informative Marker Identification via Retrospective Regression

Estimation de la fréquence allélique et identification de marqueur informatif sur l'ascendance à l'aide d'une régression rétrospective

Allele frequency estimation at a genetic marker plays a pivotal role in genetic studies. The accuracy of allele

L'estimation de la fréquence allélique pour un marqueur génétique joue un rôle essentiel dans les études génétiques. La précision de

frequency estimation impacts the accuracy and power of a genome-wide association study (GWAS). Moreover, allele frequency may differ between seemingly similar populations, which makes allele frequency estimation particularly important for identifying ancestral informative markers (AIMs). Yet, existing allele frequency estimation methods mostly rely on independent sample from a homogeneous population and cannot provide closed form solutions for the maximum likelihood estimator (MLE) of the allele frequencies. To address these challenges, we propose a retrospective regression framework that takes genotype as the response variable, and population and other covariates as the dependent variable. The regression nature of our proposed method enables it to estimate allele frequency in heterogeneous populations and accommodate sample correlation. We support our analytical findings using the 1000 Genome Project genotype data of five super-populations.

l'estimation de la fréquence allélique influence la précision et la puissance d'une étude d'association pangénomique (GWAS). De plus, la fréquence allélique peut varier selon des populations semblablement similaires, ce qui rend son estimation particulièrement importante afin d'identifier les marqueurs informatifs d'ascendance (AIMs). Pourtant, les méthodes actuelles d'estimation de la fréquence allélique se basent en grande partie sur un échantillon indépendant tiré d'une population homogène et ne peuvent pas obtenir de solution analytique pour l'estimateur de vraisemblance maximum (MLE) des fréquences alléliques. Afin d'aborder ces défis, nous proposons un cadre de régression rétrospective qui se sert du génotype en guise de variable de réponse, tandis que la variable de dépendance est basée sur la population et d'autres covariables. La nature régressive de notre méthode lui permet d'estimer la fréquence allélique dans des populations hétérogènes et facilite la corrélation d'échantillon. Nous soutenons nos résultats d'analyse grâce aux données de génotype de cinq super populations tirées du projet 1000 génomes.

[14:30-15:00]

Yongjin P. Park (The University of British Columbia) **Sishir Subedi** (University of British Columbia) **Tomokazu Sumida** (Yale University)

Probabilistic Topic Modelling to Eavesdropping Cell-Cell Communication Patterns in Spatial Gene Expression Data

Modélisation de sujet probabiliste afin de reconnaître les tendances de communication entre cellules dans des données d'expression génique spatiale

We investigate how cancer cells exploit immune systems by falsely propagating disguising messages to neighbouring immune cells using single-cell spatial transcriptomic data. We developed a scalable topic modelling approach to sort out millions of cell-cell interaction patterns while statistically learning tumour-to-microenvironment communication patterns by "eavesdropping" on gene expression patterns that co-occur between cancer and immune cells. We propose a topic model-based approach to ascertain common topics/gene expression programs embedded in millions of cell-cell interaction patterns stochastically expressed in gene/DNA accessibility vectors of ten thousand genes. The key idea is to randomly project high-dimensional data onto lower-dimensional space, build summary data, and efficiently perform topic modelling. We hypothesize and confirm that novel pathogenic gene programs can be distilled by examining genes that substantially co-expressed between cancer and immune cells.

À l'aide de données de transcriptomique spatiale à cellule unique, nous étudions comment les cellules cancéreuses exploitent les systèmes immunitaires en propageant faussement des messages aux cellules immunitaires avoisinantes. Nous avons développé une approche de modélisation thématique extensible dans le but de mettre en ordre les millions de tendances d'interactions entre cellules tout en apprenant de façon statistique quelles sont les communications entre tumeur et micro-environnement en « écoutant » les expressions géniques qui surviennent entre les cellules cancéreuses et celles qui sont immunitaires. Nous proposons une approche basée sur un modèle thématique pour établir les programmes d'expression génique et les thèmes communs incorporés dans des millions d'interactions entre cellules exprimées de manière stochastique dans des vecteurs d'accessibilité de gène-ADN de dix mille gènes. L'idée principale est de projeter aléatoirement des données de grande dimension sur un espace de faible dimension, construire des données de sommaire, puis réaliser efficacement une modélisation thématique. Nous conjecturons et confirmons que les nouveaux programmes de gène pathogène peuvent être extraits en examinant les gènes qui se sont substantiellement exprimés entre les cellules cancéreuses et immunitaires.

20th Anniversary of SSC Accreditation: Core Principles
Le 20è anniversaire de l'accréditation par la SSC : principes fondamentaux

Chair/Président: Hugh Chipman

Organizer/Responsable: Judy-Anne W. Chapman

Room/Salle: A 1046

Date: Tuesday June 4 / mardi 4 juin

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-14:00]

Peter D.M. Macdonald (McMaster University)
University Course Requirements for Accreditation
Exigences en matière de cours universitaires pour l'accréditation

Educational requirements for A.Stat. are set down in the Accreditation Document as a generic list of courses covering topics essential to the training of a statistician. The SSC maintains a list of universities with courses pre-approved by the Accreditation Committee; A.Stat. applicants who have completed these courses need not provide further documentation. Applicants for P.Stat. who are not already A.Stat. are expected to meet A.Stat. requirements, with exceptions for senior members who have made substantial contributions to the field. The list of courses in the Accreditation Document is intended to distinguish professional statisticians from others such as mathematical scientists, data scientists and those who apply statistical methodology. Because definitions of professions, university courses, and even the words we use to talk about such things, change over time, it is important that statisticians constantly re-think what defines our profession and its educational requirements.

Les exigences de formation des candidats au titre d'A.Stat. sont définies dans le Document d'accréditation sous la forme d'une liste générique de cours couvrant les sujets essentiels à la formation d'un statisticien. La SSC tient à jour une liste d'universités dont les cours sont préapprouvés par le Comité d'accréditation ; les candidats au titre d'A.Stat. qui ont suivi ces cours n'ont pas besoin de fournir d'autres documents. Les candidats au titre de P.Stat. qui ne sont pas déjà A.Stat. doivent satisfaire aux exigences de la qualification A.Stat., à l'exception des membres seniors qui ont apporté une contribution substantielle au domaine. La liste des cours figurant dans le Document d'accréditation vise à distinguer les statisticiens professionnels d'autres personnes telles que les mathématiciens, les spécialistes des données et ceux qui appliquent la méthodologie statistique. Étant donné que les définitions des professions, des cours universitaires et même des mots que nous utilisons pour parler de ces choses changent avec le temps, il est important que les statisticiens repensent constamment ce qui définit notre profession et ses exigences en matière de formation.

[14:00-14:30]

Tony Panzarella (University of Toronto)
The Statistical Society of Canada's Code of Ethical Statistical Practice: An Effective Road Map to Promoting High Professional Standards
Le Code de déontologie statistique de la Société statistique du Canada : Une feuille de route efficace pour promouvoir des normes professionnelles élevées

Twenty years ago the Statistical Society of Canada adopted a Code of Ethical Statistical Practice as part of its formal Accreditation Program. The code comprises four elements: 1) Responsibility to Society, 2) Responsibility to Employers/ Clients, 3) Responsibility

Il y a vingt ans, la Société statistique du Canada a adopté un code de déontologie statistique dans le cadre de son programme d'accréditation officiel. Ce code comprend quatre éléments : 1) Responsabilité envers la société, 2) Responsabilité envers les employeurs et les clients, 3) Responsabilité envers les autres prati-

to Other Statistical Practitioners, and 4) Professionalism. As a P.Stat. accredited statistician since 2008, I use examples from a career as a consulting biostatistician in a tertiary care academic hospital to demonstrate that the code is an effective road map to promoting high professional standards, and argue that it should be ubiquitous in the education and training of new statistical practitioners.

ciens de la statistique et 4) Professionnalisme. En tant que statisticien accrédité P.Stat. depuis 2008, j'utilise des exemples tirés d'une carrière de biostatisticien consultant dans un hôpital universitaire de soins tertiaires pour démontrer que le code est une feuille de route efficace pour promouvoir des normes professionnelles élevées, et je soutiens qu'il devrait être omniprésent dans l'éducation et la formation des nouveaux praticiens de la statistique.

[14:30-15:00]

Milena Kurtinecz (Bayer Pharmaceuticals)

Accessible Variety of Professional Development

Formation professionnelle diversifiée et accessible

Maintenance of professional competencies is a core requirement of SSC Accreditation. Year-round Professional Development (PD) is facilitated with professional practice tools in the Accredited-only area of the SSC website which includes slide decks, and now a Workshop video, from past SSC Annual Meetings, Regional Workshops, International meetings, and links to upcoming resources. A Book Club gives working applied statisticians an opportunity for PD by working through an applied textbook. Networking is encouraged with the searchable Accreditation database.

Le maintien des compétences professionnelles est une exigence fondamentale de l'accréditation à la Société statistique du Canada (SSC). La formation professionnelle offerte tout au long de l'année est facilitée par des outils pratiques professionnels accessibles dans la zone réservée aux membres accrédités sur le site Web de la SSC, y compris des diapositives, des vidéos des réunions annuelles antérieures de la SSC, des ateliers régionaux et des rencontres internationales ainsi que des liens vers des ressources à venir. Un club de lecture donne aux statisticiens en emploi dans le domaine de la statistique appliquée des possibilités de formation professionnelle à l'aide d'un manuel de statistique appliquée. Le réseautage est encouragé par la base de données consultable de l'accréditation.

Chair/Président: Jemila Seid Hamid

Organizer/Responsable: Jemila Seid Hamid

Room/Salle: A 1043

Date: Tuesday June 4 / mardi 4 juin

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-15:00]

Pierre R. L. Dutilleul (McGill University)

"Spatial, temporal and multidimensional Statistics: Estimation, testing, and applications in the environmental sciences" – An overview with novelties

« Statistique spatiale, temporelle et multidimensionnelle : estimation, test et applications aux sciences de l'environnement » - une vue d'ensemble avec des nouveautés

In this SSC 2023 Impact Award Address, I will (1) offer a tour of ¿25 years of research results, and (2) report new results for an extension of recently published work with Prof. K. Shimizu and Dr. T. Imoto. (1) I will speak about modified tests of significance in correlation and regression analyses with space-time data; mono- and multi-fractal analyses of natural structures; estimation and testing under the matrix and tensor normal distributions; multi-scale multivariate analysis with spatial data; and multi-frequential periodogram analysis, with applications in plant ecology and tree biology, the soil sciences and seismology. (2) I will show how to extend to 3+ years the 1-year and 2-year cases in Dutilleul et al. (2024, JABES, DOI: 10.1007/s13253-023-00599-2). Over 3+ years, the preferred direction of tree radial growth, as inferred from CT images, can be compared statistically by modeling the double variance-covariance structure (angular and among years) with a Kronecker product.

Dans cette allocution du Prix pour impact de la SSC 2023, je (1) proposerai un tour d'horizon de plus de 25 ans de résultats de recherche et (2) présenterai de nouveaux résultats pour une extension des travaux récemment publiés avec le professeur K. Shimizu et le docteur T. Imoto. (1) Je parlerai des tests modifiés de signification dans les analyses de corrélation et de régression avec des données spatio-temporelles ; des analyses monofractales et multifractales des structures naturelles ; de l'estimation et des tests sous les distributions normales matricielles et tensorielles ; de l'analyse multivariée multi-échelle avec des données spatiales ; et de l'analyse des périodogrammes multifréquentiels, avec des applications à l'écologie végétale et à la biologie des arbres, aux sciences du sol et à la sismologie. (2) Je montrerai comment étendre à 3+ ans les cas de 1 an et 2 ans de Dutilleul et al. (2024, JABES, DOI : 10.1007/s13253-023-00599-2). Sur 3+ ans, la direction préférée de la croissance radiale de l'arbre, déduite des images CT, peut être comparée statistiquement par une modélisation de la double structure de variance-covariance (angulaire et entre les années) avec un produit de Kronecker.

Chair/Président: Jinko Graham

Room/Salle: C 4036

Date: Tuesday June 4 / mardi 4 juin

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-13:45]

Hong Gu (Dalhousie University) **Chaoyue Liu** (Dalhousie University) **Toby J. Kenney** (Dalhousie University) **Robert Beiko** (Dalhousie University) **Zesheng Jia** (Dalhousie University)

The Community Coevolution Model and Machine Learning Approach for Phylogenetic Comparative Analysis

Modèle de coévolution communautaire et approche d'apprentissage automatique pour l'analyse comparative phylogénétique

Organismal traits can evolve in a coordinated way, with correlated patterns of gains and losses reflecting important evolutionary associations. Phylogenetic profiles treat individual genes as traits distributed across sets of genomes and can be used to identify functionally linked genes or lateral gene transfer. We propose the Community Coevolution Model (CCM) to analyze the evolutionary associations of traits (genes) based on phylogenetic profiles with traits evolving as a community with interactions, and the transition rate for each trait depends on the current states of other traits. We show that CCM is more efficient and fits real data better than other methods resulting in higher likelihood scores with fewer parameters. We further introduce a machine learning method (CNN-CCM) to estimate the parameters for large community CCM and demonstrate its efficiency through both simulations and real data.

Les caractéristiques des organismes peuvent évoluer de manière coordonnée, avec des schémas corrélés de gains et de pertes reflétant d'importantes associations évolutives. Les profils phylogénétiques considèrent les gènes individuels comme des traits distribués dans des ensembles de génomes et permettent d'identifier des gènes fonctionnellement liés ou un transfert latéral de gènes. Nous proposons le modèle de coévolution communautaire (MCC) pour analyser les associations évolutives de traits (gènes) basées sur les profils phylogénétiques, les traits évoluant comme une communauté avec des interactions, et le taux de transition pour chaque trait dépendant de l'état actuel des autres traits. Nous montrons que le MCC est plus efficace et s'adapte mieux aux données réelles que les autres méthodes, ce qui se traduit par des scores de vraisemblance plus élevés avec moins de paramètres. Nous introduisons en outre une méthode d'apprentissage automatique pour estimer les paramètres des MCC de grandes communautés et démontrons son efficacité à l'aide de simulations et de données réelles.

[13:45-14:00]

Jiaqi Bi (University of Western Ontario) **Osvaldo Espin-Garcia** (University of Western Ontario) **Yun-Hee Choi** (University of Western Ontario)

Correlated Shared Frailty Model Incorporating Ascertainment Correction with Missing Covariates in Family-Based Studies

Modèle de fragilité partagée corrélée intégrant la correction de l'incertitude avec des covariables manquantes dans les études familiales

In the analysis of clustered survival data arising from family-based studies with missing covariates, current multiple imputation (MI) methods do not handle the hierarchical structure of the data and the ascertainment of families. Especially when the time-to-event and the proband information in a genomics study should be con-

Dans l'analyse des données de survie en grappes provenant d'études familiales avec des covariables manquantes, les méthodes actuelles d'imputation multiple (IM) ne gèrent pas la structure hiérarchique des données et la vérification des familles, notamment lorsque le temps écoulé jusqu'à l'événement et les informations sur le probant dans une étude génomique doivent être condi-

ditioned when sampling the missing data distribution. We propose a Monte Carlo Expectation Maximization (MCEM) method and further adapt it into an MI method considering the family structure and the proband information using kinship matrix. Through simulations of family-clustered survival data with covariates missing at random (MAR) and an application to breast cancer families recruited from the Breast Cancer Family Registries with missing PRS and mutation gene status, our study aims to evaluate the effectiveness of the proposed method by comparing its performance to a complete case analysis.

tionnés lors de l'échantillonnage de la distribution des données manquantes. Nous proposons une méthode de maximisation de l'espérance de Monte Carlo (MCEM) et l'adaptons ensuite en une méthode IM tenant compte de la structure familiale et des informations sur le probant à l'aide de la matrice de parenté. Grâce à des simulations de données de survie regroupées par famille avec des covariables manquantes au hasard (MAR) et à une application aux familles de cancer du sein recrutées dans les registres familiaux du cancer du sein avec des données manquantes sur le SRP et le statut du gène de mutation, notre étude vise à évaluer l'efficacité de la méthode proposée en comparant ses performances à celles d'une analyse de cas complète.

[14:00-14:15]

Chenyang Li (Western University) **Osvaldo Espin-Garcia** (Western University)

Optimizing Linear Polygenic Risk Score Combinations for Two-Phase Re-sequencing Study Design

Optimiser les combinaisons de scores de risques polygéniques linéaires pour une étude de reséquençage à deux phases

The complexity of genetic architecture of traits increases the challenges in polygenic risk score (PRS) construction and reduces accuracy of prediction, given that a single PRS method might not summarize the genomic susceptibility of traits comprehensively. Recent work has developed two-phase designs in re-sequencing studies where only informative subsamples are selected for cost-effective data collection. Here, we propose an optimization approach integrating multiple PRS methods in a two-phase design for re-sequencing studies. Set in linear regression framework, our model utilizes a constrained residual dependent sampling (RDS) design to accommodate a mixture of two PRS methods: LassoSum and LD-pred-inf, which posit contrasting assumptions with respect to the trait underlying genetic architecture. The method is evaluated against the traditional RDS designs with single or both PRS methods via simulation and applied to data from Northern Finland Birth Cohort of 1966 study.

La complexité de l'architecture génétique des traits augmente les difficultés liées à la construction d'un score de risque polygénique (PRS) et réduit la précision de la prévision, étant donné qu'une seule méthode PRS pourrait ne pas résumer de manière exhaustive la sensibilité génomique des traits. Des travaux récents ont permis de mettre au point des modèles en deux phases pour des études de reséquençage dans lesquelles seuls les sous-échantillons informatifs sont sélectionnés pour une collecte de données rentable. Nous proposons ici une approche d'optimisation comportant plusieurs méthodes de PRS dans un plan en deux phases pour des études de reséquençage. Notre modèle s'établit dans un cadre de régression linéaire et emploie un échantillonnage de dépendance résiduel restreint (RDS) pour s'ajuster à un mélange de deux méthodes PRS : LassoSum et LD-pred-inf, qui postulent des hypothèses contrastantes en ce qui concerne l'architecture génétique de trait sous-jacente. La méthode est évaluée contre les modèles de RDS traditionnels avec l'une ou les deux méthodes PRS par l'entremise de simulations et appliquée à des données de l'étude « Northern Finland Birth Cohort » de 1966.

[14:15-14:30]

Patrick McMillan (University of Guelph) **Zeny Feng** (University of Guelph) **Lewis Lukens** (University of Guelph)

Improving Crop Variety Recommendations for Farmers: An Integrated Approach using Machine Learning and Genetics

Recommandations pour améliorer la variété des types de cultures agricoles : une approche intégrée utilisant l'apprentissage machine et la génétique

Crop variety selection is one of the most important factors influencing on-farm yields. Identifying suitable varieties for farms can be difficult as the relative performance of varieties often varies across environments due to genotype by environment interactions. This research seeks to address this issue by improving the accu-

Le choix de types de cultures variés est un des plus importants facteurs qui influent sur le rendement agricole. L'identification de variétés appropriées pour les fermes peut se révéler difficile puisque le rendement relatif des variétés peut différer selon les environnements en raison des interactions génotype-environnement. Cette recherche vise à s'attaquer à ce problème en améliorant

racy of variety recommendations using machine learning and single nucleotide polymorphism (SNP) data to better capture the genotype by environment interaction effect. We implement Bayesian additive regression trees (BART) to analyze 13 years of variety trials from across Ontario. We find that the BART model is able to consistently provide significantly better variety recommendations relative to the mixed effects models commonly used in variety trials. This improvement in accuracy was in part due to the BART model's ability to capture SNP-SNP interactions and nonlinear SNP effects better than the mixed effects model.

les recommandations sur la diversification des cultures, en utilisant l'apprentissage machine et des données de polymorphisme nucléotide simple (SNP) pour mieux saisir le génotype par l'effet d'interaction génotype-environnement. Nous implémentons des arbres de régression additive bayésienne (BART) pour l'analyse d'essais de variétés faits pendant 13 ans dans l'ensemble de l'Ontario. Nous découvrons que le modèle BART peut fournir de façon consistante de meilleures recommandations de variétés relativement aux modèles à effets mixtes d'utilisation courante dans les essais sur les variétés. Cette amélioration de l'exactitude relève en partie de la capacité du modèle BART de saisir les interactions SNP-SNP et de meilleurs effets SNP non linéaires que ceux du modèle à effets mixtes.

[14:30-14:45]

Yuan Sun (Lunenfeld-Tanenbaum Research Institute, Sinai Health, Toronto, Canada) **Laurent Briollais** (Lunenfeld-Tanenbaum Research Institute, Sinai Health, Toronto, Canada; Dalla Lana School of Public Health, University of Toronto, Toronto, Canada) **Xuming He** (Department of Statistics and Data Science, Washington University in St. Louis, St. Louis, USA)

A Two-Stage Model for Genome-Wide Association Study

Modèle à deux étapes pour études d'association à l'échelle du génome

Many diseases are influenced by the marginal effects of genetic covariates (G) and environmental covariates (E), as well as their interactions. These interactions are most commonly addressed by adding the GxE terms to the model. However, including the GxE terms may complicate the model, especially when the dimension of G is large. Furthermore, GxE only captures one specific type of interactions, whereas the true interactions can be more general. In this project, we propose a two-stage model as a solution to the aforementioned problems. In the first stage, we calculate the conditional percentile for each individual, adjusting for all the E factors with a global quantile regression model. In the second stage, we select G factors that are associated with the conditional percentile. By modeling the impact of genes and the environment separately in two stages, our proposed method is can identify associated gene markers that potentially have complex interactions with the environment.

De nombreuses maladies sont influencées par les effets marginaux de covariables génétiques (G) et environnementales (E), ainsi que par leurs interactions. Ces dernières sont le plus souvent traitées en ajoutant les termes GxE au modèle. Cependant, l'inclusion de termes GxE peut compliquer le modèle, en particulier lorsque la dimension de G est importante. En outre, GxE ne rend compte que d'un type spécifique d'interactions, alors que les véritables interactions peuvent être plus générales. Dans ce projet, nous proposons un modèle à deux étapes pour résoudre les problèmes susmentionnés. Dans un premier temps, nous calculons le percentile conditionnel pour chaque individu, en tenant compte de tous les facteurs E à l'aide d'un modèle de régression quantile global. Dans un deuxième temps, nous sélectionnons les facteurs G qui sont associés au percentile conditionnel. En modélisant l'impact des gènes et de l'environnement séparément en deux étapes, la méthode que nous proposons permet d'identifier les marqueurs génétiques associés qui ont potentiellement des interactions complexes avec l'environnement.

[14:45-15:00]

Brady Ryan (University of Michigan) **Michael Boehnke** (University of Michigan) **Ryan Welch** (University of Michigan) **Christian Fuchsberger** (Eurac Research)

Using External Reference Panel and Single-Variant Summary Statistics for Rare-Variant Aggregation Tests

Utilisation d'un panel de référence externe et de statistiques récapitulatives à variante unique pour les tests d'agrégation de variants rares

Rare-variant aggregation tests are often used to increase

Les tests d'agrégation de variants rares sont souvent utilisés afin

the power to detect rare-variant associations, and can be performed using publicly available GWAS summary statistics. Accurate estimates of single-variant test statistic covariances are needed, but are often unavailable due to data size and inability to share individual-level genetic data. In this study we extend a previously proposed method of estimating covariance from an external reference panel to rare-variant aggregation tests. We apply our method to UK Biobank exome sequence data and observe squared Pearson correlation coefficients comparing aggregation test -log10(p-values) generated using individual-level data and UK Biobank external reference panels between 0.984 and 0.996 at a reference panel of 1,000 across multiple simulation settings. By using an African American reference panel (n=1,000) from the InPSYght study, we also show that our approach is robust to misspecification of the reference panel ancestry.

d'accroître la puissance de détection des associations de variants rares et on peut les faire à l'aide de statistiques récapitulatives accessibles au public. Bien qu'elles soient nécessaires, des estimations exactes des covariances de tests statistiques à variante unique sont souvent non disponibles en raison de la taille des données et de l'incapacité de partager des données génétiques individuelles. Notre étude étend à des tests d'agrégation de variants rares une méthode proposée antérieurement pour l'estimation de la covariance à partir d'un panel de référence externe. Nous appliquons notre méthode à des données de séquençage de l'exome de la UK Biobank et observons le carré de coefficients de corrélation de Pearson, en comparant un test d'agrégation -log10(valeurs-p) généré à l'aide de données individuelles et de panels de référence externes de la UK Biobank entre 0,984 et 0,996 d'un panel de référence de 1 000 parmi de multiples paramètres de simulation. En utilisant un panel de référence afro-américain (n = 1 000) tiré de l'étude InPSYght, nous montrons également que notre approche est robuste à toute spécification erronée de l'ascendance du panel de référence.

Infectious Disease
Maladies infectieuses

Chair/Président: Lisa M. Lix

Room/Salle: C 3053

Date: Tuesday June 4 / mardi 4 juin

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-13:45]

Jeffrey W. Peitsch (University of Calgary)

Directionally Dependent Individual Level Models for Infectious Disease

Modèles au niveau individuel avec dépendance directionnelle pour une maladie infectieuse

Outbreaks of infectious diseases, such as COVID-19, can pose direct threats to public health and may have serious health and economic consequences. In many epidemic systems, infections tend to spread in one general direction more than others, which can be due to migration and behaviour patterns in human and animal epidemics, or due to wind patterns for plant epidemics. Mathematical individual level models (ILMs) have been developed for infectious disease transmission but have not traditionally considered these directional tendencies. I will introduce a class of ILMs that consider the directional dynamics of disease transmission. In these directionally dependent ILMs, the probability of an individual being infected depends on both the direction and distance between susceptible and infectious individuals. I will discuss the characteristics of these directionally dependent ILMs, and how they can be fitted in a Bayesian Markov chain Monte Carlo (MCMC) framework.

Des éclosions de maladies infectieuses, comme la COVID-19, représentent une menace à la santé publique et peuvent avoir de graves conséquences économiques et sur la santé. Dans plusieurs systèmes épidémiques, les infections ont davantage tendance à se propager dans une direction générale plus que dans d'autres, ce qui peut être causé par les migrations ou des schémas de comportements dans les épidémies animales et humaines, ou les configurations des vents pour les épidémies végétales. Les modèles mathématiques à niveau individuel (ILMs) ont été conçus pour la transmission de maladie infectieuse, mais n'ont pas traditionnellement tenu compte de ces tendances directionnelles. Je présenterai une classe de ILMs qui tient en compte des dynamiques directionnelles d'une transmission de maladie. Dans ces ILMs avec dépendance directionnelle, la probabilité qu'un individu soit infecté dépend de la direction et la distance entre les individus infectieux et ceux qui sont vulnérables. Je présenterai les caractéristiques de ces ILMs avec dépendance directionnelle et expliquerai de quelle façon ils peuvent être ajustés dans un cadre bayésien par la méthode de Monte Carlo par chaînes de Markov (MCMC).

[13:45-14:00]

Rado Malalatiana Ramasy (Université de Montréal) **William Ruth** (Université de Montréal)

Multilevel Mediation Analysis : Deciphering the Impact of Information Sources on Adherence to Restrictive Measures during the COVID-19 Pandemic.

Analyse de médiation multiniveau : Décrypter l'impact des sources d'information sur l'adhésion aux mesures restrictives durant la pandémie de COVID-19

Our research focuses on analyzing factors that influence compliance with public health directives. We analyze data collected by the Real-Time Interactive World-Wide platform, encompassing responses from over one hundred thousand participants across eleven countries during the COVID-19 pandemic. We use a mediation anal-

Cette étude se concentre sur les facteurs influençant la conformité aux directives de santé publique. Nous analysons les données collectées par la plateforme Real-Time Interactive World-Wide, couvrant plus de cent mille participants dans onze pays durant la pandémie de COVID-19. Nous utilisons une analyse de médiation pour évaluer comment la source d'information im-

ysis to assess how peoples'preferred source of information impacts their intention to comply with future directives, considering previous adherence as a mediator. We adapt Baron and Kenny's method for calculating mediation effects, adjusting the models for confounding variables such as age, sex, and education level. We also incorporate a multi-level structure to account for the hierarchical nature of our dataset. The bootstrap allows for robust estimation of the standard errors. Preliminary results suggest that the influence of information sources on the intention to comply with directives in the future is largely mediated by the level of adherence previously observed.

pacte les intentions de se conformer aux directives futures, en considérant l'adhérence antérieure comme médiateur. Nous adaptons la méthode de Baron et Kenny pour calculer les effets de médiation, en ajustant les modèles sur les confondants comme l'âge, le sexe et le niveau d'éducation. Nous intégrons une approche multiniveau pour tenir compte de la nature hiérarchique des données. Le bootstrap permet une estimation robuste des erreurs standard. Les résultats suggèrent que l'influence des sources d'information sur l'intention de se conformer aux directives est largement médiée par le niveau d'adhérence précédemment observé.

[14:00-14:15]

Gyanendra Pokharel (The University of Winnipeg)

Predictive Probability-based Gaussian Process Emulators for Infectious Disease Models

Émulateurs de processus gaussiens basés sur des probabilités prédictives pour les modèles de maladies infectieuses

Mechanistic models are vital for understanding infectious disease transmission dynamics, particularly with complex space-time data. While Bayesian Markov Chain Monte Carlo (MCMC) serves as a common fitting framework, it is impractical for large populations due to the computational demands. An alternative approach involves utilizing model classification methods for disease generation, but these methods may provide only point estimates, overlooking parameter uncertainty. In this study, we propose an emulation-based method to expedite computation time. We re-fit the model within a Bayesian MCMC framework, replacing the true likelihood of the epidemic-generating model with predictive probabilities obtained from ensemble learning classifiers. By simulating parameter estimates from the posterior distributions, our findings demonstrate improved accuracy and efficiency in inferring disease transmission dynamics, capturing the parameter uncertainty. The models are fitted to the data from the 2001 Great Britain foot-and-mouth epidemic, and simulated data.

Les modèles mécanistes sont essentiels pour comprendre la dynamique de transmission des maladies infectieuses, notamment pour des données spatio-temporelles complexes. Bien que la méthode bayésienne de Monte Carlo par chaînes de Markov (MCMC) serve couramment de cadre d'ajustement, celle-ci n'est pas pratique pour les grandes populations en raison des exigences de calcul. Une autre approche consiste à utiliser des méthodes de classification des modèles pour la génération de maladies, mais ces méthodes peuvent ne fournir que des estimations ponctuelles, sans tenir compte de l'incertitude des paramètres. Dans cette étude, nous proposons une méthode basée sur l'émulation pour accélérer le temps de calcul. Nous réajustons le modèle dans un cadre MCMC bayésien, en remplaçant la vraisemblance du modèle de génération d'épidémies par des probabilités prédictives obtenues à partir de classificateurs d'apprentissage d'ensembles. En simulant les estimations des paramètres à partir des distributions a posteriori, nos résultats démontrent une amélioration de la précision et de l'efficacité dans l'inférence de la dynamique de transmission de la maladie, ainsi qu'une prise en compte de l'incertitude des paramètres. Nous adaptons ces modèles aux données de l'épidémie de fièvre aphteuse de 2001 en Grande-Bretagne et à des données simulées.

[14:15-14:30]

Cong Jiang (University of Montreal) **Mireille Schnitzer** (University of Montreal) **Denis Talbot** (Laval University)

COVID-19 Vaccine Effectiveness Estimation Under the Test-Negative Design

Estimation de l'efficacité du vaccin contre la COVID-19 dans le cadre d'un devis test-négatif

The test-negative design (TND) is routinely used for the monitoring of seasonal flu vaccine effectiveness (VE) and has recently become integral to COVID-19 vaccine

Le devis test-négatif (TND) est couramment utilisé pour surveiller l'efficacité vaccinale (VE) contre la grippe saisonnière et est devenu essentiel à la surveillance des vaccins COVID-19. Contrai-

surveillance. Distinct from the case-control study, it recruits participants with a common symptom presentation and tests them for the target infection. Positive tests are considered "cases," while negative tests are "controls." Logistic regression has traditionally adjusted for confounders to estimate VE in TND, but it may be biased if effect modification by a confounder exists. I will review an inverse probability weighting estimator for the marginal risk ratio, a method valid under effect modification but requires parametric modeling for vaccination probability. Addressing this limitation, we propose a novel doubly robust and efficient estimator of the marginal risk ratio. We then proceed to theoretically and empirically demonstrate the parametric convergence rates achieved through machine learning of the nuisance functions.

rement à l'étude cas-témoins, il recrute des participants avec les mêmes symptômes, testés pour l'infection cible. Les tests positifs sont les cas, alors que les négatifs sont les témoins. La régression logistique ajuste traditionnellement les facteurs de confusion pour estimer la VE dans le TND, mais un biais est possible en cas de modification d'effet par un facteur de confusion. Nous examinons un estimateur de pondération par probabilité inverse pour le rapport de risque marginal, nécessitant une modélisation paramétrique pour la probabilité de vaccination. Nous proposons un nouvel estimateur doublement robuste et efficace du rapport de risque marginal, démontrant théoriquement et empiriquement les taux de convergence paramétriques obtenus par l'apprentissage automatique des fonctions de nuisance.

[14:30-14:45]

Jiaping (Olivia) Liu (University of British Columbia) **Zhenglun Cai** (University of British Columbia) **Paul Gustafson** (University of British Columbia) **Daniel J. McDonald** (University of British Columbia)

RtEstim: Effective Reproduction Number Estimation With Trend Filtering

RtEstim : estimation du nombre effectif de reproduction avec filtrage des tendances

To understand the transmissibility and spread of infectious diseases, epidemiologists turn to estimates of the effective reproduction number. While many estimation approaches exist, their utility may be limited. Challenges of surveillance data collection, model assumptions that are unverifiable with data alone, and computationally inefficient frameworks are critical limitations for many existing approaches. We propose a discrete spline-based approach RtEstim that solves a convex optimization problem—Poisson trend filtering—using the proximal Newton method. It produces a locally adaptive estimator for effective reproduction number estimation with heterogeneous smoothness. RtEstim remains accurate even under some process misspecifications and is computationally efficient, even for large-scale data. The implementation is easily accessible in a lightweight R package rtestim (dajmcdon.github.io/rtestim/).

Pour comprendre la transmissibilité et la propagation des maladies infectieuses, les épidémiologistes recourent à des estimations du nombre effectif de reproduction. Même si les approches d'estimation sont nombreuses, leur utilité peut être limitée : problèmes de collecte de données de surveillance, hypothèses du modèle invérifiables avec les données seules et cadres inefficaces sur le plan computationnel. Dans bon nombre d'approches existantes, ces limites sont critiques. Nous proposons une approche discrète basée sur les splines, RtEstim, qui résout un problème d'optimisation convexe – le filtrage de tendances de Poisson – en utilisant la méthode proximale de Newton. Elle produit un estimateur localement adaptatif pour une estimation efficace du nombre de reproduction, avec une fluidité hétérogène. RtEstim conserve son exactitude, même en cas de certaines spécifications erronées de processus, et est efficace sur le plan computationnel, même pour des données à grande échelle. L'implantation est facilement accessible dans un package R léger rtestim. (dajmcdon.github.io/rtestim/).

Chair/Président: Aya A. Mitani

Room/Salle: ED 2018B

Date: Tuesday June 4 / mardi 4 juin

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-13:45]

François A. Marshall (Self employed) **Lenin Arango-Castillo** (Bank of Mexico)

Coherent Nationwide Variation in U.S. Air Pollution: A Novel Switching Model

Variation cohérente à l'échelle nationale de la qualité de l'air aux États-Unis : Un nouveau modèle à transfère

Adverse to our lungs, fine particulate matter originates either directly (infrastructure, fields, smoke) or indirectly (chemical reactions involving sulfur / nitrogen compounds). For the U.S., a 1999-2019 time series of daily Air Quality Index (AQI) manifests a geographic dichotomy: in 10 western states, AQI spikes annually; and for the other states, the peak is more gradual. Here, it is shown that AQI switches geographically: in the 10 western states, detrended AQI is polychromatic-integrated SARIMA; and for the other states, it is fractional-ARIMA. Both 2020-2023 forecasts and a spatiotemporal residual analysis reveal good fit.

Les particules fines dommageables à nos poumons : proviennent soit directement (infrastructures, champs, fumées) soit indirectement (composés chimiques aux soufrés/azotés). Aux États-Unis, une série temporelle 1999-2019 de l'indice de la qualité de l'air (IQA) révèle une dichotomie géographique : dans 10 États à ouest, le IQA augmente rapidement chaque année ; et pour les autres États, ce pic augmente plus lentement. Dans cette présentation, nous montrons que le IQA change géographiquement : dans les 10 États de ouest, la tendance de l'IQA est une SARIMA intégrée polychromatique ; et pour les autres États, il s'agit d'un modèle fractionnaire-ARIMA. Les prévisions 2020-2023 et l'analyse résiduelle spatio-temporelle révèlent un un bon ajustement.

[13:45-14:00]

Yunhong Lyu (University of Montreal) **Bouchra Nasri** (University of Montreal) **Bruno N. Rémillard** (HEC Montréal)

Sequential Change-point Detecting with Generalized Ornstein–Uhlenbeck Processes

Détection séquentielle de points de changement à l'aide de processus d'Ornstein-Uhlenbeck généralisés

In this talk, we present a stochastic process which is suitable for positive financial data with a cyclic mean-reverting behaviour. The proposed stochastic process is a generalization of Ornstein–Uhlenbeck process. Here are the key contributions: Firstly, within the Generalized O-U process, the mean-reverting term takes the form of a periodic function, constantly fluctuating. Our emphasis lies in detecting the change-point in drift parameters within the specialized framework of the Generalized O-U process sequentially, rather than focusing on changes in mean, variance, or covariance. Additionally, we introduce several detectors designed to identify the location of the change-point and provide the asymptotic properties of these detectors under both the null and al-

Dans cet exposé, nous présentons un processus stochastique qui convient aux données financières positives ayant un comportement cyclique de retour à la moyenne. Le processus stochastique proposé est une généralisation du processus d'Ornstein-Uhlenbeck, dont les principales contributions sont les suivantes : Premièrement, dans le processus O-U généralisé, le terme de retour à la moyenne prend la forme d'une fonction périodique, qui fluctue constamment. Nous mettons l'accent sur la détection séquentielle du point de changement des paramètres de dérive dans le cadre spécialisé du processus O-U généralisé, plutôt que de nous concentrer sur les changements de la moyenne, de la variance ou de la covariance. En outre, nous introduisons plusieurs détecteurs conçus pour identifier l'emplacement du point de changement et fournissons les propriétés asymptotiques de ces détecteurs à la fois

ternative hypotheses. Our work offers both theoretical advancement and practical insights into this domain.

sous l'hypothèse nulle et sous l'hypothèse alternative. Notre travail offre à la fois des avancées théoriques et des aperçus pratiques dans ce domaine.

[14:00-14:15]

Adam B. Kashlak (University of Alberta)

Asymptotic Invariance in Randomization Tests

Invariance asymptotique dans les tests de randomisation

Symmetry is a cornerstone of much of mathematics, and many probability distributions possess symmetries characterized by their invariance to a collection of group actions. Thus, many mathematical and statistical methods rely on such symmetry holding and ostensibly fail if symmetry is broken. This talk considers under what conditions a sequence of probability measures asymptotically gains such symmetry or invariance to a collection of group actions. In particular, a Lipschitz function of a high dimensional random vector will be asymptotically invariant to the actions of certain compact topological groups under some well defined conditions. Applications to classical parametric statistical tests and their randomization counterparts are considered.

La symétrie est la pierre angulaire d'une grande partie des mathématiques, et de nombreuses distributions de probabilité possèdent des symétries caractérisées par leur invariance par rapport à un ensemble d'actions de groupe. Ainsi, de nombreuses méthodes mathématiques et statistiques s'appuient sur une telle symétrie et échouent si la symétrie est brisée. Cet exposé examine dans quelles conditions une séquence de mesures de probabilité acquiert asymptotiquement une telle symétrie ou invariance par rapport à un ensemble d'actions de groupe. En particulier, une fonction Lipschitz d'un vecteur aléatoire de grande dimension sera asymptotiquement invariante aux actions de certains groupes topologiques compacts sous certaines conditions bien définies. Nous examinerons des applications aux tests statistiques paramétriques classiques et à leurs équivalents aléatoires.

[14:15-14:30]

Klaus Peter Herrmann (Université de Sherbrooke) **Johanna G. Nešlehová** (McGill University) **Marius Hofert** (The University of Hong Kong)

Transformations of Stable Tail Dependence Functions

Transformations des fonctions de dépendance de queue stable

Stable tail dependence functions play a central role in multivariate extreme value theory. Given their importance, it is natural to consider transformations from the set of stable tail dependence functions into itself. One natural candidate for such a transformation is a pre/post-composition construction, where a function is applied to each argument of the stable tail dependence function. To preserve the necessary homogeneity of stable tail dependence functions, an appropriate inverse function is applied as a post-composition. A negative result concerning such transformations is discussed by showing that only transformations based on power functions result again in bona fide stable tail dependence functions. The impact of the result is discussed and connections to the more general question of transforming generalized extreme value distributions into generalized extreme value distributions are provided.

Les fonctions de dépendance de queue stable jouent un rôle central dans la théorie des valeurs extrêmes multivariées. Compte tenu de leur importance, il est naturel d'envisager des transformations internes de l'ensemble de ces fonctions. Une construction de pré/post-composition, au cours de laquelle une fonction est appliquée à chaque argument de la fonction de dépendance de queue stable, est une solution envisageable pour ce type de transformation. Une fonction inverse adaptée de post-composition est appliquée pour préserver l'homogénéité nécessaire des fonctions de dépendance de queue stable. Nous analysons un résultat négatif de ces transformations et démontrons que seules les transformations utilisant des fonctions de puissance permettent d'obtenir de véritables fonctions de dépendance de queue stable. Nous discutons de la portée de ces résultats et nous établissons des liens avec la question plus générale de la transformation des distributions de valeurs extrêmes généralisées en distributions de valeurs extrêmes généralisées.

Chair/Président: Himchan Jeong

Room/Salle: C 3033

Date: Tuesday June 4 / mardi 4 juin

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-13:45]

Kathleen E. Miao (University of Toronto) **Silvana Pesenti** (University of Toronto)

Robustifying Elicitable Functionals under Kullback-Leibler Misspecification

Amélioration de la robustesse des fonctions élicitables à l'aide de la divergence de Kullback-Leibler pour quantifier les erreurs de spécification

Elicitable functionals and strictly consistent scoring functions are of interest due to their utility of determining unique optimal forecasts. Yet in practice, assuming that a distribution is correctly specified is too strong a belief to reliably hold. Hence, we incorporate a notion of statistical robustness into the framework of elicitable functionals. Specifically, we propose a robustified version of elicitable functionals (REFs) by using the Kullback-Leibler divergence to quantify misspecification from a baseline distribution. We show that REFs admit unique solutions at the boundary of the uncertainty region. To assess the choice of scoring functions, we propose the class of b-homogeneous strictly consistent scoring functions, for which the REFs maintain desirable statistical properties. We explore the behaviour of the REFs numerically, demonstrating the impact of the uncertainty set, and choice of scoring function using Murphy diagrams. This is joint work with Silvana M. Pesenti.

Les fonctions élicitables et les fonctions de score strictement cohérentes sont intéressantes, car elles permettent de réaliser des prévisions optimales uniques. Cependant, dans la pratique, il n'est pas toujours adéquat de présumer qu'une distribution est correctement spécifiée. C'est pourquoi nous intégrons une notion de robustesse statistique dans le contexte des fonctions élicitables. Plus précisément, nous proposons une version plus robuste des fonctions élicitables en utilisant la divergence de Kullback-Leibler pour quantifier les erreurs de spécification d'une distribution de référence. Nous démontrons que les fonctions élicitables robustes offrent des solutions uniques à la limite de la zone d'incertitude. Pour déterminer les fonctions de notation, nous proposons une classe de fonctions de notation b-homogènes parfaitement cohérentes, pour lesquelles les fonctions élicitables robustes conservent des propriétés statistiques souhaitables. Nous explorons le comportement des fonctions élicitables robustes de manière numérique, en démontrant les effets de l'ensemble d'incertitude et du choix de la fonction de notation à l'aide de diagrammes de Murphy. Ces travaux sont menés en collaboration avec Silvana M. Pesenti.

[13:45-14:00]

Sébastien Jessup (Concordia University) **Mélina Mailhot** (Concordia University) **Mathieu Pigeon** (Université du Québec à Montréal)

Robust Extreme Thresholds Through Generalised Bayesian Model Averaging

Seuils de valeurs extrêmes robustes en utilisant l'aggrégation de modèles généralisés bayésiens

Actuarial reserves are sensitive to extreme claims, where only a handful of claims can represent a significant portion of the reserves. As such, actuarial models need to take these large claims into account. One approach is to have a composite model where a distribution is fitted to losses below a certain threshold, then the

Les réserves actuarielles sont sensibles aux valeurs extrêmes, car une poignée de réclamations peuvent représenter une portion significative des réserves. Ainsi, les modèles actuariels doivent prendre ces réclamations en compte. Une approche est d'avoir des modèles composés où une distribution est ajustée aux pertes sous un certain seuil, et les pertes excédentaires sont modélisées

excess losses are modelled using extreme value theory. This implies the selection of a "best" threshold, which is an open problem with no single best approach. In this presentation, we use a generalised Bayesian model averaging method to consider multiple thresholds simultaneously, thus eliminating the need for selecting a threshold, and allowing for more weight to different thresholds based on insured characteristics.

en utilisant la théorie des valeurs extrêmes. Ceci implique la sélection d'un "meilleur" seuil, qui reste un problème ouvert pour lequel aucune approche optimale n'a été identifiée. Dans cette présentation, on utilise une méthode d'aggrégation de modèles généralisés bayésiens pour considérer plusieurs seuils simultanément, éliminant ainsi le besoin de sélection d'un seuil, et permettant d'attribuer plus de poids à des seuils différents selon les caractéristiques de l'assuré.

[14:00-14:15]

Xiyue Han (University of Waterloo) **Alexander Schied** (University of Waterloo)

Statistical Aspects of Rough Stochastic Volatility

Aspects statistiques de la volatilité stochastique rugueuse

It was observed by Gatheral, Jaisson & Rosenbaum (2018) that the Hurst parameter of realized volatility of many financial time series is rather small. This observation motivates many subsequent studies to model the volatility process with stochastic processes that are rougher than martingales. To tackle the model ambiguity, we propose to measure the roughness of a continuous function with the roughness exponent, which is defined as the number $R \in [0,1]$, where the p-th variation of the function is infinite if $p < 1/R$ and zero if $p > 1/R$. The roughness exponent can characterize the pathwise regularity without any probabilistic assumptions and coincide with the Hurst parameter for fractional Brownian motion. The problem of estimating the roughness exponent for the volatility process has been studied in Han & Schied (2023), where an estimator for the roughness exponent is established. This talk will focus on the statistical properties of our estimator, and we will demonstrate the consistency of the estimator for several classes of Gaussian processes, including the fractional Brownian motion. The convergence rate and the central limit theorem of the estimator will also be discussed. Finally, this talk will highlight the underlying rationale for constructing our estimator, which is based on the robust approximation of the Faber-Schauder coefficients in Han & Schied (2022).

Il a été observé par Gatheral, Jaisson & Rosenbaum (2018) que le paramètre de Hurst de la volatilité réalisée de nombreuses séries temporelles financières est plutôt faible. Cette observation motive de nombreuses études ultérieures à modéliser le processus de volatilité avec des processus stochastiques qui sont plus rugueux que les martingales. Pour aborder l'ambiguïté du modèle, nous proposons de mesurer la rugosité d'une fonction continue avec l'exposant de rugosité, qui est défini comme le nombre $R \in [0,1]$, où la p-th variation de la fonction est infinie si $p < 1/R$ et zéro si $p > 1/R$. Le problème de l'estimation de l'exposant de rugosité pour le processus de volatilité a été étudié dans Han & Schied (2023), où un estimateur pour l'exposant de rugosité est établi. Cette présentation se concentrera sur les propriétés statistiques de notre estimateur, et nous démontrerons la cohérence de l'estimateur pour plusieurs classes de processus gaussiens, y compris le mouvement brownien fractionnaire. Le taux de convergence et le théorème de la limite centrale de l'estimateur seront également discutés. Enfin, cette présentation mettra en évidence la logique sous-jacente à la construction de notre estimateur, qui est basée sur l'approximation robuste des coefficients de Faber-Schauder dans Han & Schied (2022).

[14:15-14:30]

Wei Liang (University of Waterloo) **Changbao Wu** (University of Waterloo)

Model-Assisted Uplift Evaluation

Évaluation du levier assistée par un modèle

Uplift modeling is a new but rapidly developing area in machine learning and causal inference. In uplift modeling, the primary goal is to learn a forecaster, known as the uplift model, to predict the effect of a treatment at

La modélisation de levier (uplift) est un domaine nouveau mais en développement rapide dans l'apprentissage automatique et l'inférence causale. Dans la modélisation du levier, l'objectif principal est d'apprendre un modèle prédictif, appelé modèle d'uplift,

the individual level so that the profits of treatment allocation can be optimized. Various uplift measures, such as the uplift curve and the area under the uplift curve have been adopted to evaluate the performance of the uplift models. Because of the inherent issue of latent potential outcomes in causal inference, however, these uplift measures are not reliable especially when the sample size of the validation dataset is small. We explore the information of covariates to improve the efficiency and reliability of the these measures. The proposed empirical measures of uplift are more precise, at least asymptotically, under distribution-free assumptions. Asymptotic theory is also established which allows us to create valid confidence intervals of the uplift measures.

pour prédire l'effet d'un traitement à un niveau individuel afin que les bénéfices de l'allocation des traitements puissent être optimisés. Diverses mesures de levier, telles que la courbe de levier et l'aire sous la courbe de levier, ont été adoptées pour évaluer les performances des modèles de levier. Toutefois, en raison du problème inhérent aux résultats potentiels latents de l'inférence causale, ces mesures de levier ne sont pas fiables, en particulier lorsque la taille de l'ensemble de données de validation est petite. Nous explorons l'utilisation de covariables pour améliorer l'efficacité et la fiabilité de ces mesures. Les mesures empiriques proposées pour le leviersont plus précises, au moins asymptotiquement, sous des hypothèses non paramétriques. Une théorie asymptotique est également établie, ce qui nous permet de créer des intervalles de confiance valides pour les mesures de levier.

[14:30-14:45]

Emma Kroell (University of Toronto) **Silvana Pesenti** (University of Toronto) **Sebastian Jaimungal** (University of Toronto)
Optimal Reinsurance in a Monotone Mean-Variance Framework

Réassurance optimale dans le cadre de moyenne-variance monotone

We study the optimal behaviour of an insurer who purchases reinsurance over a finite continuous-time horizon. The insurer seeks a time-consistent strategy to maximize a mean-variance performance criterion. Utilising the monotone mean-variance preferences of Maccheroni et al. (2009), the insurer seeks the optimal reinsurance contract by choosing the ceded loss function. Assuming a Cramer-Lundberg loss model and the expected value premium principle, we show that an excess-of-loss reinsurance contract type is optimal for the insurer. We obtain an explicit expression for the insurer's optimal contract by solving the Hamilton Jacobi Bellman Isaacs equation and illustrate the solution numerically.

Nous étudions le comportement optimal d'un assureur qui achète de la réassurance sur un horizon de temps fini et continu. L'assureur cherche une stratégie cohérente dans le temps pour maximiser le critère moyenne-variance. En utilisant les préférences moyenne-variance monotones de Maccheroni et al. (2009), l'assureur recherche le contrat de réassurance optimal en choisissant la fonction de perte cédée. En supposant le modèle classique de Cramer-Lundberg et le principe de la valeur espérée, nous montrons qu'un contrat de réassurance de type excédent de sinistre est optimal pour l'assureur. Nous obtenons une expression explicite du contrat optimal de l'assureur en résolvant l'équation de Hamilton Jacobi Bellman Isaacs et illustrons la solution numériquement.

[14:45-15:00]

Taehan Bae (University of Regina) **Tatjana Miljkovic** (Miami University - Oxford)
The Size-biased Lognormal Mixture with the Entropy Regularized Algorithm

Mélange lognormal biaisé par la taille avec l'algorithme régularisé par l'entropie

A size-biased left-truncated Lognormal (SB-ltLN) mixture is proposed as a robust alternative to the Erlang mixture for modeling left-truncated insurance losses with a heavy tail. The weak denseness property of the weighted Lognormal mixture is studied along with the tail behavior. Explicit analytical solutions are derived for moments and Tail Value at Risk based on the proposed model. An extension of the regularized expectation–maximization (REM) algorithm with Shannon's entropy weights (ewREM) is introduced for pa-

Un mélange lognormal tronqué à gauche biaisé par la taille (SB-ltLN) est proposé comme alternative robuste au mélange d'Erlang pour modéliser les pertes d'assurance tronquées à gauche avec une queue lourde. La propriété de faible densité du mélange lognormal pondéré est étudiée en même temps que le comportement de la queue. Des solutions analytiques explicites sont dérivées pour les moments et la valeur de queue à risque sur la base du modèle proposé. Une extension de l'algorithme d'espérance-maximisation régularisée (REM) avec les poids d'entropie de Shannon (ewREM) est introduite pour l'estimation des paramètres

rameter estimation and variability assessment. The left-truncated internal fraud data set from the Operational Riskdata eXchange is used to illustrate applications of the proposed model. Finally, the results of a simulation study show promising performance of the proposed SB-ltLN mixture in different simulation settings.

et l'évaluation de la variabilité. L'ensemble de données tronqué à gauche sur la fraude interne de l'échange de données sur les risques opérationnels est utilisé pour illustrer les applications du modèle proposé. Enfin, les résultats d'une étude de simulation montrent des performances prometteuses du mélange SB-ltLN proposé dans différents contextes de simulation.

Organizer/Responsable: Tessema Astatkie

Room/Salle: CSF Whale Atrium

Date: Tuesday June 4 / mardi 4 juin

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-15:00]

Jiali Wang (University of Manitoba) **Xikui Wang** (University of Manitoba)

Actuarial Study and Statistical Analysis of Wildfire Insurance Claims

Étude actuarielle et analyse statistique des réclamations d'assurance liées aux feux de forêt

The wildfire season of 2023 was one of the most devastating in Canadian history and caused significant losses to the insurance industry. This study investigates the impact of temperature, precipitation, and area burned on the insurance losses (Personal, Commercial and Auto insurance respectively) caused by wildfires. Two kinds of models are applied in our study: a fat tail regression model which may predict wildfire aggregate claims and a Generalized Linear Model (GLM) that can be integrated into the frequency-severity analysis. The anomalous North Atlantic Sea surface temperature (SST) pattern during the months before the fire may also have an impact on precipitation and become a variable for predicting the loss of wildfires. By simulating future situations, this study can provide pricing references for actuarial practices such as catastrophe excess of loss reinsurance treaties or other catastrophe insurance products that lack the use of claims history.

La période des feux de forêt de 2023 au Canada a été l'une des plus dévastatrices de son histoire et a entraîné de lourdes pertes pour le secteur de l'assurance. Cette étude porte sur l'impact de la température, des précipitations et de la zone incendiée sur les pertes d'assurance (personnelle, commerciale et automobile respectivement) causées par les feux de forêt. Deux types de modèles sont appliqués dans cette étude : un modèle de régression à queue grasse qui peut prédire les réclamations globales liées aux feux de forêt et un modèle linéaire généralisé (GLM) qui peut être intégré à une analyse fréquence-gravité. Le modèle anormal de température de surface de l'océan Nord-Atlantique (SST) dans les mois précédant les feux peut aussi avoir un impact sur les précipitations et devenir une variable pour la prédiction des pertes liées aux feux de forêt. Par simulation de situations futures, cette étude peut fournir des références de tarification pour les pratiques actuarielles, telles que la réassurance en excédent de sinistre ou autres produits d'assurance sinistre lorsque l'utilisation d'un historique des réclamations manque.

[13:30-15:00]

Roberto Primo Curti (Thompson Rivers University) **Md. Erfanul Hoque** (Thompson Rivers University) **Sean Hellingman** (Thompson Rivers University)

Exploring the Complexity of Collectible Asset Valuation and Forecasting - Insights from Magic: The Gathering

Exploration de la complexité de l'évaluation et de la prévision des actifs de collection – Intuition grâce au jeu Magic : The Gathering

In the realm of collectibles and creative investment assets, traditional models for forecasting prices often struggle with the unique dynamics presented by such niche markets. Unlike conventional financial assets, the value of these assets is influenced by a complex interplay of collectability, utility, and community-driven changes. Existing methods typically overlook the granular impact of diverse factors in a deterministic frame-

Dans le domaine des objets de collection et de l'investissement dans la création, les modèles traditionnels de prévision des prix ont du mal avec la dynamique unique que présentent de tels marchés à créneaux. Contrairement aux actifs financiers conventionnels, la valeur de ces actifs est influencée par une interaction complexe entre recouvrement, utilité et changements d'ordre communautaire. Les méthodes existantes ne prennent généralement pas en compte l'impact granulaire de divers facteurs dans un

work. Responding to these challenges, the investigation specifically investigates the influence of various predictors on price dynamics within such a framework. However, it delves into a broader set of influences from both within and outside the system. This novel method reveals the complex dynamics affecting Magic: The Gathering card prices, a card game with over 30 years of history and more than 25,000 unique cards, each impacting the game uniquely. By addressing the gaps in traditional models, the proposed framework comprehensively understands the complex effects of factors on asset valuation, offering a comprehensive framework for impact assessment on collectibles and utilitarian items. A Shiny app, complementing the research, was designed to demonstrate the process and enhance data understanding.

cadre déterministe. En réponse à ces problèmes, l'enquête examine précisément l'influence de divers prédicteurs sur la dynamique des prix dans un tel cadre. Elle s'attarde cependant à un plus grand ensemble d'influences à la fois à l'intérieur et à l'extérieur du système. Cette nouvelle méthode révèle la dynamique complexe influant sur le prix des cartes du jeu Magic : The Gathering, un jeu de cartes qui a plus de 30 ans d'histoire et qui a plus de 25 000 cartes uniques, chacune ayant un impact différent sur le déroulement du jeu. Pour parer aux lacunes des modèles traditionnels, le cadre proposé comprend la globalité des effets complexes des facteurs d'évaluation des actifs, offrant ainsi un cadre global pour l'évaluation d'impact des objets de collection et des articles utilitaires. En complément à l'enquête, une application Shiny a été conçue pour illustrer le processus et accroître la compréhension des données.

[13:30-15:00]

Negar Kalanpour (Memorial University of Newfoundland) **Armin Hatefi** (Memorial University of Newfoundland)

Shrinkage Estimators for Proportional Hazards Mixture Cure Models

Estimateurs avec rétrécissement pour des modèles de mélange pour guérison à risques proportionnel

Survival analysis is essential for modelling time-to-event data, particularly in medical research. Mixture cure models are widely used methods to model the latency and incidence components of the patients. This research focuses on the mixture model properties in the semi-parametric estimation of the Cox proportional hazard models in the presence of the collinearity problem in both latency and incidence parts, where the maximum likelihood method may lead to unreliable estimates. To address this issue, we propose shrinkage Ridge and Liu-type methods to estimate the coefficient of the underlying model. To do so, we developed new EM algorithm to incorporate the shrinkage methods for both components. Through various numerical studies, we show that the proposed shrinkage methods cope with the collinearity problem in latency and incidence components and lead to more reliable estimates in the semi-parametric setting. The developed methods are finally applied to a real data example.

L'analyse de survie est essentielle à la modélisation de données spatiotemporelles en recherche médicale. Les modèles de mélange pour guérison sont couramment employés afin de modéliser les composantes d'incidence et de latence des patients. Cette recherche se concentre sur les propriétés de modèle de mélange dans l'estimation semi-paramétrique des modèles de risque proportionnel de Cox en présence d'un problème de colinéarité dans les composantes d'incidence et de latence, où la méthode du maximum de vraisemblance peut générer des estimations peu fiables. Dans le but d'aborder ce problème, nous proposons des méthodes avec rétrécissement ridge et de type Liu pour estimer les coefficients du modèle sous-jacent. Pour ce faire, nous avons conçu un nouvel algorithme EM pour intégrer les méthodes avec rétrécissement aux deux composantes. Par l'entremise de nombreuses études numériques, nous démontrons que les méthodes avec rétrécissement proposées s'adaptent au problème de colinéarité dans les composantes d'incidence et de latence et génèrent des estimations fiables dans un cadre semi-paramétrique. Les méthodes conçues sont enfin appliquées à un exemple basé sur des données réelles.

[13:30-15:00]

Vihotogbé Edouard Houssou (Polytechnique Montreal)

Probabilistic Spatial Interpolation of Meteorological Data : Exploitation of Spatial Patterns Provided by Regional Climate Models

Interpolation spatiale probabiliste de données météorologiques : exploitation des motifs spatiaux fournis par les modèles de climat régionaux

In a context of increasing change, it is important to model and understand weather fluctuations affecting the frequency and intensity of extreme climate phenomena in order to better adapt to them. For this, high spatial resolution meteorological data are necessary to feed hydrological models and be used to conduct impact studies. Spatial interpolation methods for providing this data have certain limitations such as the biases present in the auxiliary information they use. To overcome these limitations, we propose to develop an original statistical way to use exclusively and systematically the spatial patterns present in Regional Climate Model (RCM) data to interpolate observations. Statistical methods such as Principal Component Analysis (PCA) and generalized linear regression (GLM) are combines in the approach to address this problem. The aim of this project is to apply the method in different contexts and compare the results to those obtained with existing methods.

Dans un contexte de changements croissants, il est important de modéliser et de comprendre les fluctuations météorologiques affectant la fréquence et l'intensité des phénomènes climatiques extrêmes afin de mieux s'y adapter. Pour cela, des données météorologiques à haute résolution spatiale sont nécessaires pour alimenter les modèles hydrologiques et être utilisées pour mener des études d'impact. Les méthodes d'interpolation spatiale permettant de fournir ces données présentent certaines limites telles que les biais présents dans les informations auxiliaires qu'elles utilisent. Pour surmonter ces limitations, nous proposons de développer une manière statistique originale d'utiliser exclusivement et systématiquement les modèles spatiaux présents dans les données du modèle climatique régional (MCR) pour interpoler les observations. L'objectif de ce projet est d'appliquer la méthode dans différents contextes et de comparer les résultats à ceux obtenus avec les méthodes existantes.

[13:30-15:00]

Parham Pishrobat (The University of British Columbia) **William Welch** (University of British Columbia) **Stefan Schrunner** (Norwegian University of Life Sciences)

Introducing Dynamic Kernel Regression for Enhancing Hydrological Inference

Présentation d'une régression dynamique à noyaux pour améliorer l'inférence hydrologique

This study introduces Dynamic Kernel Regression (DKR) model, a novel framework designed to enhance hydrological inference and prediction using readily collectible climate variables. Stream flow directly results from current and past rainfalls, but other climate variables like temperature impose a non-constant effect over time. The DKR model effectively accounts for lagged and accumulative effects of rainfall on streamflow and the variability introduced by temperature fluctuations, thus effectively capturing streamflow's temporal dynamics. The parameters of the DKR model represent the characteristics of different lagged kernels, where each kernel represents a distinct flow path. The DKR model provides direct interpretability on the properties of each flow path, including their location, spread, and weight over time. Results from simulation studies and real-world applications imply that accounting for temperature effect dynamically improves the model's fitness and predictive performance.

Cette étude introduit un modèle de régression dynamique à noyaux (DKR), un nouveau cadre conçu pour améliorer l'inférence et la prédiction hydrologiques en utilisant des variables de climat facilement collectables. L'écoulement fluvial résulte directement des précipitations actuelles et passées, mais d'autres variables climatiques, comme la température, imposent un effet non constant avec le temps. Le modèle DKR prend efficacement en compte les effets décalés et cumulatifs des précipitations sur l'écoulement pluvial et la variabilité qu'entraîne les fluctuations de température et, par conséquent, il saisit efficacement la dynamique temporelle de l'écoulement fluvial. Les paramètres du modèle DKR représentent les caractéristiques de différents noyaux décalés, chaque noyau représentant un chemin d'écoulement distinct. De plus, ce modèle fournit une interprétabilité directe des propriétés de chaque chemin d'écoulement, y compris leur emplacement, leur étendue et leur poids avec le temps. Des résultats d'études en simulation et des applications du monde réel laissent entendre que la prise en compte de l'effet de la température améliore de façon dynamique l'ajustement et la performance prédictive du modèle.

[13:30-15:00]

Andrew Putman (Ontario Tech University) **Shilpa Dogra** (Ontario Tech University)

Initial Validity and Reliability Testing of the SGBA-5: A Measurement Tool for Facilitating Sex- And Gender-Based Analyses in Health Sciences Research

Tests initiaux de validité et de fiabilité du SGBA-5 : Un outil de mesure pour faciliter les analyses fondées sur le sexe et le genre

dans la recherche en sciences de la santé

Background: A hurdle to incorporating Sex- and Gender-Based Analysis is a lack of easily implemented measurement tools. To address this, we created the Sex- and Gender-Based Analysis Tool – 5 item [SGBA-5]. Objectives: Assess the validity and reliability of the SGBA-5 for health research where sex or gender are not primary variables. Methods: A Delphi consensus study was conducted with Canadian researchers [n=14]. A 2-arm [students, n=89; older adults, n = 71] test-retest study was then conducted. Results: Agreement was reached for the sex item [93%] and consensus non-agreement for gendered aspect of health items [identity: 64%, expression: 64%, roles: 50%, relations: 57%]. The test-retest study found all items reliable on both arms [sex: $\kappa = 1.00$, gendered: ICC(A,1) ¿ .850]. Conclusion: The novel SGBA-5 demonstrated reliability for all items and validity of the sex item; the gendered aspects of health items may be valid. Future research will further assess the SGBA-5.

Contexte : L'absence d'outils de mesure faciles constitue un obstacle à l'intégration de l'analyse comparative fondée sur le sexe et le genre. C'est pourquoi nous avons créé le Sex- and Gender-Based Analysis Tool – 5 item [SGBA-5]. Objectif : Évaluer la validité et la fiabilité du SGBA-5 pour la recherche lorsque le sexe ou le genre ne sont pas des variables primaires. Méthodes : Une étude de Delphi avec des chercheurs canadiens [n=14] et une étude de testage-retestage à deux bras [étudiants, n=89 ; adultes âgés, n=71] a été menée. Résultats : On est parvenu à un accord sur la question de sexe [93%] et l'absence d'accord sur les questions de genre [identité : 64%, expression : 64%, rôles : 50%, relations : 57%]. L'étude test-retest a trouvé toutes les questions fiables [sexe : $\kappa = 1$, genre : ICC(A,1)¿0.850]. Conclusion : Le SGBA-5 a démontré sa fiabilité et sa validité de la question sur le sexe ; les questions de genre peuvent être valides. Des recherches futures devront évaluer davantage le SGBA-5.

Modelling Rainfall Extremes
Modélisation des précipitations extrêmes

Chair/Président: Michaël Lalancette

Organizer/Responsable: Léo Belzile

Room/Salle: A 1046

Date: Tuesday June 4 / mardi 4 juin

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-16:00]

Debbie J. Dupuis (HEC Montréal) **Luca Trapin** (University of Bologna)

Mixed-frequency Extreme Value Regression: Estimating the Effect of Mesoscale Convective Systems on Extreme Rainfall Intensity

Régression à fréquences mixtes pour valeurs extrêmes : estimation de l'effet des complexes convectifs de méso-échelle sur l'intensité des précipitations extrêmes

Understanding and modeling the determinants of extreme hourly rainfall intensity is of utmost importance for the management of flash-flood risk. Increasing evidence shows that mesoscale convective systems (MCS) are the principal driver of extreme rainfall intensity in the United States. We use extreme value statistics to investigate the relationship between MCS activity and extreme hourly rainfall intensity in Greater St. Louis, an area particularly vulnerable to flash floods. Using a block maxima approach with monthly blocks, we find that the impact of MCS activity on monthly maxima is not homogeneous within the month/block. To appropriately capture this relationship, we develop a mixed-frequency extreme value regression framework accommodating a covariate sampled at a frequency higher than that of the extreme observation.

Comprendre et modéliser les déterminants de l'intensité des précipitations horaires extrêmes est très importante pour la gestion du risque de crue éclair. Les données montrent que les complexes convectifs de méso-échelle (MCS) sont le principal moteur de l'intensité extrême des précipitations aux États-Unis. Nous utilisons des statistiques de valeurs extrêmes pour étudier la relation entre l'activité MCS et l'intensité des précipitations horaires extrêmes dans la grande région de Saint-Louis, une zone particulièrement vulnérable aux crues éclair. En utilisant une approche de maximum de blocs avec des blocs mensuels, nous constatons que l'impact de l'activité MCS sur les maximums n'est pas homogène au sein du mois/bloc. Pour capturer de manière appropriée cette relation, nous développons un cadre de régression à fréquences mixtes pour valeurs extrêmes prenant en compte une variable explicative échantillonnée à une fréquence supérieure à celle de l'observation extrême.

[16:00-16:30]

Mélina Mailhot (Concordia University) **Mathieu Pigeon** (Université du Québec à Montréal) **Sébastien Jessup** (Concordia University)

Combination Methods on Extreme and Skewed Data

Méthodes de combinaison de données extrêmes et asymétriques

Focussing on extreme and skewed data, we investigate different methods, from non-parametric to Bayesian approaches, and identify assumptions under which some combination techniques are performing better. In this presentation, we apply multiple model combination methods to an ensemble of experts in a pooling approach and use the differences in outputs from the different

Nous étudions différentes méthodes, des approches non paramétriques aux approches bayésiennes, en mettant l'accent sur les données extrêmes et asymétriques, puis nous identifions les hypothèses dans lesquelles certaines techniques de combinaison sont plus efficaces. Dans cette présentation, nous mettons en œuvre plusieurs méthodes de combinaison de modèles pour un ensemble d'experts dans une approche de mise en commun. Ensuite, nous

combinations to illustrate how one can gain additional insight from using multiple methods. An illustration with extreme precipitation will be presented. Then, we present a generalized method, in order to fit skewed errors, and apply this new algorithm to a simulated insurance dataset.

utilisons les différences entre les résultats des différentes combinaisons pour illustrer comment on peut obtenir d'autres données supplémentaires en utilisant plusieurs méthodes. Nous illustrons ces méthodes à l'aide d'un exemple concernant les précipitations extrêmes. Enfin, nous présentons une méthode généralisée pour prendre en compte les erreurs asymétriques, puis nous appliquons ce nouvel algorithme à un ensemble de données d'assurance simulées.

[16:30-17:00]

Léo Belzile (HEC Montréal) **Rishikesh Yadav** (HEC Montréal)

Can Climate Model Output Adequately Represent Extreme Rainfall?

Est-ce que les précipitations extrêmes de modèles climatiques sont fidèles à la réalité ?

Global circulation models (GCM) provide state-of-the-art information about the behaviour of the climate under different emission scenarios via catalogues of simulations, yet they are not calibrated for extreme events. We consider the latter by filtering exceedances at a single location using the conditional spatial extremes model to model left-censored zero-inflated precipitation data. We use Gaussian Markov random field residual processes to ensure the extreme model can be fitted to a large number of measurement sites and consider estimation of both the margins and the dependence structure parameters under the Bayesian paradigm, using MCMC methods through data augmentation for estimation. Using the joint model, we compare extremal properties of rainfall fields of statistically downscaled GCM output from the Pacific Climate Impact Consortium with measurements from station data in British Columbia and Washington, finding important discrepancies between the two.

Les modèles de circulation générale (MCG) permettent de représenter le comportement du climat sous différents scénarios d'émission par le biais de catalogues de simulations. Nous considérons la calibration des extrêmes de champs de précipitations des sorties climatiques en filtrant les données pour lesquelles un excès de seuil survient à un site. Nous modélisons ces dernières à l'aide du modèle d'extrêmes spatiaux conditionnels, en traitant l'absence de précipitation comme de la censure à gauche et en utilisant un processus résiduel Gaussien markovien pour faciliter l'ajustement à un grande nombre de sites. Les paramètres des marges et du modèle conjoint sont estimés par MCMC. À l'aide du modèle conjoint, nous comparons les propriétés extrémales des champs de précipitations des sorties désagrégées fournies par le PCIC avec des données de stations météos en Colombie-Britannique et dans l'État de Washington. D'importantes divergences entre les deux sources sont constatées.

The Bayesian Edge: Novel Applications of Bayesian Methods to Clinical Research, Indirect Treatment Comparison, and Public Health.
L'avantage bayésien : nouvelles applications des méthodes bayésiennes à la recherche clinique, à la comparaison indirecte des traitements et à la santé publique

Chair/Président: Audrey Béliveau

Organizer/Responsable: Aaron Springford

Room/Salle: C 2045

Date: Tuesday June 4 / mardi 4 juin

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-16:00]

Linke Li (University of Toronto Dalla Lana School of Public Health)

Efficient Computation Methods for Expected Value of Sample Information in Bayesian Clinical Trial Designs

Méthodes de calcul efficaces de la valeur attendue de l'information d'échantillonnage (EVSI) dans des essais cliniques bayésiens

Due to the inherent conflict between numerous viable research projects and limited funding, decision-makers must prioritize research questions, with a key strategy being the identification of studies that yield the greatest economic benefit. The Expected Value of Sample Information (EVSI) is a Bayesian decision metric capable of quantifying the value of a clinical trial design from a health economic perspective. However, conventional EVSI estimation methods are computationally demanding. To address this, we developed a computational approach that efficiently and accurately estimates EVSI for fixed-sample-size trial designs using a Gaussian approximation-based method. Additionally, we introduce two innovative computational methods that employ machine learning to evaluate EVSI effectively for adaptive trial designs. These methods can streamline the trial design process, enhance research efficiency, and maximize the societal economic benefits of research.

En raison du conflit inhérent entre la multitude de projets de recherche viables et le financement limité, les décideurs doivent placer en ordre de priorité les sujets de recherche, et une des stratégies principales est d'identifier les études qui auront le plus grand avantage économique. La valeur attendue de l'information d'échantillonnage (EVSI) est une mesure de décision bayésienne capable de quantifier la valeur d'une conception d'essai clinique du point de vue économique de la santé. Les méthodes d'estimation EVSI conventionnelles sont cependant très exigentes en termes de calcul. Pour y remédier, nous avons développé une approche computationnelle qui estime avec efficacité et exactitude la valeur attendue de l'information d'échantillonnage pour des conceptions d'essai avec taille d'échantillon fixe en utilisant une méthode basée sur une approximation gaussienne. Nous présentons en outre deux méthodes de calcul innovantes qui emploient l'apprentissage machine pour évaluer efficacement la EVSI pour des conceptions d'essai clinique adaptatif. Ces méthodes peuvent rationaliser le processus de conception des essais, accroître l'efficacité de la recherche et maximiser les avantages économiques de la recherche pour la société.

[16:00-16:30]

Yiran Wang (University of Waterloo) **Martin Lysy** (University of Waterloo) **Audrey Béliveau** (University of Waterloo)

Plant-Capture Methods for Estimating Population Size from Uncertain Plant Captures

Méthodes de capture de plantes pour estimer la taille de population provenant de captures incertaines de plantes

Plant-capture is a variant of classical capture-recapture methods, which is used to estimate the size of a population. In this method, decoys referred to as "plants" are introduced into the population in order to estimate the capture probability. The method has shown consider-

La capture de plante est une variante de la méthode classique de « capture-recapture », utilisée pour estimer la taille d'une population. Cette méthode comprend des leurres nommés « plantes » qui sont insérés dans la population afin d'estimer la probabilité de capture. Elle a réussi considérablement à estimer les

able success in estimating population sizes from limited samples in many epidemiological, ecological, and demographic studies. However, previous plant-recapture studies have not systematically accounted for uncertainty in the capture of each individual plant. In this work, we propose various approaches to formally incorporate the uncertainty arising from the capture status of plants and the heterogeneity between multiple survey sites into the plant-capture model. We further introduce two inference methods, compare their performance in simulation studies and apply our methods to analyze real data from the "S-night" study conducted by the US Census Bureau.

tailles de population à partir d'échantillons limités dans plusieurs études épidémiologiques, écologiques et démographiques. Cependant, les études « plante-recapture » antérieures n'ont pas systématiquement tenu compte de l'incertitude relative à la capture de chaque plante. Dans le cadre de ce travail, nous proposons plusieurs approches pour formellement incorporer l'incertitude provenant de l'état de capture des plantes et de l'hétérogénéité entre plusieurs sites d'enquête dans le modèle de plante-capture. Nous présentons d'ailleurs deux méthodes d'inférence, comparons leur performance dans des études en simulation et appliquons notre méthode afin d'analyser des données réelles tirées de l'étude « S-night » menée par le Bureau de recensement des États-Unis.

[16:30-17:00]

Emma K Mackay (Inka Health)

Bayesian borrowing approaches to address the challenges of evaluating efficacy/effectiveness in rare indications: applications to basket trials and pediatric studies

Approches d'emprunt bayésiennes pour relever les défis de l'évaluation de l'efficacité dans les indications rares : applications aux essais panier et études pédiatriques

With the development of new therapies for rare diseases, including those targeting increasingly rare cancer mutations, evaluating efficacy in a clinical trial setting can be a challenge due to the difficulty of recruiting enough patients. This has led to novel trial designs which reduce control group allocation by borrowing from historical trial control arms, use of information borrowing from adult trial populations to supplement pediatric trials, and uptake of basket trials which recruit patients with multiple disease types which share a common mutation or biomarker that is targeted by the therapy. This talk will outline the challenges of rare disease settings, and discuss several recent applications of Bayesian hierarchical models, power priors, and meta-analytic predictive priors to facilitate partial information borrowing–or "dynamic borrowing"–to improve precision/power when evaluating efficacy and conducting indirect treatment comparisons.

Avec le développement de nouvelles thérapies pour les maladies rares, y compris celles ciblant des mutations cancéreuses de plus en plus rares, l'évaluation de l'efficacité dans les essais cliniques peut poser des problèmes en raison de la difficulté à recruter suffisamment de patients. Cela a conduit à de nouvelles conceptions d'essais qui réduisent l'allocation des groupes de contrôle en empruntant des bras de contrôle d'essais historiques, à l'utilisation d'informations empruntées à des populations d'essais adultes pour compléter les essais pédiatriques, et à l'adoption d'essais panier qui recrutent des patients atteints de plusieurs types de maladies et qui partagent une mutation ou un biomarqueur commun qui est ciblé par la thérapie. Cet exposé présentera les défis posés par les maladies rares et discutera de plusieurs applications récentes de modèles hiérarchiques bayésiens, d'a priori de puissance et d'a priori prédictifs méta-analytiques pour faciliter l'emprunt partiel d'informations - ou « emprunt dynamique » - afin d'améliorer la précision/puissance lors de l'évaluation de l'efficacité et de la réalisation de comparaisons indirectes de traitements.

Machine Learning in Causal Inference: Modern Health Research Paradigms
Apprentissage automatique en inférence causale : paradigmes modernes de la recherche en santé

Chair/Président: Mohammad Ehsanul Karim

Organizer/Responsable: Mohammad Ehsanul Karim

Room/Salle: ED 2018A

Date: Tuesday June 4 / mardi 4 juin

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-16:00]

Lan Wen (University of Waterloo)

Estimating the Average Causal Effects of Dietary Substitution Strategies

Estimation des effets causaux moyens des stratégies de substitution alimentaire

According to the 2020-2025 Dietary Guidelines, achieving a healthy dietary pattern for most people will require changes in food and drink choices, with some of these changes involving simple substitutions. For instance, the Dietary Guidelines suggest opting for chicken for processed red meat lower sodium intake, and/or switching from refined grains to whole grains to boost dietary fiber intake. The question about food substitution (e.g., replacing processed red meat with chicken) seeks to estimate the counterfactual outcome under a hypothetical strategy that depends on the natural value of treatment in the factual world. We will show conditions under which the average causal effects of substitution strategies can be identified. We provide efficient estimators for our proposed dietary substitution strategy that can be used in conjunction with machine learning methods, and demonstrate the methodology via simulation studies and an application utilizing data from the Nurses' Health Study.

Selon les Directives alimentaires 2020-2025, atteindre un modèle alimentaire sain pour la plupart des gens nécessitera des changements dans les choix alimentaires et de boissons, certains de ces changements impliquant de simples substitutions. Par exemple, les Directives alimentaires recommandent de préférer le poulet à la viande rouge transformée pour réduire l'apport en sodium, et/ou de passer des céréales raffinées aux céréales complètes pour augmenter l'apport en fibres alimentaires. La question de la substitution alimentaire (par exemple, remplacer la viande rouge transformée par du poulet) cherche à estimer le résultat contre-factuel sous une stratégie hypothétique qui dépend de la valeur naturelle du traitement dans le monde factuel. Nous montrerons les conditions sous lesquelles les effets causaux moyens des stratégies de substitution peuvent être identifiés. Nous fournissons des estimateurs efficaces pour notre stratégie de substitution alimentaire proposée qui peuvent être utilisés conjointement avec des méthodes d'apprentissage automatique, et démontrons la méthodologie via des études de simulation et une application utilisant des données de l'étude de santé des infirmières.

[16:00-16:30]

Robert W. Platt (McGill University) **Rubiya Akter** (McGill University) **Enrico Ripamonti** (University of Milan-Bicocca)

Lookback Periods in Observational Epidemiology: Statistical Considerations

Périodes rétrospectives en épidémiologie par observation : considérations statistiques

In pharmacoepidemiology and other observational research, determining how far before the index date to assess confounders (the appropriate lookback period) is crucial. This talk discusses using propensity score matching and targeted maximum likelihood estimation (TMLE) to evaluate the lookback period for confounding control. There is a tradeoff inherent in varying the

En pharmaco-épidémiologie et dans d'autres recherches par observation, il est crucial de déterminer combien de temps avant la date d'indexation il faut évaluer les facteurs de confusion (la période rétrospective appropriée). Cet exposé traite de l'utilisation de l'appariement des scores de propension et de l'estimation ciblée par maximum de vraisemblance (TMLE) pour évaluer la période rétrospective pour le contrôle des facteurs de confusion. Il existe

lookback; the longer the lookback period the higher the likelihood of missing data. The discussion includes statistical considerations in choosing the lookback period, especially in situations when important confounding variables may be measured long before the index date. We will demonstrate the application of the two proposed approaches through simulations and a worked example. We show that in general, longer lookbacks are useful, but that in practice the gains may be small.

un compromis inhérent à la variation de la période rétrospective ; plus celle-ci est longue, plus la probabilité de données manquantes est élevée. La discussion inclut des considérations statistiques dans le choix de la période rétrospective, en particulier dans les situations où d'importantes variables confusionnelles peuvent être mesurées bien avant la date d'indexation. Nous démontrerons l'application des deux approches proposées à l'aide de simulations et d'un exemple concret. Nous montrons qu'en général, les périodes rétrospectives plus longues sont utiles, mais qu'en pratique, les gains peuvent être faibles.

[16:30-17:00]

Mireille Schnitzer (Université de Montréal) **Cong Jiang** (Université de Montréal) **Miceline Mésidor** (INRS-Institut Armand-Frappier) **Yan Liu** (Université de Montréal) **Edgar Ortiz Brizuela** (McGill University) **Mabel Carabali** (McGill University) **Denis Talbot** (Université Laval)

Methods for the test-negative design: application and analysis of vaccine effectiveness during the pandemic and new approaches

Méthodes pour le devis test négatif : l'application et l'analyse de l'efficacité de la vaccination pendant la pandémie et nouvelles approches

The test-negative design has been hailed for its rapid implementation with electronic health data that allows for almost live estimates of vaccine effectiveness, which proved very useful during the COVID-19 pandemic. Its validated application involves using logistic regression to analyze the data of symptomatic individuals who receive a test for an infection of interest. Under assumptions, this produces estimates of vaccine effectiveness against disease. However, certain applications of the design have not been theoretically validated and might lead to biased results. We first provide results of a systematic review of TND methods in peer-reviewed articles published between January 1, 2020, and January 25, 2022, during the COVID-19 pandemic. Using DAGs and data simulations, we demonstrate the bias that may arise from a popular version of the design that includes people without symptoms. We then present a novel non-parametric method for estimation of vaccine effectiveness against disease.

Le devis d'étude « test négatif » a été salué pour sa mise en œuvre rapide avec des données de santé électroniques qui permettent d'obtenir des estimations presque en direct de l'efficacité du vaccin, ce qui s'est avéré très utile pendant la pandémie de COVID-19. Son utilisation validée consiste à l'application de la régression logistique pour analyser les données d'individus symptomatiques qui reçoivent un test pour l'infection d'intérêt. Sous réserve d'hypothèses, cela produit des estimations de l'efficacité du vaccin contre la maladie. Cependant, certaines applications de ce modèle n'ont pas été validées théoriquement et pourraient conduire à des résultats biaisés. Nous fournissons d'abord les résultats d'une revue systématique des méthodes TND dans des articles évalués par des pairs publiés entre le 1er janvier 2020 et le 25 janvier 2022 pendant la pandémie de COVID-19. À l'aide des DAGs et des simulations de données, nous démontrons le biais pouvant découler d'une version populaire du modèle qui inclue des personnes asymptomatiques. Nous présentons ensuite une nouvelle méthode non paramétrique pour estimer l'efficacité d'un vaccin contre la maladie.

Chair/Président: Yanglei Song

Organizer/Responsable: Yanglei Song

Room/Salle: A 1045

Date: Tuesday June 4 / mardi 4 juin

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-16:00]

Yajun Mei (University)

Active Learning in Sequential Analysis and Change-Point Detection

Apprentissage actif en analyse séquentielle et détection de changement de régime

Sequential or streaming data occur in many real-world problems ranging from clinical trials and biosurveillance to environmental monitoring and network security to finance and economics and so on. Often it is challenging to develop nearly optimal procedures or algorithms for sequential or streaming data, partly due to uncertainties on model parameters and limited data available to make rapid decisions. In this talk, we present our latest research that develop efficient sequential analysis and change-point detection algorithms via active learning of unknown parameters. Two specific examples will be discussed: one is futility stopping boundaries of the adaptive group sequential trials, and the other is efficient change-point detection algorithms for monitoring high-dimensional linear regression models via implicit regularization. Numerical simulations and case studies will be presented to demonstrate the usefulness of our proposed algorithms.

Les données séquentielles ou en continu sont présentes dans de nombreux problèmes du monde réel, des essais cliniques et de la biosurveillance à la surveillance de l'environnement et à la sécurité des réseaux, en passant par la finance et l'économie, etc. Il est souvent difficile de développer des procédures ou des algorithmes presque optimaux pour les données séquentielles ou en continu, en partie à cause des incertitudes sur les paramètres du modèle et du peu de données disponibles pour prendre des décisions rapides. Dans cet exposé, nous présentons nos dernières recherches qui développent des algorithmes efficaces d'analyse séquentielle et de détection de changement de régime via l'apprentissage actif de paramètres inconnus. Nous discuterons de deux exemples spécifiques : l'un concerne les limites d'arrêt de futilité des essais séquentiels de groupe adaptatifs, et l'autre concerne des algorithmes efficaces de détection de changement de régime pour le suivi de modèles de régression linéaire en grande dimension via une régularisation implicite. Nous présenterons des simulations numériques et des études de cas pour démontrer l'utilité des algorithmes proposés.

[16:00-16:30]

Jay Bartroff (University of Texas at Austin)

Group Sequential Testing of a Treatment Effect Using a Surrogate Marker

Test séquentiel de groupe d'un effet de traitement à l'aide d'un marqueur de substitution

The identification of surrogate markers is motivated by their potential to make accurate decisions sooner about a treatment effect. Many existing methods combine surrogate marker and primary outcome information, rely on parametric methods with unrealistic assumptions, or utilize the surrogate marker at only a single time point. I will talk about recent work using a nonparametric test

Il est intéressant de pouvoir identifier des marqueurs de substitution car ceux-ci permettent de prendre des décisions précises plus rapidement sur l'effet d'un traitement. De nombreuses méthodes existantes combinent les informations sur les marqueurs de substitution et les résultats primaires, s'appuient sur des méthodes paramétriques avec des hypothèses irréalistes ou n'utilisent le marqueur de substitution qu'à un seul point dans le temps. Je vous par-

of the treatment effect with group sequential monitoring of the surrogate marker measured repeatedly over time. The main statistical challenge is deriving the properties of the correlated surrogate-based nonparametric test statistic at multiple time points and computing stopping boundaries that allow for early stopping for a significant treatment effect, or for futility. The performance of our procedure is shown through simulation studies and on data from two different AIDS clinical trials. This is joint work with Layla Parast.

lerai de travaux récents utilisant un test non paramétrique de l'effet du traitement avec un suivi séquentiel de groupe du marqueur de substitution mesuré de manière répétée dans le temps. Le principal défi statistique consiste à dériver les propriétés de la statistique du test non paramétrique basé sur le marqueur de substitution corrélé à plusieurs points temporels et à calculer les limites d'arrêt qui permettent un arrêt précoce en cas d'effet significatif du traitement ou de futilité. Je démontrerai la performance de cette procédure par des études de simulation et sur des données provenant de deux essais cliniques différents sur le Sida. Ce travail a été réalisé en collaboration avec Layla Parast.

[16:30-17:00]

Georgios Fellouris (University of Illinois, Urbana-Champaign) **Yiming Xing** (University of Illinois, Urbana-Champaign)

Centralized and Asynchronous Sequential Multiple Testing

Essais séquentiels multiples centralisés et asynchrones

We will consider the sequential testing of the distributions of multiple data streams, and we will propose an asynchronous and centralized formulation for this problem. According to it, the decisions for the various testing problems can be made at different times, and it is possible to utilize data from all sources to decide when to stop sampling in each stream and which hypothesis to select for the corresponding testing problem. We will introduce a novel testing procedure that controls the generalized familywise error rates of both kinds and minimizes asymptotically the expected total sample size, simultaneously under every configuration, as both target error rates go to zero. The proposed approach will be compared with existing ones, synchronous and decentralized, and it will be illustrated with simulation studies.

Nous analysons les essais séquentiels des distributions de flux de données multiples et nous proposons une formulation asynchrone et centralisée pour ce problème. Selon cette formulation, il est possible de prendre les décisions relatives aux différents problèmes d'essais à des moments différents et d'utiliser des données provenant de toutes les sources pour décider quand arrêter l'échantillonnage dans chaque flux et quelle hypothèse sélectionner pour le problème d'essais correspondant. Nous présentons une nouvelle procédure d'essai qui permet de contrôler les taux d'erreurs des types I et II globaux généralisés et de minimiser asymptotiquement la taille totale prévue de l'échantillon, simultanément dans toutes les configurations, lorsque les deux taux d'erreurs cibles deviennent nuls. Nous comparerons notre approche avec les approches actuelles, synchrones et décentralisées, puis nous l'illustrerons à l'aide d'études de simulation.

Inference Methods in Stochastic Processes with Change-points: Recent Advances
Méthodes d'inférence dans les processus stochastiques avec les points de changement : avancées récentes

Chair/Président: Sévérien Nkurunziza

Organizer/Responsable: Sévérien Nkurunziza

Room/Salle: A 1043

Date: Tuesday June 4 / mardi 4 juin

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-16:00]

Zhou Zhou (University of Toronto) **Weichi Wu** (Tsinghua University) **David Veitch** (University of Toronto)

Asynchronous Jump Testing and Estimation in High Dimensions Under Complex Temporal Dynamics

Test de saut asynchrone et estimation en grande dimension selon des dynamiques temporelles complexes

Most high dimensional changepoint detection methods rely on the assumptions that the error process is stationary and changepoints occur synchronously across dimensions. The violation of these assumptions, which in applied settings is increasingly likely as the dimensionality of the time series being analyzed grows, can dramatically curtail the power or the accuracy of these methods. We propose MJPD-HD, a high dimensional multiscale jump detection method which is able to detect jumps in an otherwise smoothly varying mean function for high dimensional time series with nonstationary noise where the jumps across dimensions may not occur at the same time. A multiplier bootstrap procedure is proposed to estimate the critical values of our max-type test statistic with a controlled Type I error rate. An application to seismic and financial time series demonstrates MJPD-HD's ability to accurately detect jumps in real-world high dimensional time series with complex temporal dynamics.

La plupart des méthodes de détection de point de rupture dimensionnel dépendent de l'hypothèse selon laquelle le processus d'erreur est stationnaire et les points de rupture surviennent de manière synchrone à travers les dimensions. Lorsque cette hypothèse se révèle fausse, ce qui est de plus en plus probable en pratique au fur et à mesure que la dimension des séries temporelles analysées croît, la puissance et la précision de ces méthodes s'avèrent fortement réduites. Nous proposons MJPD-HD, une méthode de détection de saut multiéchelle de grande dimension qui peut détecter les sauts dans une fonction moyenne à variation lisse pour des séries temporelles de grande dimension avec bruit stationnaire où les sauts à travers les dimensions peuvent ne pas survenir en même temps. Nous suggérons une procédure par bootstrap multiplicative dans le but d'estimer les valeurs critiques de nos statistiques de test de type max avec un taux d'erreur contrôlé de type I. Son application sur des séries temporelles financières et sismiques démontre la capacité de MJPD-HD de détecter les sauts avec précision dans des séries temporelles réelles de grande dimension avec des dynamiques temporelles complexes.

[16:00-16:30]

Mai Ghannam (University of Ottawa) **Sévérien Nkurunziza** (University of Windsor)

Estimation and Inference in a Tensor Regression Model with Change-Points

Estimation et inférence dans le modèle de régression tensoriel avec points de rupture

In this talk, we consider an estimation problem in a tensor regression model with multiple change-points. Under a dependence structure of the error and covariates that is as weak as an L2-mixingale array, we establish the asymptotic properties of the unrestricted tensor estimator (UE) and restricted tensor estimator (RE). Moreover, we propose a class of shrinkage estimators (SEs)

Lors de cet exposé, nous examinons un problème d'estimation dans un modèle de régression tensoriel avec plusieurs points de rupture. Selon une structure de dépendance de l'erreur et des covariables qui est faible en tant que réseau L2-mixingale, nous établissons les propriétés asymptotiques de l'estimateur tensoriel non restreint (UE) et de celui restreint (RE). En outre, nous proposons une classe d'estimateurs de rétrécissement (SES) dans le

in the case of tensor regression, and we derive sufficient conditions for the SEs to dominate the UE. In addition, we consider an inference problem in the model for the special case of a possible change-point. Specifically, we consider a general hypothesis testing problem on a tensor parameter and the studied testing problem includes as a special case testing the absence of a change-point. To this end, we derive a test for testing the restriction and its asymptotic power and we prove that the proposed test is consistent. Finally, we present some simulation and real data results that corroborate the theoretical results.

cas d'une régression tensorielle, et nous dérivons suffisamment de conditions pour que le SES surpasse l'UE. De plus, nous abordons un problème d'inférence dans le modèle pour le cas particulier d'un point de rupture possible. Spécifiquement, nous évaluons un problème de test d'hypothèse général sur un paramètre tensoriel et le problème de test étudié comprend en guise d'exception un test avec absence de point de rupture. Pour ce faire, nous dérivons un test pour tester la restriction et sa puissance asymptotique et démontrons que le test proposé est constant. En conclusion, nous présentons certains résultats de données réelles et simulées qui confirment les résultats théoriques.

[16:30-17:00]

Rogemar S. Mamon (The University of Western Ontario) **Fuqi Chen** (Health Canada)

Determination of Multiple Change Points in a Multi Dimensional Mean-reverting Process

Détermination de points de changement multiples dans un processus de retour à la moyenne multidimensionnel

The accurate estimation of multiple structural-change locations is a fundamental consideration in the natural sciences and engineering. We determine the unknown change points in a multivariate Ornstein-Uhlenbeck process, which is typically used to model a system that has the tendency to revert back to some stationary mean. Two approaches are put forward, namely, the maximum-likelihood-based technique and the sum-of-squared-error method. For each approach, certain asymptotic properties of the location estimators are established. When the number of change points is unknown, our methodology assumes that the percentage change difference in the likelihood values also increases as the number of estimated change points increases. Numerical implementation is included involving simulated and actual financial data on exchange rates and commodity futures prices.

L'estimation précise de multiples points de changement structurel est une considération fondamentale en sciences naturelles et ingénierie. Nous déterminons les points de changement inconnus dans un processus multivarié d'Ornstein-Uhlenbeck, processus généralement utilisé pour modéliser un système qui a tendance à revenir à une certaine moyenne stationnaire. Nous proposons deux approches, à savoir la technique basée sur le maximum de vraisemblance et la méthode de la somme des erreurs quadratiques. Pour chaque approche, nous établissons certaines propriétés asymptotiques des estimateurs de localisation. Lorsque le nombre de points de changement est inconnu, notre méthodologie suppose que la différence de changement en pourcentage dans les valeurs de vraisemblance augmente également à mesure que le nombre de points de changement estimés augmente. Nous incluons une application numérique à des données financières simulées et réelles sur les taux de change et les prix à terme des matières premières.

Publishing for Early Career Researchers (Panel)
Publication pour les chercheurs en début de carrière (Table Ronde)

Chair/Président: James H. McVittie

Organizer/Responsable: Johanna G. Nešlehová

Room/Salle: SN 2109

Date: Tuesday June 4 / mardi 4 juin

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-17:00]

Hugh Chipman (Acadia University) **Josée Dupuis** (McGill University) **Richard A. Lockhart** (Simon Fraser University) **Grace Y. Yi** (University of Western Ontario)

Panel on Publishing for Early Career Researchers

Table ronde sur la publication pour les chercheurs en début de carrière

This panel aims to help early career researchers navigate the scientific publishing landscape: as authors, reviewers, and eventually as associate editors. Four panelists with extensive editorial experience will share their views on topics such as: good practices in article preparation and the role of a cover letter, challenges of publishing multidisciplinary collaborative articles, the choice of a suitable publishing venue, pros and cons of double-blind refereeing, how to handle negative feedback and survive harsh referee reports, understanding the difference between an invited revision and a resubmission, good practices of refereeing, the role of an associate editor and when to consider becoming one. Open access publishing options and techniques for identifying predatory journals will also be discussed. The panel will close with a short presentation of the specifics of The Canadian Journal of Statistics by its current Editor-in-Chief.

Cette table ronde a pour but d'aider les chercheurs en début de carrière à explorer le monde de l'édition scientifique, à titre d'auteurs, de réviseurs et éventuellement d'éditeurs associés. Fort d'une longue expérience de l'édition, quatre panélistes partageront leurs points de vue sur des sujets, comme les bonnes pratiques de préparation d'un article et le rôle d'une lettre de présentation, les difficultés de publier des articles en collaboration multidisciplinaire, le choix d'une revue appropriée, les aspects positifs et négatifs de l'arbitrage en double aveugle, les moyens de traiter les commentaires négatifs et de survivre aux rapports d'arbitrage sévères, la compréhension de la différence entre une révision par invitation et une resoumission, les pratiques exemplaires d'arbitrage, le rôle de l'éditeur associé et le moment opportun pour songer à le devenir. Nous aborderons aussi les options et techniques de publication en libre accès pour identifier les revues prédatrices. L'actuelle rédactrice en chef de la Revue canadienne de statistique conclura par une courte présentation des caractéristiques de cette publication.

Chair/Président: Zeinab Mashreghi

Room/Salle: A 2071

Date: Tuesday June 4 / mardi 4 juin

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-15:45]

Nahid Sadr (Université de Sherbrooke) **Marius Hofert** (The University of Hong Kong) **Klaus Peter Herrmann** (Université de Sherbrooke)

Index-mixed Copulas: A New Class of Multivariate Copulas

Les copules mixtes d'indices : une nouvelle classe de copules multivariées

Copulas hold significant importance in the context of multivariate statistics and risk management, as they model the relationship between variables irrespective of their marginal distributions. In this talk, we aim to introduce a new class of multivariate copulas named index-mixed copulas, that are in a sense a generalization of copula mixtures, and show a remarkable degree of analytical tractability. Properties investigated include measures of concordance such as Spearman's rho, Kendall's tau, and concordance orderings. As the construction is based on a stochastic representation, sampling algorithms can be given as well. An interesting feature of index-mixed copulas is that they allow one to provide an interpretation of the family of Eyraud-Farlie-Gumbel-Morgenstern (EFGM) copulas, which are popular due to their tractability. Through the lens of index-mixing, one can see the limited range of concordance of EFGM copulas, while this is not the case for index-mixed copulas in general.

Les copules sont d'une importance significative en statistiques multivariées et en gestion des risques, modélisant la relation entre les variables sans tenir compte de leur distribution marginale. Cette présentation introduira les copules mixtes d'indices, une nouvelle classe construite à partir de copules de base et d'un vecteur d'indices aléatoires, montrant un degré remarquable de flexibilité analytique. Les propriétés examinées comprennent des mesures de concordance tels que le coefficient de Kendall, le rho de Spearman, et les ordres de concordance. Puisque l'approche est fondée sur la représentation stochastique, des algorithmes d'échantillonnage peuvent également être donnés. Une caractéristique des copules mixtes d'indices est qu'elles permettent d'interpréter la famille de copules Eyraud-Farlie-Gumbel-Morgenstern (EFGM). À travers la mixité des indices, on peut constater que les copules EFGM ne modélisent qu'une plage limitée de concordance, ce qui n'est pas le cas pour les copules mixtes d'indices.

[15:45-16:00]

Nikola Surjanovic (University of British Columbia) **Saifuddin Syed** (University of Oxford) **Alexandre Bouchard-Côté** (University of British Columbia) **Trevor Campbell** (University of British Columbia)

Exploration-agnostic Geometric Ergodicity of Parallel Tempering

Ergodicité géométrique à exploration-agnostique de l'atténuation parallèle

Non-reversible parallel tempering (NRPT) is an effective algorithm for sampling from target distributions with complex geometry, such as those arising from posterior distributions of weakly identifiable and high-dimensional Bayesian models. In this work we establish geometric ergodicity of NRPT under a model of efficient local exploration. The rates that we obtain are

L'atténuation parallèle irréversible (NRPT) est un algorithme efficace pour l'échantillonnage de distributions cibles comportant une géométrie complexe, comme celles générées par des distributions a posteriori de modèles bayésiens de haute dimension et faiblement identifiables. Dans le cadre de ce travail, nous établissons l'ergodicité géométrique de la NRPT selon un modèle d'exploration locale efficace. Les taux obtenus sont limités en termes de

bounded in terms of an easily-estimable divergence, the global communication barrier (GCB), that was recently introduced in the literature. We obtain analogous ergodicity results for classical reversible parallel tempering, providing new evidence that NRPT dominates its reversible counterpart. Our results are based on an analysis of the hitting time of a continuous-time persistent random walk, related to Telegrapher's equations, which is also of independent interest.

divergence facilement estimable, c'est-à-dire l'obstacle à la communication globale (GCB) qui a récemment été présentée dans la documentation. Nous obtenons des résultats d'ergodicité analogues pour l'atténuation parallèle réversible classique, ce qui est une nouvelle preuve que la NRPT est supérieure à sa version réversible. Nos résultats se basent sur une analyse du temps d'atteinte d'une promenade aléatoire persistante en temps continu, relatif aux équations des télégraphistes, qui est aussi un intérêt indépendant.

[16:00-16:15]

Samuel Valiquette (Université de Sherbrooke) **Éric P. Marchand** (Université de Sherbrooke) **Gwladys Toulemonde** (Université de Montpellier) **Frédéric Mortier** (CIRAD) **Jean Peyhardi** (Université de Montpellier)

Multivariate Discrete Tree Pólya Splitting Distributions

Modèle multivarié discret Tree Pólya splitting

The analysis of multivariate count data is fundamental in various fields. An appropriate model must be able to be flexible enough for inducing correlation, but also simple for inference and interpretation. One solution is the Splitting model which randomly divides the sum of the discrete vector into its components. This simple approach has several interesting properties. However, its dependency structure must be similar for each component. To overcome this problem, we propose a generalization of this model called Tree Pólya splitting. This new model uses a similar approach, but the splitting is represented by a tree structure. In this presentation, we define this model and present various properties such as marginals, factorial moments or dependency structure.

L'analyse des données de comptage multivariées est fondamentale dans divers domaines. Un modèle approprié doit être en mesure d'être flexible pour induire la corrélation, mais également simple pour l'inférence et l'interprétation. Une solution est le modèle Splitting. Celui-ci modélise la somme des composantes du vecteur discret et ensuite la répartition de celle-ci. Cette approche simple possède plusieurs propriétés intéressantes. Cependant, sa structure de dépendance doit être similaire pour chaque composante. Afin de remédier à ce problème, nous proposons une généralisation de ce modèle nommé Tree Pólya splitting. Cette nouvelle approche utilise un principe similaire, mais modélise la répartition à l'aide d'une structure d'arbre. Dans cette présentation, nous définissons ce modèle et nous présentons différentes propriétés comme les marginales, les moments factoriels ou la structure de dépendance.

[16:15-16:30]

Evan Reynolds (Carleton University) **Song Cai** (Carleton University)

Application of Lasso Methods to Parameter Estimation in Density Ratio Models

Application des méthodes Lasso à l'estimation des paramètres dans les modèles de rapport de densité

In the absence of sufficiently large samples, due to expensive sampling costs or other constraints, statisticians can gain power for statistical inference by pooling the information of multiple available samples. If the underlying distributions of the samples being pooled are assumed to share some latent characteristics, one option at the statisticians' disposal is the semi-parametric Density Ratio Model (DRM). An area of interest concerning DRM inference is determining the pre-specified basis function of the model prior to estimating the model parameters. It has been shown that misspecification of this function can have adverse effects with respect to bias and mean-squared error of the estimates. This paper investigates the application of a Group Lasso penalty

En l'absence d'échantillons suffisamment grands, en raison de coûts d'échantillonnage élevés ou d'autres contraintes, les statisticiens peuvent gagner en puissance pour l'inférence statistique en regroupant les informations de plusieurs échantillons disponibles. Si les distributions sous-jacentes des échantillons regroupés sont supposées partager certaines caractéristiques latentes, l'une des options à la disposition des statisticiens est le modèle semi-paramétrique du rapport de densité (DRM). Un domaine d'intérêt concernant l'inférence DRM est la détermination de la fonction de base pré-spécifiée du modèle avant l'estimation des paramètres du modèle. Il a été démontré qu'une mauvaise spécification de cette fonction peut avoir des effets négatifs en ce qui concerne le biais et l'erreur quadratique moyenne des estimations. Cet article étudie l'application d'une pénalité Lasso

to the parameter estimation problem in order to obtain sparse solutions and simplify an overspecified basis function. Simulated results of this novel estimator, computed with multiple algorithms, will be presented.

de groupe au problème de l'estimation des paramètres afin d'obtenir des solutions éparses et de simplifier une fonction de base surspécifiée. Des résultats simulés de ce nouvel estimateur, calculé avec plusieurs algorithmes, seront présentés.

[16:30-16:45]

Christine Allard (Université de Sherbrooke) **Éric P. Marchand** (Université de Sherbrooke)

Bayesian and Minimax Estimators of Loss

Estimateurs bayésiens et minimax de perte

We study the problem of loss estimation that involves the choice of a first-stage estimator of the parameter of interest, incurred loss, and the choice of a second-stage estimator of this loss. We consider both a sequential version where the first-stage estimator and loss are fixed and optimization is performed at the second-stage level, and a simultaneous version with a loss designed for the evaluation of the estimators of the pair composed of the parameter and the loss together. We explore various Bayesian solutions and provide minimax estimators in both cases. The analysis is carried out for several probability models (multivariate normal, Gamma, Poisson, negative binomial), and relates to different choices of the first and second-stage losses. The minimax findings make use of a least favourable sequence of priors and depend critically on particular Bayesian solution properties, namely cases where the second-stage estimator is constant.

Nous étudions le problème de l'estimation de la perte, intégrant le choix d'un estimateur du paramètre au premier stage, la perte subie et le choix d'un estimateur de la perte au second stage de l'estimation. Le problème est abordé sous deux angles, soit avec une approche séquentielle où l'estimateur au premier stage et la perte sont fixés et l'optimisation est effectuée au second stage et avec une approche simultanée avec une perte conçue pour l'évaluation des estimateurs de la paire formée du paramètre et de la perte ensemble. Nous explorons diverses solutions bayésiennes et donnons des estimateurs minimax. L'analyse englobe maints modèles (normale multivariée, Gamma, Poisson, binomiale négative) et se rattache aux choix de pertes faits aux deux stages de l'estimation. Les résultats minimax utilisent une suite de lois a priori la moins favorable et reposent de manière critique sur des propriétés de solutions bayésiennes, notamment les cas où l'estimateur au second stage est constant.

[16:45-17:00]

Ian Waudby-Smith (Carnegie Mellon University) **Martin Larsson** (Carnegie Mellon University) **Aaditya Ramdas** (Carnegie Mellon University)

Distribution-Uniform Strong Laws of Large Numbers

Uniformité de la loi forte des grands nombres sur des familles de distributions

We revisit the question of whether the strong law of large numbers (SLLN) holds uniformly in a rich family of distributions, culminating in a distribution-uniform generalization of the Marcinkiewicz-Zygmund SLLN. These results can be viewed as extensions of Chung's distribution-uniform SLLN to random variables with uniformly integrable q^{th} absolute central moments for $0 < q < 2;\ q \neq 1$. Furthermore, we show that uniform integrability of the q^{th} moment is both sufficient and necessary for the SLLN to hold uniformly at the Marcinkiewicz-Zygmund rate of $n^{1/q-1}$. These proofs centrally rely on novel distribution-uniform analogues of some familiar almost sure convergence results including the Khintchine-Kolmogorov convergence theorem, Kolmogorov's three-series theorem, a stochastic gener-

Nous réexaminons la question de déterminer si la loi forte des grands nombres (LFGN) s'applique uniformément à une famille riche de distributions, ce qui aboutit à une généralisation uniforme de la LFGN de Marcinkiewicz et Zygmund. Ces résultats peuvent être considérés comme des extensions de la LFGN uniforme de Chung aux variables aléatoires avec des moments centraux et absolus d'ordre q qui sont uniformément intégrables pour $0 < q < 2; q \neq 1$. De plus, nous montrons que la condition d'intégrabilité uniforme du moment d'ordre q est à la fois suffisante et nécessaire pour que la LFGN s'applique uniformément aux taux de Marcinkiewicz et Zygmund de $n^{1/q-1}$. Ces preuves s'appuient principalement sur de nouveaux analogues uniformes de résultats de convergence presque sûre, y compris le théorème de convergence de Khintchine et Kolmogorov, le théorème des trois séries de Kolmogorov, une généralisation stochastique du lemme de Kronecker,

alization of Kronecker's lemma, and the Borel-Cantelli lemmas.

et les lemmes de Borel et Cantelli.

Chair/Président: Khurram Nadeem

Room/Salle: A 2065

Date: Tuesday June 4 / mardi 4 juin

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-15:45]

Dayi Li (University of Toronto)

Bayesian Optimization Sequential Surrogate (BOSS) Algorithm: Fast Bayesian Inference for a Broad Class of Bayesian Hierarchical Models

Algorithme de substitution séquentiel d'optimisation bayésienne : inférence bayésienne rapide pour une vaste catégorie de modèles hiérarchiques bayésiens

Approximate Bayesian inference based on Laplace approximation and quadrature methods have become increasingly popular for their efficiency to fit latent Gaussian/extended latent Gaussian models (LGM/ELGM), which encompass various popular hierarchical models. However, many useful models belong to the LGM/ELGM frameworks only if some parameters are fixed. Such models are termed conditional LGM/ELGMs (Gomez-Rubio and Rue, 2018). Existing methods to fit conditional LGM/ELGM rely on grid search or Markov-chain Monte Carlo (MCMC) to explore the unnormalized posterior density. Such procedures become computationally prohibitive beyond simple scenarios, as each evaluation of the density requires fitting a separate LGM/ELGM. In this work, we introduce the Bayesian optimization sequential surrogate (BOSS) algorithm to reduce the computational resource for fitting conditional LGM/ELGMs. With orders of magnitude fewer evaluations compared to grid or MCMC methods, Bayesian optimization generate sequential design points that capture the majority of the posterior mass of the conditioning parameter, which subsequently yields an accurate surrogate posterior distribution that can be normalized with negligible computational cost. We illustrate the efficiency and accuracy and practical utility of the proposed method through extensive simulation studies and real-world applications.

L'inférence bayésienne approximative reposant sur les méthodes d'approximation de Laplace et de quadrature devient de plus en plus populaire en raison de son efficacité à ajuster les modèles gaussiens latents et les modèles gaussiens latents étendus, qui englobent divers modèles hiérarchiques populaires. Cependant, de nombreux modèles utiles ne font partie du cadre des modèles gaussiens latents/modèles gaussiens latents étendus que si certains paramètres sont fixés. Ces modèles sont appelés modèles gaussiens latents/modèles gaussiens latents étendus conditionnels (Gomez-Rubio et Rue, 2018). Les méthodes actuelles d'ajustement des modèles gaussiens latents/modèles gaussiens latents étendus reposent sur la recherche en grille ou la méthode de Monte Carlo par chaîne de Markov pour explorer la densité a posteriori non normalisée. Lorsque les scénarios ne sont pas simples, ces méthodes deviennent compliquées sur le plan informatique, car chaque évaluation de la densité nécessite l'adaptation d'un modèle gaussien latent distinct. Dans le cadre de ces travaux, nous proposons l'algorithme de substitution séquentiel d'optimisation bayésienne afin de réduire les ressources informatiques nécessaires à l'ajustement des modèles gaussiens et des modèles gaussiens latents étendus conditionnels. Grâce à un nombre d'évaluations inférieur de plusieurs ordres de grandeur par rapport aux méthodes de recherche en grille ou méthode de Monte Carlo par chaîne de Markov, l'optimisation bayésienne génère des points de conception séquentiels qui saisissent la majorité de la densité a posteriori du paramètre de conditionnement, ce qui permet d'obtenir une distribution de substitution a posteriori précise pouvant être normalisée à un coût informatique minime. Nous illustrons l'efficacité, la précision et l'utilité pratique de la méthode proposée au moyen d'études de simulation approfondies et d'applications réelles.

[15:45-16:00]

Hui Shen (McGill University) **Eric Kolaczyk** (McGill University)

Consistent Identification of Top-K Nodes in Noisy Networks

Identification consistante de liens « top-k » dans des réseaux bruités

In applied network analysis, one of the key questions involves identifying the most important nodes, typically characterized by various centrality measures. Nevertheless, any inaccuracies inherent in the measurements used for network construction or in the construction of the network itself will inevitably affect the centrality measures, potentially obscuring the key nodes. In this work, we rigorously study the influence of network noise on the recovery of Top-K nodes, focusing on degree centrality. We derive the conditions for consistent recovery and evaluate the feasibility of these conditions under a number of canonical network models. Additionally, we present findings on the infeasibility of detecting vital nodes under certain conditions and demonstrate the implications for network applications. For scenarios that fall between consistency and infeasibility under noise, we propose a confidence set that includes the vital nodes with high probability.

En analyse de réseau, l'une des questions fondamentales à résoudre est l'identification des noeuds les plus importants, généralement caractérisés par plusieurs mesures de centralité. Néanmoins, toute inexactitude propre aux mesures utilisées pour la construction d'un réseau influencera inévitablement les mesures de centralité et pourrait voiler les noueds clés. Dans le cadre de ce travail, nous étudions rigoureusement l'influence du bruit de réseau sur la récupération des noeuds top-k, en nous concentrant sur le degré de centralité. Nous dérivons les conditions pour l'obtention d'une récupération consistante et évaluons la faisabilité de ces conditions selon certains modèles de réseaux canoniques. De plus, nous présentons des résultats en lien avec l'infaisabilité de détecter des noeuds cruciaux sous certaines conditions et démontrons les conséquences relatives à l'application de réseaux. Pour des scénarios qui se tombent entre la consistance et l'infaisabilité du au bruit, nous proposons un ensemble de confiance qui inclut les noeuds importants avec une grande probabilité.

[16:00-16:15]

Shenita Pramij (Memorial University of Newfoundland) **Candemir Cigsar** (Memorial University of Newfoundland) **Yildiz Yilmaz** (Memorial University of Newfoundland)

Mediation Analysis for Recurrent Event Data

Analyse de médiation pour les données d'événements récurrents

Estimating the effects of exposure variables on outcomes has been a topic of high interest, particularly in the fields of medicine, epidemiology and social sciences. In some settings, however, these effects may be mediated by intermediate variables. Mediation analysis methods are used to estimate the direct effects of exposures, as well as indirect effects that may occur through mediating variables. We introduce a method to estimate the controlled direct effect of an exposure in recurrent event processes where measured and unmeasured mediator-outcome confounders may be present. Unlike traditional methods based on additive models, we focus on multiplicative models for event counts, which may include internal covariates representing a dependency on previous event occurrences. We present the results of simulation studies conducted to investigate the finite sample properties of the controlled direct effect estimator. Finally, we illustrate our method using a hospital readmission dataset.

L'estimation des effets des variables d'exposition sur les variables réponses est un sujet qui a soulevé beaucoup d'intérêt, en particulier dans les domaines de la médecine, de l'épidémiologie et des sciences sociales. Dans certains contextes, ces effets peuvent cependant être médiés par des variables intermédiaires. Les méthodes d'analyse de médiation sont utilisées pour l'estimation des effets directs des expositions, de même que des effets indirects que peuvent produire des variables médiatrices. Nous présentons une méthode pour l'estimation de l'effet direct contrôlé de l'exposition dans des processus d'événements récurrents lorsque des facteurs de confusion médiateur-réponse mesurés et non mesurés peuvent être présents. Contrairement aux méthodes traditionnelles basées sur des modèles additifs, nous nous intéressons plutôt à des modèles multiplicatifs de comptage d'événements qui peuvent comprendre des covariables internes représentant une dépendance à des occurrences d'événements antérieurs. Nous présentons les résultats d'études de simulation menées pour examiner les propriétés non asymptotiques de l'estimateur des effets directs contrôlés. Finalement, nous illustrons notre méthode à l'aide d'un ensemble de données de réadmission

à l'hôpital.

[16:15-16:30]

Kevin Granville (University of Windsor) **Douglas Woolford** (University of Western Ontario) **Charmaine B. Dean** (University of Waterloo)

Investigating changes in the timing of Ontario's wildland fire season: a spatial perspective

Étude des changements dans le temps de la saison des feux de forêt en Ontario : une perspective spatiale

Changes in fire regimes can raise wildland fire risk due to increases in frequency, size, and intensity of wildland fires. They may also result in a longer fire season, the portion of the year when most ignitions are observed. A method is proposed to investigate spatial trends in the timings of the start and end of the fire season across Ontario between 1960 – 2022. Using interpolation on the times and locations of historical fires, an algorithm is introduced to create smooth spatial surfaces of fire season start/end dates across our study region, allowing us to test for evidence of monotonic trends. We consider changes to the proportions of the year before, during, and after the fire season, while also contrasting trends based on all ignitions, only human-caused ignitions, or only lightning-caused ignitions. In Northwestern Ontario, we find evidence of human-caused ignitions starting earlier and lightning-caused ignitions continuing later, resulting in longer fire seasons.

Toute modification des régimes d'incendie peut accroître le risque d'incendie de forêt en raison de l'augmentation de leur fréquence, de leur taille et de leur intensité. Elle peut également entraîner un allongement de la saison des feux, cette partie de l'année où l'on observe le plus grand nombre de nouveaux feux. Nous proposons une méthode pour étudier les tendances spatiales des dates de début et de fin de la saison des incendies en Ontario entre 1960 et 2022. En interpolant les dates et lieux des incendies historiques, nous introduisons un algorithme pour créer des surfaces spatiales lisses des dates de début et de fin de la saison des incendies sur notre région d'étude, ce qui nous permet de tester la présence de tendances monotones. Nous examinons les changements dans les proportions de l'année avant, pendant et après la saison des feux, tout en contrastant les tendances basées sur tous les nouveaux feux, uniquement ceux d'origine humaine, ou uniquement ceux causés par la foudre. Dans le nord-ouest de l'Ontario, nous constatons que les incendies d'origine humaine commencent plus tôt et que les incendies provoqués par la foudre se poursuivent plus tard, ce qui se traduit par des saisons des feux plus longues.

[16:30-16:45]

Camila P. E. de Souza (University of Western Ontario) **Pedro H. T. O. Souza** (Universidade Federal do Paraná) **Ronaldo Dias** (Universidade de Campinas)

Bayesian Variable Selection for Function-on-Scalar Regression Models: a Comparative Analysis

Sélection de variables bayésiennes pour des modèles de régression fonction-sur-scalaire : une analyse comparative

In this work, we developed a new Bayesian method for variable selection in function-on-scalar regression (FOSR). Our method uses a hierarchical Bayesian structure and latent variables to enable an adaptive covariate selection in FOSR. Extensive simulation studies show the proposed method's accuracy in estimating the coefficients and high capacity to select variables correctly. Furthermore, we conducted a substantial comparative analysis with the main competing methods, the BGLSS method, the group LASSO, the group MCP, and the group SCAD. Results demonstrate that the proposed methodology is superior in correctly selecting covariates compared with the existing competing methods while maintaining a satisfactory level of goodness of fit. We also considered a COVID-19 dataset from Brazil as an application and obtained satisfactory results. In

Dans le cadre de ce travail, nous élaborons une nouvelle méthode bayésienne pour la sélection de variables d'une régression fonction-sur-scalaire (FOSR). Notre méthode utilise une structure bayésienne hiérarchique et des variables latentes pour permettre une sélection de variables adaptative dans une FOSR. Des études de simulations approfondies démontrent la précision de la méthode proposée pour l'estimation des coefficients et sa capacité à sélectionner correctement des variables. De plus, nous menons une analyse comparative substantielle avec les principales méthodes concurrentes : la méthode BGLSS, le lasso par groupe, le MCP par groupe et une SCAD par groupe. Les résultats démontrent que la méthode proposée surpasse les autres méthodes pour la sélection adéquate de variables tout en conservant un niveau satisfaisant d'adéquation. Nous avons aussi examiné un ensemble de données de la COVID-19 provenant du Brésil en guise de sujet d'application et avons obtenu des résultats satisfaisants. Bref, le modèle de sélection

short, the proposed Bayesian variable selection model is highly competitive, showing significant predictive and selective quality.

de variables bayésiennes proposé est grandement concurrentiel et possède une capacité de sélection et de prédiction considérable.

[16:45-17:00]

W. John Braun (The University of British Columbia)

Monte Carlo Integration of a First Order Differential Equation

Intégration Monte Carlo d'une équation différentielle du premier ordre

Monte Carlo integration of a definite integral is a straightforward application of the Law of Large Numbers applied to simulated data. Integrating a differential equation requires different tactics. A new method based on the Mean Value Theorem is presented and compared with a recent approach proposed in the literature. The new method appears to work well, even on stiff equations.

L'intégration Monte Carlo d'une intégrale définie est une application directe de la loi des grands nombres à des données simulées. L'intégration d'une équation différentielle requiert une tactique différente. Nous présentons une nouvelle méthode basée sur le théorème de la valeur moyenne et la comparons à une approche récente proposée dans la littérature. La nouvelle méthode semble bien fonctionner, même pour les équations rigides.

Chair/Président: Qingrun Zhang

Room/Salle: C 4036

Date: Tuesday June 4 / mardi 4 juin

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-15:45]

Rose Garrett (University of Toronto) **Eleanor M. Pullenayegum** (The Hospital for Sick Children)

Parametric modelling of irregular longitudinal data: a simulation study

Modélisation paramétrique de données longitudinales irrégulières : une étude en simulations

Parametric models of irregular longitudinal data are needed for Bayesian inference and may be more robust to misspecification of the visit process compared to semiparametric approaches. Robustness of univariate parametric models has been assessed under the assumption of a memoryless visit process; however, it would be useful to assess this more generally, exploring scenarios where the intensity of a visit depends on the time elapsed since the last visit. We investigate the robustness of univariate parametric models under an informative visit process that is based on the physician's recommendation on when the patient should return, and compare the performance of the univariate approach to a novel parametric joint model. Using simulation studies, we show that in most cases, the standard linear mixed model estimates parameters of interest with little bias, and for the specific cases where the univariate model is biased, we show how our joint model can eliminate or reduce the bias.

Les modèles paramétriques de données longitudinales irrégulières sont nécessaires pour l'inférence bayésienne et peuvent être plus résistants aux erreurs de classification du processus de visite par rapport aux approches semi-paramétriques. Nous avons évalué la robustesse des modèles paramétriques univariés selon l'hypothèse d'un processus de visite sans mémoire ; cependant, il serait plus utile de les évaluer de façon générale, en explorant des scénarios où l'intensité d'une visite dépend du temps écoulé depuis la dernière visite. Nous étudions la robustesse des modèles paramétriques univariés selon un processus de visite informative basée sur la recommandation du médecin concernant la date de retour du patient, et comparons la performance de l'approche univariée au nouveau modèle conjoint paramétrique. Au moyen d'études de simulations, nous démontrons que, dans la plupart des cas, le modèle mixte linéaire estime les paramètres pertinents avec peu de biais. Dans les cas spécifiques où le modèle univarié est biaisé, nous montrons de quelle façon notre modèle conjoint peut éliminer ou réduire les biais.

[15:45-16:00]

George Stefan (University of Toronto / The Hospital for Sick Children) **Eleanor M. Pullenayegum** (University of Toronto / The Hospital for Sick Children)

Methods for Irregularly Measured Longitudinal Data Subject to Informative Dropout

Méthodes pour des données longitudinales irrégulièrement mesurées assujetties à l'abandon informatif

Longitudinal data are commonly encountered in biomedical research, including randomized trials and retrospective cohort studies. Subjects are typically followed over a period of time and may be scheduled for follow-up at pre-determined time points. However, subjects may miss their appointments or return at non-specified times, leading to irregularity in the visit process. Inverse-intensity weighted generalized estimating equations

Les données longitudinales sont couramment utilisées en recherche biomédicale, y compris dans les essais randomisés et les études de cohorte rétrospectives. Le suivi des sujets est généralement étendu sur une certaine période et peut être prévu à des points temporels prédéterminés. Les sujets peuvent cependant manquer un rendez-vous ou se présenter à des moments non spécifiés, ce qui entraîne l'irrégularité du processus de visite. Des équations d'estimation généralisées à pondération par

(IIW-GEEs) have been developed as one method to account for this irregularity, whereby estimates from a visit intensity model are used as weights in a GEE model with an independent correlation structure. We have shown that currently available methods can be biased for situations in which the health outcome of interest may influence a subject's dropout from the study. We have extended the IIW-GEE framework to adjust for informative censoring and have demonstrated via simulation studies that this bias can be significantly reduced.

intensité inverse (IIW-GEE) ont été développées comme méthode pour la prise en compte de cette irrégularité, par laquelle les estimations d'un modèle d'intensité du processus de visite sont utilisées comme poids dans un modèle GEE avec une structure de corrélation indépendante. Nous avons montré que les méthodes actuellement disponibles peuvent être biaisées dans des situations où les résultats de santé d'intérêt peuvent influer sur l'abandon de l'étude par un sujet. Nous avons étendu le cadre IIW-GEE pour l'ajustement de la censure informative et avons montré à l'aide d'études en simulation que ce biais peut être notablement réduit.

[16:00-16:15]

Marc Angelo Parsons (McGill University) **Andrea Benedetti** (McGill University) **Russell Steele** (McGill University)

Comparing Fractional Polynomial and Spline Meta-Regression Models to Estimate Longitudinal Trajectories in the Presence of Heterogeneity in the Number and Timing of Assessments Between Studies

Une comparaison des modèles de meta-régression employant des bases polynomiales fractionnaires et splines pour l'estimation des trajectoires longitudinales dans la présence de l'hétérogénéité dans le calendrier de l'évaluation des mesures entre les études

Systematic reviews are often interested in the change in health outcomes over time, aggregated over multiple timepoints and studies. Meta-regression may be used to estimate such trajectories. However, it is often the case that included studies present different outcome assessment patterns over time. It is not known how well existing meta-regression methods perform in the presence of such heterogeneity. Simulated data from an individual participant data meta-analysis will be used to compare two existing methods to two proposed extensions. White et al. (2019) proposed using fractional polynomial (FP) basis expansions within linear mixed models to meta-estimate trajectories. We propose two extensions which use spline rather than FP expansions. Results from a simulation that considers varying the level of heterogeneity in the timing and number of assessments between studies will be presented.

Les revues systématiques s'intéressent souvent aux changements des mesures de santé au fil du temps, regroupées sur plusieurs périodes et études. La méta-régression peut être utilisée pour estimer de telles trajectoires. Cependant, il arrive souvent que les études incluses mesurent les résultats sur des calendriers différents. On ne sait pas dans quelle mesure les méthodes existantes fonctionnent en présence d'une telle hétérogénéité. Les données simulées d'une méta-analyse des données individuelles des participants seront utilisées pour comparer deux méthodes existantes à deux extensions proposées. White et al. (2019) ont proposé d'utiliser des expansions de bases polynomiales fractionnaires (FP) dans des modèles mixtes pour meta-estimer les trajectoires. Nous proposons deux extensions qui utilisent des modèles spline plutôt que FP. Les résultats d'une simulation qui prend en compte la variation dans le calendrier d'évaluations entre les études seront présentés.

[16:15-16:30]

Kecheng Li (University of Waterloo) **Richard J. Cook** (University of Waterloo)

Design and Sequential Analysis of Transfusion Trials

Conception et analyse séquentielle d'essais transfusionnels

Transfusion medicine trials often involve repeated administration of blood products for therapeutic and prophylactic reasons. Patients with immune thrombocytopenia will require prophylactic platelet transfusion to mitigate the risk of bleeding, but the transfusion frequency may vary across individuals. We consider group sequential designs of transfusion trials when the goal is to assess the relative effectiveness of a new blood prod-

Les essais de médecine transfusionnelle impliquent souvent l'administration répétée de produits sanguins pour des raisons thérapeutiques et prophylactiques. Les patients atteints de thrombocytopénie immunitaire auront besoin d'une transfusion prophylactique de plaquettes pour réduire le risque de saignement, mais la fréquence des transfusions peut varier d'un individu à l'autre. Nous considérons des plans séquentiels de groupe pour des essais transfusionnels lorsque l'objectif est d'évaluer l'efficacité relative

uct on the response to transfusions. A multivariate probability mass function is used to model the intervention effect and dependence of the binary responses across serial transfusions from each individual. The sample size formula is derived to ensure power requirements are met when analyses are based on generalized estimating equations and robust variance estimation. Strategies for interim monitoring using error spending functions are developed based on a robust covariance matrix for estimates of treatment effect over successive analyses.

d'un nouveau produit sanguin sur la réponse aux transfusions. Nous utilisons une fonction de densité multivariée pour modéliser l'effet de l'intervention et la dépendance des réponses binaires entre les transfusions en série pour chaque individu. Nous dérivons une formule pour la taille de l'échantillon qui garantit que les exigences en matière de puissance sont satisfaites lorsque les analyses sont basées sur des équations d'estimation généralisées et une estimation robuste de la variance. Nous développons des stratégies de contrôle intermédiaire utilisant des fonctions de dépense d'erreur sur la base d'une matrice de covariance robuste pour estimer l'effet du traitement au cours d'analyses successives.

[16:30-16:45]

Hensley Hubert Mariathas (Memorial University of Newfoundland) **Shabnam Asghari** (Memorial University of Newfoundland) **Oliver Hurley** (Memorial University of Newfoundland)

An Application of Interrupted Time Series Modeling using Autoregressive Integrated Moving Average for Evaluation of Quality Improvement Intervention

Application d'une modélisation de série chronologique interrompue à l'aide d'une moyenne mobile autorégressive intégrée (ARIMA) pour l'évaluation d'une intervention d'amélioration de la qualité

Interrupted time series analysis (ITSA) has emerged as a prevalent method for evaluating the effects of policy or broad healthcare interventions over time. Within ITSA, segmental regression analysis offers a robust evaluation of intervention impacts. However, for interventions exhibiting seasonality and autocorrelation, the Autoregressive Integrated Moving Average (ARIMA) model presents a valuable alternative. This presentation explains the foundational theory of ARIMA models and their application in assessing the efficacy of SurgeCon, a pragmatic emergency department (ED) management platform implemented to reduce wait times and improve patient flow in Newfoundland and Labrador EDs without significant changes to workforce volume and composition. Additionally, we detail the process of model selection, fit assessment, and result interpretation, providing insights into the effectiveness of SurgeCon and the utility of ARIMA in evaluating complex healthcare interventions.

L'analyse de série chronologique interrompue (ITSA) est devenue une méthode prévalente pour l'évaluation des effets d'interventions politiques ou générales en soins de santé dans le temps. Dans l'ITSA, l'analyse de régression segmentée offre une évaluation robuste de l'impact des interventions. Pour les interventions montrant une saisonnalité et une autocorrélation, le modèle ARIMA est cependant une solution de rechange utile. Cette présentation explique la théorie fondamentale des modèles ARIMA et leur application pour l'évaluation de l'efficacité de SurgeCon, une plateforme des gestion pragmatique d'un service d'urgence, implémentée afin de réduire le temps d'attente et améliorer le flux de patients dans les urgences de Terre-Neuve-et-Labrador, sans changements marqués dans le volume et la composition de la main-d'œuvre. De plus, nous expliquons en détail le processus de sélection du modèle, l'évaluation de l'ajustement et l'interprétation des résultats, fournissant ainsi un éclairage sur l'efficacité de SurgeCon et l'utilité des modèles ARIMA pour l'évaluation d'interventions complexes en soins de santé.

[16:45-17:00]

Mathilde Dicaire-Cartier (Université de Montréal) **Janie Coulombe** (Université de Montréal)

Estimating the Causal Effect of a Cumulative Exposure on a Continuous Outcome in Studies Prone to Confounding and Irregular Visits

Estimation de l'effet causal d'une exposition cumulative sur une réponse continue dans les études enclines à la confusion et aux visites irrégulières

Non-experimental data, such as electronic medical records, are often used for causal inference to estimate the effect of an exposure on an outcome variable. How-

Les données non expérimentales, comme celles de dossiers médicaux électroniques, sont souvent utilisées pour faire de l'inférence causale afin d'estimer l'effet d'une exposition sur une

ever, these data do not come from a study design ensuring a balance of patient characteristics between exposure groups. Patients are also observed irregularly over time. These imbalances can bias the estimation of causal effects. Methods have recently been proposed to address these challenges, but they mostly focused on acute treatment effects. In this presentation, we propose a methodology to consistently estimate the causal effect of a cumulative exposure over time on a continuous response. It allows for the consideration of delayed treatment effects and the causal effect is estimated from irregularly measured responses. Confounding and bias induced by irregular observation times are addressed through weighting methods.

réponse. Ces données ne proviennent pas d'un plan d'étude assurant une balance des caractéristiques des patients entre les groupes d'exposition. Les patients sont aussi observés de façon irrégulière dans le temps. Ces déséquilibres peuvent biaiser l'estimation de l'effet causal. Des méthodes ont été proposées pour adresser ces défis, mais elles se sont surtout concentrées sur des effets de traitement aigus. Dans cette présentation, nous proposons une méthodologie permettant d'estimer l'effet causal d'une exposition cumulée dans le temps sur une réponse continue. Elle permet de prendre en compte les effets de traitement délayés estimés à partir d'une réponse mesurée irrégulièrement. La confusion et le biais induit par les temps d'observation irréguliers sont traités à partir de méthodes de pondération.

Chair/Président: Himchan Jeong

Room/Salle: C 3053

Date: Tuesday June 4 / mardi 4 juin

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-15:45]

Toby J. Kenney (Dalhousie University) **Yun Cai Hong Gu** (Dalhousie University)

New Methods and Applications of Deconvolution

Nouvelles méthodes et applications de déconvolution

Deconvolution is the problem of estimating the distribution of a random variable from a sample with additive measurement error. Most penalised MLE methods restrict the space of possible solutions to a fixed a priori finite basis. However, we find a finite-dimensional subspace that contains the penalised MLE and use standard optimisation methods to obtain the infinite-dimensional MLE. This method is consistent and performs well in simulations. We also present a novel application of deconvolution - bootstrap sampling. Our motivating example is a likelihood ratio test to determine the rank of Nonnegative Matrix Factorisation (NMF). We estimate the null distribution via a bootstrap. Optimisation for NMF often finds a local optimum. A computationally more efficient approach than taking multiple starting points, is to bootstrap with optimisation error, then use deconvolution to remove it. This approach performs as well using multiple starting points, with much shorter computation time.

La déconvolution est le problème d'estimation de la distribution d'une variable aléatoire avec une erreur de mesure additive. La plupart des méthodes d'estimateur du maximum de vraisemblance (MLE) limitent l'espace des solutions possibles à une base finie a priori fixe. Nous trouvons cependant un sous-espace à dimension finie qui contient un MLE pénalisé et utilisons des méthodes d'optimisation pour obtenir le MLE à dimension infinie. Cette méthode est cohérente et performante dans les simulations. Nous présentons également une nouvelle application de déconvolution – échantillonnage bootstrap. L'exemple motivant cette application est un test du ratio de vraisemblance pour déterminer le rang de la factorisation matricielle non négative (NMF). Nous estimons la distribution nulle par bootstrap. L'optimisation de la NMF trouve souvent un optimum local. Une approche d'une plus grande efficacité computationnelle que celle des points de départ multiples est un bootstrap avec erreur d'optimisation puis l'utilisation d'une déconvolution pour l'enlever. Cette approche fonctionne aussi bien en utilisant des points de départ multiples, avec un temps de calcul plus court.

[15:45-16:00]

Jervis Gallanosa (University of Manitoba) **Yuliya V. Martsynyuk** (University of Manitoba)

On More Powerful Nonparametric Tests for Change in the Mean with Better Controlled Type I Errors

Tests non paramétriques plus puissants pour un changement de la moyenne avec erreurs de type 1 mieux contrôlées

Let n independent, chronologically ordered observables either form a random sample with a finite variance, or be such that the first k ones have a common mean that is different from the common mean of the last n-k, $1 \leq k < n$, where the time k of the change in the mean is usually unknown. We study numerically the finite-sample power functions of nonparametric tests for such a change that are based on convergence in distribution of sup- and integral-functionals of an appropriately

Soit n observations indépendantes en ordre chronologique qui forment, ou bien un échantillon aléatoire avec variance finie, ou qui sont telles que les k premières observations ont une moyenne commune qui est différente de la moyenne commune des dernières n-k, $1 \leq k < n$, où le temps k du changement de la moyenne est habituellement inconnu. Nous étudions de façon numérique les fonctions de puissance de tests non paramétriques pour détecter un tel changement pour des échantillons de taille fini qui sont basées sur la convergence en loi des fonctionnelles supérieures et

weighted tied-down partial sums process. For each test, a three-way trade-off is observed among its type I errors, the power for detecting the change on the tails of the sample, and the power for detecting the change in the middle, more pronounced for samples from highly skewed and/or heavy-tailed distributions. By choosing suitable weight functions, we obtain new sup- and integral-based tests that are at least as powerful for various times k as those in the literature and have better controlled type I errors.

intégrales d'un processus de sommes partielles liées adéquatement pondérées. Parmi les erreurs de type 1 de chaque test, un compromis à trois voies est observé, la puissance de détection des changements sur les queues de l'échantillon et la puissance de détection des changements au centre, plus prononcée dans les échantillons de distributions très asymétriques ou à queue lourde. Le choix de fonctions de pondération adéquates permet d'obtenir de nouveaux tests, basés sur les fonctionnelles supérieures et intégrales, au moins aussi puissants pour divers temps k que ceux figurant dans la documentation, et avec des erreurs de type 1 mieux contrôlées.

[16:00-16:15]

Ethan Lawler (Dalhousie University) **Joanna Elizabeth Mills Flemming** (Dalhousie University) **Chris Field** (Dalhousie University)

Automatic Outlier Detection and Robust Filtering for Multivariate, Irregular, and Heteroscedastic State-Space Models

Détection automatique des valeurs aberrantes et filtrage robuste pour modèles espace-état multivariés, irréguliers et hétéroscédastiques

Modern state-space models are typically employed to predict the true state of a system using noisy observations of that system. While outliers may represent a small percentage of the data, the large size of modern datasets means that the total number of outliers can be enough to significantly impact the statistical analysis. Further, the size of the datasets often precludes manual removal of the outliers. We introduce a semi-parametric state-space model paired with a robust estimation procedure that automatically detects and downweights outlier observations to improve prediction of the true state of the system. Our model is built using the R package Template Model Builder for efficient computation of a score-weighted marginal log-likelihood function using the Laplace approximation. We discuss our implementation and usage of the model for filtering marine grey seal movement paths where observations are multivariate, irregularly spaced in time, and with heteroscedastic observation error.

Les modèles espace-état modernes sont généralement utilisés pour prédire l'état réel d'un système à partir d'observations bruitées de celui-ci. Bien que les valeurs aberrantes puissent ne représenter qu'un faible pourcentage des données, la taille importante des ensembles de données modernes signifie que le nombre total de ces valeurs peut avoir un impact significatif sur l'analyse statistique. En outre, la taille des ensembles de données empêche souvent l'élimination manuelle des valeurs aberrantes. Nous introduisons un modèle espace-état semi-paramétrique associé à une procédure d'estimation robuste qui détecte et pondère automatiquement les observations aberrantes afin d'améliorer la prédiction de l'état réel du système. Notre modèle est construit à l'aide du paquet R Template Model Builder pour le calcul efficace d'une fonction de log-vraisemblance marginale pondérée par le score à l'aide de l'approximation de Laplace. Nous discutons de notre mise en œuvre et de l'utilisation du modèle pour filtrer les trajectoires de déplacement des phoques gris marins lorsque les observations sont multivariées, irrégulièrement espacées dans le temps et avec une erreur d'observation hétéroscédastique.

[16:15-16:30]

Armin Hatefi (Memorial University of Newfoundland) **Moein Yoosefi** (Memorial University of Newfoundland)

Shrinkage Methods for Contaminated Mixture Models with Matrix-valued Data

Méthodes de rétrécissement pour modèles de mélange contaminés avec des données matricielles

In this research, we develop shrinkage model-based clustering methods with mixture models arising from contaminated matrix-valued data. Despite the popularity in multivariate settings, due to their two-way dependent structures, not only the matrix-valued data are prone to high dimensional, but also the input data are often contaminated by the presence of outliers. By impos-

Dans cette présentation, nous développons des méthodes de regroupement basées sur des modèles de rétrécissement avec des modèles de mélange contaminés provenant de données matricielles. Malgré leur popularité dans les contextes multivariés, en raison de leurs structures dépendantes à double sens, non seulement les données matricielles sont susceptibles d'avoir une dimension élevée, mais les données d'entrée sont souvent contaminées

ing sparsity structures, we develop two-layered missing data mechanisms to simultaneously detect the outlier matrices and handle the regularization steps to estimate accurately the mean signal of the clusters. Through extensive numerical studies, we evaluate the performance of the proposed methods and show they lead to more accurate estimation and prediction. Finally, the methods are applied to real data examples.

par la présence de valeurs aberrantes. En imposant des structures de parcimonie, nous développons des mécanismes de données manquantes à deux niveaux pour détecter simultanément les matrices aberrantes et gérer les étapes de régularisation afin d'estimer avec précision le signal moyen des groupes. À l'aide d'études numériques approfondies, nous évaluons les performances des méthodes proposées et montrons qu'elles permettent une estimation et une prédiction plus précises. Enfin, nous appliquons nos méthodes à des exemples de données réelles.

[16:30-16:45]

Jia Wei He **Ayesha Ali** (University of Guelph)

Proximal Projection for Doubly Sparse Regularized Models

Projection proximale pour les modèles régularisés à double parcimonie

Regularization is often used for variable selection in linear models. Doubly sparse regression incorporating graphical structure among predictors (DSRIG) exploits the underlying structure associated with the predictor graph by decomposing the estimated coefficient vector into a sum of latent variables, corresponding to the sum of each node's contribution to the coefficient vector, and performs regularization on the latent variables rather than on the coefficient vector directly. We propose a novel proximal projection to replace the predictor duplication used in DSRIG and reparametrize the penalty function to permit a clear user-defined trade-off between the L1 and L2 penalties. Through simulation, we evaluate the performance of our approach to predictor duplication, among other methods, and present results on real world data. Preliminary results suggest that our method exhibits stable performance relative to DSRIG with predictor duplication and other singly sparse regression models.

La régularisation est souvent utilisée pour la sélection de variables dans des modèles linéaires. La régression à double parcimonie intégrant une structure graphique parmi les variables explicatives (DSRIG) exploite la structure sous-jacente associée au graphe des variables explicatives. Pour ce faire, elle décompose le vecteur de coefficients estimé en une somme de variables latentes, qui correspond à la somme de la contribution de chaque nœud à ce vecteur, puis elle régularise les variables latentes au lieu de régulariser directement le vecteur de coefficients. Nous proposons une nouvelle méthode de projection proximale pour remplacer la duplication des variables explicatives utilisée dans DSRIG, puis nous paramétrons de nouveau la fonction de pénalité pour permettre à l'utilisateur de faire un choix éclairé entre les pénalités L1 et L2. Ensuite, nous évaluons, à l'aide de simulation, les résultats de notre approche par rapport à la duplication de variables explicatives, parmi d'autres méthodes, et nous présentons les résultats obtenus par rapport aux résultats avec des données réelles. Les résultats préliminaires suggèrent que notre méthode est stable par rapport à la DSRIG utilisant la duplication de variables explicatives et d'autres modèles de régression à parcimonie unique.

[16:45-17:00]

Kai Yang (McGill University) **Masoud Asgharian** (McGill University) **Celia Greenwood** (McGill University)

Tsallis Entropy-Based Method for Sparse Statistical Machine Learning on Correlated Data and a Proximal Conjugate Gradient Algorithm for Nonconvex Nonsmooth Objective Function

Méthode basée sur l'entropie de Tsallis pour l'apprentissage machine statistique éparse de données corrélées et algorithme du gradient conjugué proximal pour une fonction objective non lisse et non convexe

The principle of maximizing entropy when applied to Tsallis entropy leads to the q-Gaussian distribution. This framework broadens the scope of bell curve distributions, offering a robust model for analyzing heavy-tailed data. q-Gaussian distributions have been effectively applied across various fields, including physics, finance, and biology. This paper presents a novel ap-

Le principe de maximisation de l'entropie lorsqu'il s'applique à l'entropie de Tsallis nous mène à la loi q-gaussienne. Ce cadre étend la portée des lois de courbe en cloche, offrant un modèle robuste pour l'analyse des données à queue lourde. Les distributions q-gaussiennes ont été efficacement appliquées à divers domaines, dont la physique, la finance et la biologie. Cet article présente une nouvelle approche qui utilise la loi q-gaussienne mul-

proach that utilizes the multivariate q-Gaussian distribution in conjunction with oracle penalties to accurately model correlated data. Our simulations highlight the benefits of this method over traditional approaches. Additionally, we propose a new proximal conjugate gradient technique designed for optimizing nonconvex and nonsmooth objective functions, which demonstrates a markedly faster convergence rate than existing alternatives.

tivariée en conjugaison avec des méthodes de pénalité Oracle pour la modélisation adéquate des données corrélées. Des simulations mettent en lumière les avantages de cette méthode par rapport aux méthodes traditionnelles. De plus, nous proposons une nouvelle technique du gradient conjugué proximal conçue pour optimiser les fonctions objectives non lisses et non convexes, dont le taux de convergence est nettement plus rapide que celui d'autres techniques existantes.

Chair/Président: Hedayat Fathi

Room/Salle: ED 2018B

Date: Tuesday June 4 / mardi 4 juin

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-15:45]

Skye Paphora Griffith (Queen's University) **Wesley Burr** (Trent University) **Glen Takahara** (Queen's University)

Boundary Correction and Smoothing Methods for the Spectrograms of Uniformly Modulated Processes

Méthodes de correction aux frontières et méthodes de lissage pour les spectrogrammes de processus uniformément modulés

Among nonstationary time series, the class of Uniformly Modulated Processes (UMPs) have Evolutionary Power Spectra (EPS) that are the outer product of the modulating function of time and the spectrum of a stationary process. In this paper we propose a smoothing procedure on an estimator of the EPS of a general nonstationary process to obtain an estimator of the EPS of a UMP. This is done by first obtaining isolated estimates of the modulating function of time and the spectrum of the stationary component of UMPs, then combining them to form our smoothed estimator of the EPS. Via simulation, we compare our estimator to the High Resolution Spectrogram, and also to an estimator that minimizes the Frobenius norm between the general estimator and one that has the structure of an EPS of a UMP. Moreover, we build upon previously established time-boundary correction methods for sliding window spectrograms who have been modified via nonstationary quadratic inverse theory.

Parmi les séries temporelles non stationnaires, la classe des processus uniformément modulés (PUM) a des spectres de puissance évolutifs (SPE) qui sont données par le produit extérieur de la fonction de modulation du temps et du spectre d'un processus stationnaire. Dans cet article, nous proposons une procédure de lissage sur un estimateur du SPE d'un processus non stationnaire général afin d'obtenir un estimateur du SPE d'un PUM. Pour ce faire, nous obtenons d'abord des estimations isolées de la fonction de modulation du temps et du spectre de la composante stationnaire des PUM, puis nous les combinons pour former notre estimateur lissé du SPE. Par simulation, nous comparons notre estimateur au spectrogramme à haute résolution, ainsi qu'à un estimateur qui minimise la norme de Frobenius entre l'estimateur général et un estimateur qui a la structure d'un SPE d'un PUM. En outre, nous nous appuyons sur des méthodes de correction aux frontières temporelles précédemment établies pour les spectrogrammes à fenêtre glissante qui ont été modifiées par la théorie inverse quadratique non stationnaire.

[15:45-16:00]

Bowei Ding (University of Calgary) **Jingjing Wu** (Department of Mathematics and Statistics, University of Calgary) **Rohana J. Karunamuni** (Department of Mathematical and Statistical Sciences, University of Alberta)

Minimum Profile Hellinger Distance Estimation of Covariate Models

Estimation du profil de distance de Hellinger minimal pour modèles à covariables

Covariate models, such as linear models, generalized linear models and single-index models are widely used in statistical applications. Because of their flexibility, covariate models are being increasingly exploited as a convenient and semi-parametric way to model data that consists of a response variable and covariate variables that affect the response variable. This study investigates efficient and robust estimations for both general covari-

Les modèles à covariables, tels que les modèles linéaires, les modèles linéaires généralisés et les modèles à indice unique, sont largement utilisés dans les applications statistiques. En raison de leur souplesse, les modèles à covariables sont de plus en plus exploités comme un moyen pratique et semi-paramétrique de modéliser des données qui consistent en une variable de réponse et des variables covariées qui affectent celle-ci. Cette étude s'intéresse aux estimations efficaces et robustes pour les

ate models and single-index models, as the most commonly used covariate models in statistical applications. For this purpose, we employ the minimum distance approach. In particular, the minimum Hellinger distance approach introduced by Beran (1977) produces estimators that are asymptotically efficient at the model density and simultaneously possess excellent robustness properties. In this study, we construct several minimum profile Hellinger distance estimators for the considered models. We investigate the asymptotic properties of the proposed estimators. To ease the calculation, a computational algorithm is also developed. Finite-sample performance regarding both efficiency and robustness of the proposed estimators are examined using Monte Carlo simulation studies and real data analysis.

modèles à covariables générales et les modèles à indice unique, qui sont les plus couramment utilisés en statistique. Pour ce faire, nous utilisons l'approche de la distance minimale. En particulier, l'approche de la distance de Hellinger minimale introduite par Beran (1977) produit des estimateurs qui sont asymptotiquement efficaces pour la densité du modèle et qui possèdent simultanément d'excellentes propriétés de robustesse. Dans cette étude, nous construisons plusieurs estimateurs du profil de distance de Hellinger minimal pour les modèles considérés. Nous étudions les propriétés asymptotiques des estimateurs proposés. Pour faciliter le calcul, nous développons également un algorithme computationnel. Nous examinons les performances (efficacité et robustesse) des estimateurs proposés sur échantillon fini à l'aide d'études de simulation de Monte Carlo et d'analyses de données réelles.

[16:00-16:15]

Jingyue Huang (University of Waterloo) **Changbao Wu** (University of Waterloo) **Leilei Zeng** (University of Waterloo)

Empirical likelihood approaches to estimating quantile treatment effects

Approches de vraisemblance empirique pour l'estimation des effets de traitement par quantile

The average treatment effect offers valuable insights in causal inference problems for guiding precise decision-making. However, understanding quantile treatment effects (QTE) is equally crucial as it illuminates how interventions affect different segments of a population. In this talk, we introduce one pseudo empirical likelihood and two sample empirical likelihood approaches, augmented with model-calibration constraints, to construct doubly robust estimators for QTE. Two types of model-calibration constraints are proposed, one utilizing multiple imputation of potential outcomes and another directly modelling the indicator functions. For each of the six frameworks, bootstrap-calibrated confidence intervals using the point estimator and empirical likelihood ratio are computed, respectively. Computation challenges and simulation results will be presented. This is a joint work with my PhD supervisors, Dr. Changbao Wu and Dr. Leilei Zeng.

L'effet de traitement moyen offre des indications précieuses dans les problèmes d'inférence causale et permet de guider une prise de décision précise. Cependant, une compréhension des effets de traitement par quantile (ETQ) est tout aussi cruciale car elle permet de comprendre comment les interventions affectent les différents segments d'une population. Dans cet exposé, nous présentons une approche de pseudo-vraisemblance empirique et deux approches de vraisemblance empirique d'échantillon, augmentées par des contraintes de calibrage de modèle, pour construire des estimateurs doublement robustes pour l'EQT. Nous proposons deux types de contraintes de calibrage de modèle, l'un utilisant l'imputation multiple des résultats potentiels et l'autre modélisant directement les fonctions indicatrices. Pour chacun des six cadres, nous calculons des intervalles de confiance calibrés par bootstrap via l'estimateur ponctuel et le rapport de vraisemblance empirique, respectivement. Nous présenterons les difficultés de calcul et les résultats des simulations. Ce travail a été réalisé en collaboration avec mes directeurs de thèse, Changbao Wu et Leilei Zeng.

[16:15-16:30]

Saba Saghatchi (University of Calgary) **Xuewen Lu** (University of Calgary) **Jingjing Wu** (University of Calgary)

Variable Selection for Generalized Odds Rate Non-Mixture Cure Models with Current Status Data

Sélection de variable pour les modèles généralisés de non-mélange avec taux de guérison avec des données d'état actuel.

Current status data are common in clinical trials and epidemiological studies. In some cases, there exists a cured sub-population, where individuals never experience the event of interest. In practice, one may encounter a large number of risk factors, so variable selection is desirable

Les données sur l'état actuel sont courantes dans les essais cliniques et les études épidémiologiques. Dans certains cas, il existe une sous-population traitée dans laquelle les individus ne subissent jamais l'événement d'intérêt. En pratique, on peut rencontrer un grand nombre de facteurs de risque, de sorte que la sélection de

in model building. This paper studies variable selection for the generalized odds rate non-mixture cure models with current status data when the number of covariates diverges with the sample size. To estimate the unknown function, the sieve method based on Bernstein Polynomials is adopted. To facilitate computation, we implement a penalized expectation maximization (EM) algorithm. In theory, we show that the proposed penalized method possesses the oracle properties. Furthermore, we conduct a simulation study to assess the finite sample performance of the proposed method. Finally, the method is applied to the Wisconsin Prognostic Breast Cancer (WPBC) dataset to identify the risk factors of cancer disease.

variable est souhaitable dans la construction du modèle. Cet article étudie la sélection de variable pour les modèles généralisés de non-mélange avec taux de guérison avec des données d'état actuel lorsque le nombre de covariables diverge avec la taille de l'échantillon. Pour estimer la fonction inconnue, la méthode du tamisage basée sur les polynômes de Bernstein est adoptée. Pour faciliter le calcul, nous implémentons un algorithme de maximisation de l'espérance pénalisée (EM). En théorie, nous montrons que la méthode pénalisée proposée possède les propriétés d'oracle. En outre, nous menons une étude de simulation pour évaluer la performance de la méthode proposée sur un échantillon fini. Enfin, la méthode est appliquée aux données de Wisconsin Prognostic Breast Cancer (WPBC) afin d'identifier les facteurs de risque de la maladie cancéreuse.

[16:30-16:45]

Hao He (University of Ottawa) **Hao He** (University of Ottawa) **David Haziza** (University of Ottawa) **Song Cai** (Carleton University)

Empirical Likelihood for Density Ratio Model with Missing Data

Vraisemblance empirique d'un modèle de ratio de densité avec données manquantes

The semiparametric density ratio model (DRM) is a powerful tool for inference problems regarding multiple samples that are collected from possibly different populations. Empirical likelihood (EL) based on fully observed data has been studied extensively for statistical inference under the DRM. However, to the best of our knowledge, no method has been proposed for dealing with missing data under the DRM. In this work, we propose an EL method for inference under the DRM with data that are missing at random. We impute the missing data with a kernel imputation procedure, and construct a dual profile-EL function based on the imputed data. The resulting estimators of the DRM parameters are shown to be consistent and asymptotically normal. Results from a simulation study show that the proposed method outperforms competitors, especially when the underlying distributions of the data are skewed.

Le modèle de ratio de densité semiparamétrique (DRM) est un outil fort utile pour les problèmes d'inférence relatifs aux échantillons multiples qui sont collectés dans des populations possiblement différentes. La vraisemblance empirique (EL) basée sur des données entièrement observées a été longuement étudiée pour l'inférence statistique dans un DRM. À notre connaissance, aucune méthode n'a été proposée pour traiter les données manquantes dans un DRM. Nous proposons ici une méthode de vraisemblance empirique pour l'inférence dans un DRM avec des données manquantes au hasard. Nous utilisons une procédure d'imputation par noyau des données manquantes et élaborons une fonction EL à profil double basée sur les données imputées. On voit que les estimateurs qui résultent des paramètres du DRM sont cohérents et normaux sur le plan asymptotique. Les résultats d'une étude en simulation montrent que la méthode proposée surpasse les méthodes concurrentes, en particulier lorsque les distributions sous-jacentes de données sont asymétriques.

[16:45-17:00]

Louis Arsenault-Mahjoubi (Simon Fraser University) **Jean-François Bégin** (Simon Fraser University)

A generalized Computational Method for Nonlinear Non-Gaussian Filtering in Finance

Une méthode computationnelle généralisée pour le filtrage nonlinéaire et non gaussien en finance

Financial market dynamics exhibit nonlinearities, discontinuities or jumps, and are best modelled with several latent variables. Researchers often use particle filtering to estimate the likelihood function of these complicated models as it is flexible, simple to implement, and avoids the curse of dimensionality. However, recent

Les marchés financiers ont des comportements non linéaires, des discontinuités ou des sauts, et sont mieux modélisés avec plusieurs variables latentes. Les chercheurs utilisent souvent le filtre particulaire pour obtenir un estimé de la vraisemblance de ces modèles compliqués, car ce filtre est flexible, simple et évite le fléau de la dimension. Par contre, dans beaucoup de cas pratiques, des

work demonstrates that, in many practical scenarios, deterministic filters based on numerical integration offer faster and smoother estimates of the likelihood function. Integration by parts-based methods are able to accelerate numerical integration in deterministic filters. Yet, integration by parts has only been applied to limited modelling frameworks in the past. In this talk, I will present a generalization of these filtering methods. I then use the generalized filter to accelerate the estimation of models that allow for the leverage effect and multiple latent factors — both key empirical features of market data— in simulation and empirical studies.

filtres déterministes sont plus rapides et offrent des évaluations plus lisses de la fonction de vraisemblance. L'intégration par parties peut accélérer l'intégration numérique dans les filtres déterministes. Par contre, cette méthode a seulement été appliquée dans des cas très limités. Ici, je vais présenter une méthode générale basée sur l'intégration par parties. Je vais ensuite utiliser cette généralisation pour accélérer l'estimation de modèles qui incorporent l'effet levier et plusieurs variables latentes – des éléments importants en finance empirique— dans des études de simulation et empirique.

Chair/Président: Anand N Vidyashankar

Room/Salle: C 2033

Date: Tuesday June 4 / mardi 4 juin

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-15:45]

Ana Carolina da Cruz (University of Western Ontario) **Camila P. E. de Souza** (University of Western Ontario)

Variational Bayes for Basis Function Selection for Functional Data Representation with Correlated Errors

Un algorithme variationnel bayésien pour la sélection des fonctions de base pour la représentation de données fonctionnelles avec des erreurs corrélées

Functional data analysis (FDA) has found extensive application across various fields, driven by the increasing recording of data continuously over a time interval or at several discrete points. FDA provides the statistical tools specifically designed for handling such data. Over the past decade, Variational Bayes (VB) algorithms have gained popularity in FDA, primarily due to their speed advantages over MCMC methods. This work proposes a VB algorithm for basis function selection for functional data representation while allowing for a complex error covariance structure. We assess and compare the effectiveness of our proposed VB algorithm with MCMC via simulations. We also apply our approach to a publicly available dataset. Our results show the accuracy in coefficient estimation and the efficacy of our VB algorithm to find the true set of basis functions. Notably, our proposed VB algorithm demonstrates a performance comparable to MCMC but with substantially reduced computational cost.

L'analyse de données fonctionnelles (FDA) a trouvé des applications extensives dans différents domaines en raison de la croissance de collectes de données en continu sur un intervalle de temps ou à un grande nombre discret de points. La FDA fournit des outils statistiques spécialement conçus pour manipuler telles données. Au long de la décennie précédente, des algorithmes variationnels bayésiens (VB) ont gagné en popularité dans la FDA, essentiellement en raison de leur avantage en terme de temps de calcul par rappport aux méthodes MCMC. Ce travail propose un algorithme VB pour la sélection des fonctions de base pour la représentation des données fonctionnelles tout en permettant une structure complexe dans la covariance des erreurs. Nous évaluons et comparons l'efficacité de l'algorithme VB proposé avec le MCMC par des simulations. Nous appliquons aussi notre approche à un jeu de données publiques. Nos résultats démontrent l'exactitude de l'estimation des coefficients et l'efficacité de notre algorithme VB pour trouver les vraies fonctions de base. On peut souligner que l'algorithme VB proposé montre une performance comparable à celle du MCMC mais avec une diminution considérable du coût de calcul.

[15:45-16:00]

Mélanie Raymond (Université du Québec à Montréal)

Building Ancestral Recombination Graphs with Reinforcement Learning

Construire des Généalogies de population en utilisant l'apprentissage par renforcement

The ancestral recombination graph (ARG) is used to represent the genetic relationship between a sample of individuals. It plays a key role in biological and genetic studies. But since we cannot go back in time, it is impossible to know the real relationship between a set of genetic sequences. So, we have to infer it. Many approaches have been proposed over the years. In this talk, we will explore a new one: Reinforcement Learn-

Le graphe de recombinaison ancestral (ARG) est utilisé pour représenter la relation génétique qui relie un ensemble d'individus. Comme on ne peut voyager dans le temps, il est impossible de connaître cette relation. Il faut donc l'inférer. Différentes approches ont été proposées au fil des ans. Dans la présentation, nous en explorons une nouvelle : l'apprentissage par renforcement. Nous nous basons sur les méthodes utilisées pour apprendre le chemin le plus court pour sortir d'un labyrinthe. Plusieurs

ing (RL). RL can be used to learn the shortest way out of a maze. We use the same principle. Many methods for building ARGs are based on the assumption that the most likely graph is among the shortest ones. So, we are looking for the shortest path between a set of genetic sequences and their most recent common ancestor. In this talk, we introduce the problem thoroughly, discuss genetic sequences representation, explain how RL can be used to build ARGs for a given sample, and show how our model can generalize its learning to samples not seen during training.

méthodes pour construire des ARGs reposent sur l'hypothèse que le graphe le plus probable est parmi les plus courts. Nous cherchons donc le chemin le plus court entre un ensemble de séquences génétiques et leur ancêtre commun le plus récent. Dans cette présentation, nous présentons le problème en détail, discutons de la représentation des séquences génétiques, expliquons comment apprendre à construire des ARGs pour un échantillon donné et montrons comment généraliser les apprentissages.

[16:00-16:15]

Ashani N. Wickramasinghe (University of Manitoba) **Saman Muthukumarana** (University of Manitoba) **Matt Schaubroeck** (ioAirFlow)

Hotspot Analysis in Buildings using Moran's I Statistic

Analyse des points chauds dans les immeubles avec la statistique I de Moran

Hotspot analysis can identify areas with a higher concentration of events compared to the expected number in a given random distribution of events. Hotspot analysis has been widely used in various research disciplines, such as public health, crime, and environmental quality-related research. In this study, we applied hotspot analysis and Local Moran's I indices to pinpoint high and low-temperature clusters within commercial buildings. Our study demonstrates that hotspot analysis is applicable inside a building. Using the Moran's I statistic, we introduced a method to determine the optimal number of neighbors. Additionally, random forests were employed to identify important placement features influencing hot/cold spots. The vacancy of the place and the presence of windows emerged as crucial factors in our application. In conclusion, our combined application of Moran's I statistic and hotspot analysis effectively identifies hot and cold areas within a building, when the number of neighbors is accurately identified and there are no outliers or abnormal sensors.

L'analyse des points chauds permet de recenser les zones où la concentration d'événements est plus élevée que le nombre attendu dans une distribution aléatoire donnée d'événements. L'analyse des points chauds a été largement utilisée dans diverses disciplines de recherche, telles que la santé publique, la criminalité et la recherche liée à la qualité de l'environnement. Dans cette étude, nous avons appliqué l'analyse des points chauds et les statistiques I locales pour identifier les grappes de températures élevées et basses dans les immeubles commerciaux. Notre étude démontre que l'analyse des points chauds est applicable à l'intérieur d'un immeuble. Grâce à la statistique I de Moran, nous avons élaboré une méthode permettant de déterminer le nombre optimal de voisins. De plus, nous avons utilisé des forêts aléatoires pour identifier les caractéristiques importantes de l'emplacement qui influencent les points chauds et froids. Par ailleurs, l'inoccupation du lieu et la présence de fenêtres sont apparues comme des facteurs cruciaux dans notre application. En conclusion, notre application combinée de la statistique I de Moran et de l'analyse des points chauds permet d'identifier efficacement les zones chaudes et froides d'un immeuble lorsque le nombre de voisins est correctement identifié et qu'il n'y a pas de valeurs aberrantes ou de capteurs anormaux.

[16:15-16:30]

Jasper Zhongyuan Zhang (University of Toronto) **Rafal Kustra** (University of Toronto) **Davide Chicco** (University of Toronto and Università di Milano-Bicocca)

Identifying Clinically Relevant Clusters within Cognitive State Research among a Large Adult Population

Identification de groupes cliniquement pertinents dans la recherche sur l'état cognitif au sein d'une large population adulte

Dementia, affecting millions worldwide, gains importance with aging populations. Our study, leveraging the Canadian Longitudinal Study on Aging (CLSA) data, aimed to cluster patients by cognitive changes over

La démence, touchant des millions de personnes dans le monde, devient plus importante avec le vieillissement des populations. Notre étude, utilisant les données de l'Étude longitudinale canadienne sur le vieillissement (ELCV), visait à regrouper les pa-

time, identifying patterns linked to dementia subtypes and exploring demographic influences on these clusters through regression-adjusted cognitive measures. We analyzed cognitive, combined cognitive-demographic data, and adjusted cognitive measurements at baseline and follow-up, employing k-means, Partition Around Medoids, and Hierarchical clustering. Cluster quality was evaluated using the Silhouette coefficient, Dunn Index, and Adjusted Rand Index (ARI). Findings suggest demographic data's subtle yet significant impact on clustering, especially after adjusting cognitive measurements, indicating the importance of demographic factors in dementia patient clustering. This approach may improve care strategies by acknowledging demographic variations.

tients selon les changements cognitifs dans le temps, identifiant des motifs liés aux sous-types de démence et explorant les influences démographiques sur ces groupes via des mesures cognitives ajustées par régression. Nous avons analysé des données cognitives et démographiques combinées, et mesures cognitives ajustées au départ et au suivi, utilisant k-means, Partition Around Medoids et clustering hiérarchique. La qualité des groupes a été évaluée avec le coefficient Silhouette, l'indice de Dunn et l'Indice Rand Ajusté (IRA). Les résultats suggèrent un impact subtil mais significatif des données démographiques sur le regroupement, surtout après ajustement des mesures cognitives, soulignant l'importance des facteurs démographiques dans le regroupement des patients atteints de démence. Cette approche pourrait améliorer les stratégies de soins en reconnaissant les variations démographiques.

[16:30-16:45]

Alex Stringer (University of Waterloo) **Jeffrey Negrea** (University of Waterloo)

Testing Variance Components the Easy Way

Tester les composantes de la variance en toute simplicité

We present a methodology for estimating variance components in linear mixed models that accommodates testing general linear hypotheses on the components. This allows, for example, testing that two groups have the same variance and testing that a subset of variance components are zero, among many other applications. We derive an accurate approximation to the null distribution of the normalized residual likelihood ratio statistic for an arbitrary number of components as well as an algorithm for fast sampling from its exact finite-sample distribution. Because splines fit by penalized regression can be written as linear mixed models, all of our results apply to semi-parametric models fit with penalized splines, and this is discussed and results presented. In particular, we can test whether spline curves are zero, polynomial, equal to each other, and/or equally smooth. This is joint work with Jeffrey Negrea.

Nous présentons une méthode d'estimation des composantes de la variance dans les modèles mixtes linéaires qui permet de tester des hypothèses linéaires générales sur les composantes. Cela permet, par exemple, de tester que deux groupes ont la même variance et de tester qu'un sous-ensemble de composantes de la variance est nul, parmi de nombreuses autres applications. Nous obtenons une approximation précise de la distribution sous l'hypothèse nulle de la statistique du rapport de vraisemblance résiduel normalisé pour un nombre arbitraire de composantes, ainsi qu'un algorithme d'échantillonnage rapide à partir de sa distribution exacte sur échantillon fini. Comme les splines ajustées par régression pénalisée peuvent être écrites comme des modèles mixtes linéaires, tous nos résultats s'appliquent aux modèles semi-paramétriques ajustés avec des splines pénalisées, ce dont nous discutons avec présentation des résultats. En particulier, nous pouvons tester si les courbes splines sont nulles, polynomiales, égales entre elles et/ou également lisses. Ce travail a été réalisé en collaboration avec Jeffrey Negrea.

[16:45-17:00]

Kelly Ramsay (York University) **Shojaeddin Chenouri** (University of Waterloo)

Changepoint Detection in the Variability of Multivariate and Functional Data

Détection de points de changement dans la variabilité des données fonctionnelles et multivariées

We consider the problem of robustly detecting changepoints in the variability of a sequence of independent multivariate functions and vectors. We present novel changepoint procedures, called the functional and multivariate Kruskal–Wallis for covariance (FKWC and

Nous étudions le problème de la détection robuste des points de changement dans la variabilité d'une séquence de fonctions et vecteurs multivariés indépendants. Nous présentons de nouvelles procédures de points de changement appelées procédures de points de changement des covariables fonctionnelles et multivariées de

MKWC) changepoint procedures, based on rank statistics and data depth. The MKWC and FKWC changepoint procedures allow the user to test for at most one changepoint or an epidemic period, or to estimate the number and locations of an unknown amount of changepoints in the data. We show that when the "signal-to-noise" ratio is bounded below, the changepoint estimates produced by the MKWC and FKWC procedures attain the minimax localization rate for detecting general changes in distribution in the univariate setting. We also provide the behavior of the proposed test statistics for the AMOC and epidemic setting under the null hypothesis, and, as a simple consequence of our main result, these tests are consistent.

Kruskal-Wallis (FKWC et MKWC), basées sur les statistiques de rang et la profondeur de données. Les procédures de points de changement MKWC et FKWC permettent à l'utilisateur de tester au plus un changement (AMOC) ou une période épidémique, ou d'estimer le nombre et les emplacements d'une quantité inconnue de points de changement dans les données. Nous montrons que si le ratio « signal sur bruit » est délimité par le bas, l'estimation des points de changement produite par les procédures MKWC et FKWC atteint le taux de localisation minimax pour la détection des changements généraux de la distribution dans le contexte univarié. Nous présentons également le comportement statistique des tests proposés dans le contexte AMOC et épidémique en fonction d'une hypothèse nulle et la cohérence de ces tests comme simple conséquence de notre résultat principal.

Chair/Président: Tharshanna Nadarajah

Room/Salle: C 3033

Date: Tuesday June 4 / mardi 4 juin

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-15:45]

Jesse Ghashti (University of British Columbia - Okanagan) **Jeffrey Andrews** (University of British Columbia) **John R.J. Thompson** (University of British Columbia)

A Bootstrap Augmented k-means Algorithm for Fuzzy Partitions

Algorithme à K moyennes augmenté pour les partitionnements de données diffus

Fuzzy c-means (FCM) algorithms partition data into probabilistic cluster assignments by selecting a so-called fuzzy parameter. Although FCM is built on the efficient and straightforward framework of k-means clustering, its drawbacks include the required a priori knowledge of fuzziness to select the optimal fuzzy parameter, potential local optima entrapment, and sensitivity to initial cluster centres. In this talk, we present a bootstrap augmented k-means clustering algorithm that incorporates bootstraps into the loss function optimization scheme, allowing for probabilistic cluster assignments without tuning parameters and reducing the impact of random initializations with iterated refinement of cluster centres. We show that the proposed algorithm more accurately models the preordained uncertainty of cluster allocations for simulated data. We demonstrate, under satisfied model assumptions, that this augmented algorithm will mathematically match or exceed the performance of FCM variants.

Les algorithmes des c-moyennes partitionnent les données en groupes probabilistes en sélectionnant un paramètre diffus. Bien que les c-moyennes reposent sur le cadre efficace et simple du partitionnement des K moyennes, elles présentent des inconvénients, dont la nécessité d'une connaissance préalable du partitionnement diffus pour sélectionner le paramètre flou optimal, le risque de piégeage des valeurs optimales locales et la sensibilité aux centres de grappes initiaux. Dans cette présentation, nous proposons un algorithme à k-moyennes augmenté de bootstrap qui intègre des bootstraps dans le schéma d'optimisation de la fonction de perte, ce qui permet des attributions probabilistes de grappes sans réglage des paramètres et réduit l'impact des initialisations aléatoires grâce à une amélioration itérée des centres de grappes. Nous montrons que l'algorithme proposé modélise plus précisément l'incertitude préétablie des attributions de grappes de données simulées. Nous démontrons, à partir d'hypothèses de modèles fiables, que cet algorithme augmenté atteindra ou dépassera mathématiquement l'efficacité des variantes de la méthode des c-moyennes.

[15:45-16:00]

Ladan Tazik (University of British Columbia) **W. John Braun** (University of British Columbia)

Local Polynomial Lp Norm Regression

Régression polynomiale locale de la norme Lp

The local least-squares estimator for a regression curve cannot provide optimal results when non-Gaussian noise is present. Theoretically and empirically, there is evidence that residuals often exhibit distributional properties different from that of a normal distribution, so it is worth considering estimation based on other norms. It is suggested to use Lp-norm estimators to minimize the residuals when residuals have non-normal kurto-

L'estimateur local des moindres carrés pour une courbe de régression ne peut pas fournir des résultats optimaux en présence d'un bruit non gaussien. Sur le plan théorique et empirique, il semble que les résidus montrent souvent des propriétés distributionnelles différentes de celles d'une distribution normale. Par conséquent, il est bon de considérer une estimation basée sur d'autres normes. On suggère d'utiliser les estimateurs de la norme Lp pour minimiser les résidus, lorsque ces derniers ont un co-

sis. We propose local polynomial Lp-norm regression, which replaces weighted least-square estimation with weighted Lp-norm estimation for fitting the polynomial locally. We also introduce a new method for estimating the parameter p from the residuals, enhancing the adaptability of the approach. Through numerical and theoretical investigation, we demonstrate our method's superiority over local least squares in one-dimensional data and promising outcomes for higher dimensions, specifically 2D.

efficient d'acuité (kurtosis) non normalisé. Nous proposons la régression polynomiale locale de la norme Lp qui remplace l'estimation des moindres carrés pondérés par une estimation de la norme pondérée Lp pour ajuster la régression polynomiale localement. Nous présentons également une méthode pour l'estimation du paramètre p des résidus, ce qui renforce l'adaptabilité de l'approche. À l'aide d'une enquête numérique et théorique, nous montrons la supériorité de notre méthode sur celle des moindres carrés locaux pour des données unidimensionnelles et faisons état de résultats prometteurs pour les plus grandes dimensions, plus précisément 2D.

[16:00-16:15]

Mohammad Kaviul Anam Khan (University of Toronto) **Rafal Kustra** (Dalla Lana School of Public Health, University of Toronto) **Olli Saarela** (Dalla Lana School of Public Health, University of Toronto)

Conditional Permutation based on Generalized Variable Importance Metric and it's Relation to Causal Inference

Permutation conditionnelle basée sur l'importance de variable généralisée et son lien à l'inférence causale

In our previous study, we developed a model-agnostic variable importance metric called the Generalized Variable Importance Metric (GVIM), based on Breiman's variable importance metric for random forests. GVIM was shown to be a function of conditional average treatment effect. However, when estimated from tree-based methods, the estimates were biased, specifically when there was strong correlation among predictors. To address this issue, we developed a new GVIM based on the conditional distribution of a chosen predictor over all the other predictors. This metric was decomposed as a function of conditional average treatment effect. We further showed the bias-variance decomposition of GVIM and the statistical properties of the GVIM estimator using multiple simulations.

Dans notre étude précédente, nous avons développé une mesure d'importance de variable à modèle agnostique appelée « mesure d'importance de variable généralisée » (GVIM) basée sur la mesure d'importance de variable de Breiman pour les forêts aléatoires. GVIM est une fonction d'effet de traitement moyen conditionnelle. Cependant, les estimations sont biaisées lorsqu'elles sont réalisées à partir de méthodes basées sur des arbres, spécifiquement lorsqu'il y avait une forte corrélation entre les prédicteurs. Pour résoudre ce problème, nous avons conçu une nouvelle GVIM fondée sur la distribution conditionnelle d'un prédicteur sélectionné avant tous les autres prédicteurs. Cette mesure a été décomposée en guise de fonction d'effet de traitement moyen conditionnelle. Enfin, nous démontrons la décomposition biais-variance de la GVIM et les propriétés statistiques de l'estimateur GVIM au moyen de plusieurs simulations.

[16:15-16:30]

Thimani Dananjana Ranathungage (University of Manitoba) **Sulalitha Bowala** (University of Manitoba) **Md. Erfanul Hoque** (Thompson Rivers University) **Aerambamoorthy Thavaneswaran** (University of Manitoba) **Ruppa Thulasiram** (University of Manitoba)

Application of a Novel Fuzzy Pattern Mining Algorithm for Sequence Data

Application d'un nouvel algorithme d'exploration de modèles flous pour des données de séquences

Recently, there has been a growing interest in using Markov chain (MC) models in studying patterns in sequence data such as DNA sequences. The process of identifying underlying processes for specified patterns is known as pattern mining. The purpose of pattern mining in large DNA data sets involves, capturing and comparing trends that have previously been associated with observations such as diseases and detecting malicious events in DNA sequences. Our objectives in this

Les modèles par chaîne de Markov (MC) ont récemment fait l'objet d'un intérêt croissant pour l'étude des données de séquences, comme celles du séquençage de l'ADN. L'exploration de modèle est un processus d'identification des processus sous-jacents à des modèles spécifiés. Le but de l'exploration de modèle dans de grands ensembles de données sur l'ADN est notamment de capturer et comparer des tendances qui ont été précédemment associées à des observations, telles que des maladies et la détection d'événements malveillants dans des séquences d'ADN. Les ob-

study include identifying a desired nucleotide pattern in a DNA sequence and where it appears in the sequence. A novel fuzzy transition probability (TP) matrix is introduced, and a novel pattern mining algorithm is proposed for sequence data of any length. The proposed algorithm, which avoids the inversion of the pattern matrix, applies to Markov chains with large state spaces. DNA sequence data with 3954 base pairs is studied by using the proposed algorithm and obtained expected waiting time for patterns of interest.

jectifs de notre étude sont notamment d'identifier un modèle de nucléotide souhaité dans une séquence d'ADN et son emplacement dans la séquence. Nous présentons une nouvelle matrice de probabilité de transition floue (TP) et proposons un nouvel algorithme d'exploration de modèle pour des données de séquences de toutes les longueurs. En plus d'éviter l'inversion de la matrice du modèle, l'algorithme proposé s'applique aux chaînes de Markov avec de grands espaces d'état. Des données de séquençage de l'ADN avec 3 954 paires de bases sont étudiées à l'aide de l'algorithme proposé et des temps d'attente prévus obtenus pour les modèles d'intérêt.

[16:30-16:45]

Sarah Organ (Dalhousie University) **Hong Gu** (Dalhousie University) **Toby J. Kenney** (Dalhousie University)

Vertex Cover Matroid Variable Selection for Controlling the False Discovery Rate and Improving Power With Correlated Predictors

Sélection de variables matroïdes de couverture par sommets pour contrôler le taux de fausses découvertes et améliorer la puissance avec des prédicteurs corrélés

Variable selection methods struggle with controlling for false discovery (FDR) while maintaining a high power when the variables are correlated. To address this problem, we rethink variable selection to allow for the selection of surrogate pairs of variables. By selecting surrogate pairs, we are considering that if the pairs of variables are highly correlated, we do not know which of the variables is a true variable, therefore choosing either variable is appropriate. This approach allows us to overcome the problems multicollinearity introduces to standard variable selection. One of the challenges we overcome in this method is how to measure true and false positive rates for methods that select choices of surrogates, rather than picking a single variable. By utilizing our chosen algorithm to measure true and false positives, simulations show our two-stage method maintains an FDR less than 5% while achieving higher power than existing methods when the correlation is greater than 0.5.

Les méthodes de sélection des variables éprouvent des difficultés à contrôler les fausses découvertes (FDR) tout en conservant une puissance élevée lorsque les variables sont corrélées. Pour répondre à ce problème, nous repensons la sélection des variables afin de permettre la sélection de paires de variables substitutives. Cette sélection nous amène à considérer que si les paires de variables sont très corrélées, nous ignorons laquelle des variables est une vraie variable, et par conséquent le choix de l'une ou l'autre est approprié. Cette approche nous permet de résoudre les problèmes que la multicolinéarité entraîne pour la sélection standard de variables. Une des difficultés surmontées avec cette méthode est la façon de mesurer les vrais et faux taux positifs pour les méthodes avec des choix de variables substitutives plutôt qu'une seule variable. À l'aide de l'algorithme choisi pour mesurer les vrais et faux taux positifs, des simulations montrent que notre méthode à deux phases maintient un taux de fausses découvertes inférieur à 5 %, tout en produisant une plus grande puissance que les méthodes existantes lorsque le corrélation est supérieure à 0,5.

[16:45-17:00]

Tia Der (The University of British Columbia) **John R.J. Thompson** (The University of British Columbia)

Iterative Mean-Shift Clustering for Change-Point Regression Estimation

Regroupement itératif par déplacement de la moyenne pour l'estimation de la régression des points de changement

In this talk, we discuss the challenges in estimating noisy regression functions with abrupt changes or "change-points", where noise surrounding change-points causes errors and overweighting during estimation. We propose an iterative mean-shift clustering approach to improve function estimation around change-

Dans cette présentation, nous abordons les défis posés par l'estimation de fonctions de régression bruitées avec des changements abrupts ou « points de changement », lorsque le bruit autour des points de changement provoque des erreurs et une surpondération lors de l'estimation. Nous proposons une approche itérative de regroupement par déplacement de la moyenne pour

points by reweighting variables based on local distributional information. This is achieved through a data sharpening approach, which uses kernels to coalesce similar points to local modes and separate dissimilar points. We show that this method does not assume the shape of the underlying data-generating process or the location and number of change-points present. We explore the reweighting effects on change-point regression function estimator convergence, including the cost to rate of convergence away from change-points and boundaries. Through simulated non-linear data, we find improvements in estimating change-point shapes across many change-point regression estimation methods.

améliorer l'estimation des fonctions autour des points de changement en repondérant les variables à partir de données locales sur la répartition. Pour ce faire, nous recourons à une méthode d'affinage des données, qui utilise des noyaux pour regrouper les points similaires en modes locaux et séparer les points dissemblables. Nous démontrons que cette méthode ne tient pas compte de la forme du processus sous-jacent de génération des données, ni de l'emplacement ou du nombre de points de changement présents. Nous explorons les effets de la repondération sur la convergence de l'estimateur de la fonction de régression des points de changement, ainsi que le coût du taux de convergence loin des points de changement et des limites. Grâce à l'utilisation de données non linéaires simulées, nous constatons des améliorations dans l'estimation des formes des points de changement pour de nombreuses méthodes d'estimation de la régression des points de changement.

NSERC Discovery Grants Information Session
Séance d'information sur les subventions à la découverte de CRSNG

Chair/Président: Saman Muthukumarana

Organizer/Responsable: Saman Muthukumarana

Room/Salle: A 1049

Date: Tuesday June 4 / mardi 4 juin

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-17:00]

Adele Ngi-Song (NSERC)

NSERC Discovery Grants Information Session

Séance d'information sur les subventions à la découverte de CRSNG

This workshop will be presented by NSERC Research Grants staff and will cover the Notification of Intent to Apply (NOI) and Full Application process, the Discovery Grant evaluation process principles (criteria and ratings), the Conference Model and tips for preparing a Discovery Grant application. Following the Workshop, there will be an opportunity for participants to ask questions.

Cet atelier, présente par le personnel des subventions à la découverte du CRSNG, couvrira l'Avis d'intention de présenter une demande de subvention à la découverte et le processus de demande détaillée, les principes du processus d'évaluation des subventions à la découverte (critères et cotes), le modèle de conférence et présentera certains conseils pour la préparation d'une demande de subvention à la découverte. À la fin de l'atelier, les participants seront invités à poser leurs questions.

LaTeX class presentation (Canadian Journal of Statistics)
Présentation de la classe LaTeX (La revue canadienne de statistique)

Chair/Président: Johanna G. Nešlehová

Organizer/Responsable: Johanna G. Nešlehová

Room/Salle: SN 2109

Date: Tuesday June 4 / mardi 4 juin

Time/Heure: 17:00-18:00

Abstract/Résumé

[17:00-18:00]

Vincent Goulet (Université Laval)

Introducing the new class for authors of The Canadian Journal of Statistics: cjs-rcs-article

Introduction de la nouvelle classe pour les auteurs de La revue canadienne de statistique : cjs-rcs-article

The Statistical Society of Canada maintains its own LaTeX class to typeset The Canadian Journal of Statistics/La revue canadienne de statistique. The latest version of the class 'TD-CJS' that you may well know is dated 13 July... 1994. A lot of water went under the bridge in the LaTeX world since then. The class and bibliography styles were outdated to the point of causing compilation problems for recent papers. In late 2022, I was commissioned to develop a class that would update not only the production underpinnings, but also the visuals of The CJS. This special presentation will be the opportunity to unveil the new class 'cjs-rcs-article', give a tour of its features, and explain how to quickly get started for your next article!

La Société statistique du Canada maintient sa propre classe LaTeX pour la composition de The Canadian Journal of Statistics/La revue canadienne de statistique. La dernière version de la classe « TD-CJS » que vous connaissez peut-être est datée du 13 juillet... 1994. Beaucoup d'eau a coulé sous les ponts dans le monde LaTeX depuis lors. La classe et les styles de bibliographie étaient dépassés au point de causer des problèmes de compilation pour les articles récents. Fin 2022, j'ai été chargé de développer une classe qui mettrait à jour non seulement les fondements de la production, mais aussi l'aspect visuel de la RCS. Cette présentation spéciale sera l'occasion de dévoiler la nouvelle classe « cjs-rcs-article », de faire un tour d'horizon de ses fonctionnalités, et d'expliquer comment rapidement démarrer votre prochain article !

SSC 2023 Gold Medal Address
Allocution du récipiendaire de la Médaille d'or de la SSC 2023

Chair/Président: Grace Y. Yi

Organizer/Responsable: Grace Y. Yi

Room/Salle: IIC 2001

Date: Wednesday June 5 / mercredi 5 juin

Time/Heure: 08:30-09:30

Abstract/Résumé

[08:30-09:30]

Charmaine B. Dean (University of Waterloo)

Optimizing research impact through interdisciplinary and collaborative research

Optimiser l'impact de la recherche par la recherche interdisciplinaire et collaborative

Interdisciplinary collaborative research is a key component of data science and for some of us, plays an important part of our roles as statisticians. It is not unusual that we become accustomed to vertical thinking whereby we use existing tools and methods in our own specialty to problem solve, losing sight of the larger interdisciplinary context of data science, and the context of the scientific challenge. The Government of Canada - Science and Technology branch has identified several key priority research challenge topics that involve cross-disciplinary work. Although statistical tools and analytics are identified in these research challenge priority areas, additionally, the development of fundamental transformative and enabling technological tools specifically for statistical methods and analytics to support research and societal advancement is also seen as a priority. This talk shares insights about the challenges and opportunities for statistics in interdisciplinary research. Specifically, monitoring viral signals in wastewater and assessing forest fire risk are given as complex, case studies that use a collaborative and interdisciplinary approach to solve difficult problems. This approach will demonstrate the significant benefits for not only optimizing research impact but for training students to become horizontal problem solvers across a wide range of research methods which will benefit them in navigating complex problems and in the development of appropriate tools for their analysis.

La recherche collaborative interdisciplinaire est un élément clé de la science des données qui, pour certains d'entre nous, joue un rôle important dans notre métier de statisticien. Il n'est pas rare que nous nous habituions tellement à penser verticalement et à résoudre nos problèmes avec les seuls outils et méthodes de notre propre spécialité que nous perdons de vue le contexte interdisciplinaire plus large de la science des données, ainsi que le contexte du défi scientifique. La Direction générale des sciences et de la technologie du Gouvernement du Canada a identifié plusieurs thèmes de recherche prioritaires qui impliquent un travail interdisciplinaire. Bien que les outils statistiques et l'analytique y soient identifiés, une autre priorité est de développer des outils technologiques fondamentaux, transformateurs et habilitants, spécifiquement conçus pour les méthodes statistiques et l'analytique, afin de soutenir la recherche et le progrès sociétal. Nous présenterons ici un aperçu des défis et des opportunités qui existent pour la statistique dans la recherche interdisciplinaire. Plus précisément, nous explorerons la surveillance des signaux viraux dans les eaux usées et l'évaluation des risques d'incendie de forêt comme des études de cas complexes qui s'appuient sur une approche collaborative et interdisciplinaire pour résoudre des problèmes difficiles. Cette approche présente des avantages significatifs non seulement pour optimiser l'impact de la recherche mais aussi pour former les étudiants à devenir des résolveurs de problèmes horizontaux avec à leur disposition un large éventail de méthodes de recherche, ce qui leur permettra de gérer des problèmes complexes et de développer des outils appropriés à leur analyse.

Presenting of Student Research Presentation and Case Study Awards
Remise des prix pour les présentations de recherche étudiantes et d'études de cas

Chair/Président: Shirley E. Mills

Organizer/Responsable: Shirley E. Mills

Room/Salle: IIC 2001

Date: Wednesday June 5 / mercredi 5 juin

Time/Heure: 09:30-09:50

Abstract/Résumé

[09:30-09:50]

Pingzhao Hu (Western University)
Student Research Presentation Awards
Prix pour les présentations de recherche étudiantes

. .

[09:30-09:50]

Chel Hee Lee (Alberta Health Services)
Case Studies in Data Analysis Awards
Prix d'études de cas en analyse de données

. .

Chair/Président: Joan X. Hu

Organizer/Responsable: Joan X. Hu

Room/Salle: C 2045

Date: Wednesday June 5 / mercredi 5 juin

Time/Heure: 10:20-11:50

Abstract/Résumé

[10:20-10:42]

Naisyin Wang (University of Michigan)

Utilizing Synthetic Components to Balance Privacy Protection and Data Utility

Utilisation de composants synthétiques pour équilibrer la protection de la vie privée et l'utilité des données

The importance of privacy protection is rising in the current practice of publicly sharing data. Different evaluation criteria in terms of privacy protection and data utilities are considered. They may or may not agree with each other. In this presentation, we illustrate ways to utilize synthetic components to balance privacy protection and data utilities for different evaluation criteria. One additional aim is to enable users easily implement statistical analysis using publicly shared data after the observations are processed with such privacy protection procedures. The efficacy and quality of the proposed procedures are illustrated theoretically and numerically via applications to biomedical datasets.

La protection de la vie privée revêt de plus en plus d'importance face à la pratique actuelle de partage public des données. Nous explorons différents critères d'évaluation en termes de protection de la vie privée et d'utilité des données, qui peuvent ou non être en accord les uns avec les autres. Dans cette présentation, nous illustrons différentes façons d'utiliser des composants synthétiques pour équilibrer la protection de la vie privée et l'utilité des données pour différents critères d'évaluation. Nous souhaitons aussi que les utilisateurs puissent facilement réaliser des analyses statistiques à partir de données partagées publiquement, une fois les observations traitées via de telles procédures de protection de la vie privée. Nous illustrons l'efficacité et la qualité des procédures proposées théoriquement et numériquement, par des applications à des ensembles de données biomédicales.

[10:42-11:05]

Jie Chen (Augusta University)

Linking Genomic Features to the Survival Time of GBM Cancer Patients

Lier les caractéristiques génomiques au temps de survie de patients atteints de cancer GBM

We focus on how to jointly analyze multiple types of genomics data, along with prognostic information, available within and across different studies. It has been a challenging and common task in modern statistical research to use all types of data to infer disease-prone genetic information and to link those features to cancer survival. We modelled the genomic, prognostic and survival datasets under a framework of an accelerated failure time with frailty (AFTF) to infer patients' survival time. Simulation results confirmed the good performance of the approach. The approach was applied to the analysis of the Cancer Genome Atlas (TCGA)

Nous nous concentrons à découvrir comment analyser conjointement plusieurs types de données génomiques, ainsi que des renseignements pronostiques, offerts à travers différentes études. Il est à la fois difficile et fréquent dans la recherche statistique moderne d'utiliser tous les types de données dans le but d'inférer l'information génétique sensible à la maladie et de faire le lien entre ces caractéristiques et la survie au cancer. Nous avons modélisé les ensembles de données génomiques, pronostiques et de survie selon un cadre de temps de défaillance accéléré avec fragilité (accelerated failure time with frailty, AFTF) afin d'inférer le temps de survie des patients. Les résultats de simulations confirment la bonne performance de l'approche. L'approche a été appliquée à

multiple genomic datasets of Giloblastoma Multiforme (GBM), a lethal brain cancer, and interesting genomic features are identified and biological interpretations are explored. This talk is based a joint work with S. Deng.

l'analyse des ensembles de données génomiques du glioblastome multiforme (GBM, un cancer fatal du cerveau) tirés de l'Atlas du génome du cancer. Nous y avons identifié des caractéristiques génomiques intéressantes et avons exploré des interprétations biologiques. Cet exposé est basé sur un travail conjoint avec S. Deng.

[11:05-11:27]

Subharup Guha (University of Florida) **Yi Li** (University of Michigan)

Causal Meta-Analysis by Integrating Multiple Observational Studies with Multivariate Outcomes

Méta-analyse causale par l'intégration de plusieurs études d'observation avec des résultats multivariés

Integrating multiple observational studies to make unconfounded causal comparisons of group potential outcomes in a large natural population is challenging. Moreover, retrospective cohorts, being convenience samples, are usually unrepresentative of the population of interest and have groups with unbalanced covariates. We propose a general covariate-balancing framework based on pseudo-populations that extends established weighting methods to the meta-analysis of multiple retrospective cohorts with multiple groups. By maximizing the effective sample sizes of the cohorts, we propose a FLEXible, Optimized, and Realistic (FLEXOR) weighting method appropriate for integrative analyses. We develop new estimators for unconfounded inferences and examine their asymptotic properties. Through simulation studies and meta-analyses of TCGA datasets, we demonstrate the versatility and reliability of the weighting strategy, especially for the FLEXOR pseudo-population.

Il est difficile d'intégrer plusieurs études d'observation dans le but d'effectuer des comparaisons causales valides des résultats potentiels des groupes dans une grande population naturelle. De plus, les cohortes rétrospectives, qui sont des échantillons de convenance, ne sont pas toujours représentatives de la population étudiée et comportent des groupes dont les covariables ne sont pas équilibrées. Nous suggérons un cadre général d'équilibrage des covariables en fonction des pseudo-populations qui étend les méthodes de pondération établies pour la méta-analyse de plusieurs cohortes rétrospectives avec plusieurs groupes. Nous proposons une méthode de pondération flexible, optimisée et réaliste (FLEXOR) qui maximise la taille effective des échantillons des cohortes et qui est adaptée aux analyses intégratives. Nous créons de nouveaux estimateurs pour les inférences sans biais de confusion et examinons leurs propriétés asymptotiques. Au moyen d'études de simulation et de méta-analyses d'ensembles de données de l'Atlas du génome du cancer, nous démontrons la polyvalence et la fiabilité de la stratégie de pondération, en particulier pour la pseudo-population de FLEXOR.

[11:27-11:50]

Rui Wang (Harvard Pilgrim Health Care Institute) **Chia-Rui Chang** (Harvard University) **Yue Song** (Harvard University) **Fan Li** (Duke University)

Covariate Adjustment in Randomized Clinical Trials with Missing Covariate and Outcome Data

Ajustement de covariables dans des essais cliniques randomisés avec covariables et données de résultats manquantes

When analyzing data from randomized clinical trials, covariate adjustment can be used to account for chance imbalance in baseline covariates and to enhance the precision of the treatment effect estimate. A practical barrier to implementing covariate adjustment is the presence of missing data. We investigate the implications of the missing data mechanism on the estimation of the average treatment effect in randomized clinical trials. We propose a weighting approach that combines inverse probability weighting for adjusting missing outcomes and overlap weighting for covariate adjustment. We conduct simulation studies to examine the finite sample performance of the proposed methods. We find that

Lors de l'analyse de données tirées d'essais cliniques randomisés, l'ajustement des covariables peut servir à tenir compte d'un déséquilibre dans les covariables de base et à hausser la précision de l'estimation d'effet de traitement. Une barrière pratique pour implanter l'ajustement de covariable signifie qu'il y a des données manquantes. Nous étudions les conséquences du mécanisme de données manquantes sur l'estimation de l'effet de traitement moyen dans les essais cliniques randomisés. Nous proposons une approche de pondération qui combine la pondération de probabilité inversée pour ajuster les résultats manquants et la pondération de chevauchement pour l'ajustement de covariables. Nous menons des études en simulations pour évaluer la performance d'échantillons finis des méthodes proposées. Nous

the proposed adjustment methods generally improve the precision of treatment effect estimates, irrespective of the imputation methods, when the adjusted covariate is associated with the outcome.

découvrons que les méthodes d'ajustement proposées améliorent généralement la précision des estimations d'effet de traitement, indépendamment des méthodes d'imputation, lorsque les covariables ajustées sont associées au résultat.

Novel Spatiotemporal Models for Complex Data in Fisheries and Ecosystem Studies
Nouveaux modèles spatio-temporels pour les données complexes des études sur les pêcheries et écosystèmes

Chair/Président: Noel Cadigan

Organizer/Responsable: Noel Cadigan, Nan Zheng, Asokan Mulayath Variyath

Room/Salle: A 1045

Date: Wednesday June 5 / mercredi 5 juin

Time/Heure: 10:20-11:50

Abstract/Résumé

[10:20-10:50]

Raphael Robert McDonald (Dalhousie University) **David Keith** (Fisheries and Oceans Canada) **Jessica Sameoto** (Fisheries and Oceans Canada) **Joanna Elizabeth Mills Flemming** (Dalhousie University)

Improving Spatio-Temporal Stock Assessment Models Through the Inclusion of Habitat Features and Drop-Camera Surveys

Améliorer les modèles spatio-temporels d'évaluation des stocks en incluant les caractéristiques de l'habitat et les sondages par caméras lestées

Stock assessment models aim to get reliable estimates of population abundance for the provision of science advice to fisheries managers. Many of these models are strong at dealing with various types of autocorrelations, but often underutilize alternative data (e.g., drop-camera surveys and seascape maps) to help improve their prediction and estimation capacities. Furthermore, how to include these novel sources of information is an open question. We focus here on the development of a spatially explicit habitat based assessment model (SEHBAM) that accounts for habitat through the inclusion of seascape maps using the Maritimes Inshore Sea Scallop as a case study. Results show improvements in estimating probabilities of encounter and more realistic catchabilities, providing an improved understanding of population dynamics. Finally, methods for the further inclusion of a drop-camera survey are proposed to account for the different resolution and precision associated with these new data.

Les modèles d'évaluations des stocks ont comme objectif d'obtenir des estimés d'abondance de population pour la provision d'avis scientifiques aux gérants de pêches. Même si plusieurs de ces modèles ont la capacité d'inclure plusieurs formes d'autocorrélations, plusieurs sources d'informations externes (e.g., sondages par caméras lestées et carte des paysages marins) restent sous-utilisées et leur inclusion est une question ouverte. Nous présentons le développement d'un modèle d'évaluation basé sur l'habitat et explicitement spatial qui utilise une carte des paysages marins à travers une étude de cas sur la pêche pour le pétoncle de mer dans les Maritimes. Nos résultats démontrent une amélioration dans l'estimation des probabilités de capture ainsi que des meilleurs capturabilités, par le fait-même améliorant notre compréhension des dynamiques de populations. Finalement, nous proposons des nouvelles méthodes pour l'inclusion d'un sondage par caméra lestée à l'intérieur du même modèle.

[10:50-11:20]

James Thorson (Alaska Fisheries Science Center)

Including Ecological Mechanism in Spatio-temporal Analysis: Habitat Preferences and Structural Multivariate Spatio-temporal Models

Intégration de mécanismes écologiques dans l'analyse spatio-temporelle : préférences en matière d'habitat et modèles structurels spatio-temporels multivariés

Spatio-temporal models are widely used in fisheries science and management. However, climate forecasting requires mechanistic representation of ecosystem link-

Les modèles spatio-temporels sont largement utilisés dans la science et la gestion de la pêche. Cependant, les prévisions climatiques nécessitent une représentation mécaniste des liens entre

ages. I therefore discuss two avenues for greater mechanism in spatio-temporal models. The first extends structural equation model to include both lagged and simultaneous effects in multivariate time-series analysis (termed "dynamic structural equation models" DSEM). I illustrate DSEM using a simple model for ecosystem linkages in the eastern Bering Sea, and also illustrate a spatial extension by estimating associations between fishes and biogenic habitat (corals and sponges). The second incorporates advection towards preferred habitats (termed "taxis") and estimates diffusive-taxis movement within a continuous-time Markov chain (CTMC). I illustrate diffusive-taxis movement using archival tagging data for Pacific cod in the eastern Bering Sea, and encourage further integration of movement and SEM in spatio-temporal models.

les écosystèmes. Je discute donc de deux pistes pour améliorer les mécanismes des modèles spatio-temporels. La première consiste à étendre le modèle d'équation structurelle afin d'inclure des effets décalés et simultanés dans l'analyse de séries temporelles multivariées (appelée « modèle d'équation structurelle dynamique » ou DSEM). J'illustre le modèle DSEM à l'aide d'un modèle simple pour les liens entre les écosystèmes dans l'est de la mer de Béring, et j'illustre également une extension spatiale en estimant les associations entre les poissons et l'habitat biogénique (coraux et éponges). La seconde incorpore l'advection vers les habitats préférés (appelés « taxis ») et estime le mouvement diffusif-taxis dans une chaîne de Markov à temps continu. J'illustre le mouvement diffusif-taxis à l'aide de données de marquage d'archives pour la morue du Pacifique dans l'est de la mer de Béring, et j'encourage une intégration plus poussée du mouvement et de la modélisation par équations structurelles dans les modèles spatio-temporels.

[11:20-11:50]

Nan Zheng (Memorial University of Newfoundland) **Noel Cadigan** (Fisheries and Marine Institute of Memorial University of Newfoundland)

Enhancing Fisheries Stock Assessment: Spatiotemporal Modeling of Zero-Inflated Nonnegative Continuous Data using the Tweedie Distribution

Modélisation spatio-temporelle de données continues non négatives et avec excès de zéros à l'aide de la distribution Tweedie pour améliorer l'évaluation des stocks de pêche

Zero-inflated nonnegative continuous (ZINC) data frequently arise in fishery studies, and survey indices of population size are an important example. The Tweedie distribution is a promising choice for modeling the distribution of ZINC data. However, the Tweedie dispersion relationship (DR) is not general enough to cover some important forms such as quadratic dispersion. Also, conventional multivariate Tweedie distributions lack computational efficiency for spatiotemporal applications. In this research we extend the Tweedie distribution to include more flexible DRs and we introduce computationally efficient methods for implementing multivariate Tweedie distributions with practical correlation structures tailored for spatiotemporal modeling. Our approaches are simple to implement and require only the Tweedie probability density function, which is available in many statistical modeling packages such as R and TMB (Template Model Builder). We present simulation studies and real data analysis to quantify the effectiveness of these novel methodologies.

Les données continues non-négatives avec excès de zéros apparaissent souvent dans les études liées à la pêche, et les indices de taille de population des enquêtes en sont un exemple significatif. La distribution Tweedie est une solution prometteuse pour la modélisation de la distribution des données continues non-négatives avec excès de zéros. Cependant, la relation de dispersion de Tweedie n'est pas suffisamment générale pour prendre en compte certaines formes importantes, comme la dispersion quadratique. De plus, les distributions Tweedie multivariées conventionnelles ne sont pas assez efficaces pour les calculs dans les applications spatio-temporelles. Dans cette étude, nous étendons la distribution Tweedie pour inclure des relations de dispersion plus souples et nous proposons des méthodes de calcul efficaces pour mettre en œuvre des distributions Tweedie multivariées avec des structures de corrélation pratiques adaptées à la modélisation spatio-temporelle. Nos approches sont simples à mettre en œuvre et ne nécessitent que la fonction de densité de Tweedie, qui se trouve dans de nombreux logiciels de modélisation statistique, comme R et Template Model Builder (TMB). Nous présentons des études de simulation et des analyses de données réelles pour quantifier l'efficacité de ces nouvelles méthodologies.

Modelling Dependence of Multivariate Extremes
Modélisation de la dépendance des extrêmes multivariés

Chair/Président: Léo Belzile

Organizer/Responsable: Debbie J. Dupuis

Room/Salle: A 1049

Date: Wednesday June 5 / mercredi 5 juin

Time/Heure: 10:20-11:50

Abstract/Résumé

[10:20-10:50]

Natalia Nolde (University of British Columbia) **Vincenzo Coia** (BGC Engineering) **Harry Joe** (University of British Columbia)

Copula-Based Conditional Tail Indices

Indices de queue conditionnels basés sur des copules

We consider regression-like situations in which the response variable is assumed to be heavy-tailed, whose upper tail behaviour is characterized by regular variation with a tail index parameter measuring the heaviness of the tail. In this context, we study the behavior of the conditional tail index, the tail index of the distribution of the response variable given predictor variables, and a related new notion of the copula conditional tail index. Our results show that in a wide range of scenarios, the conditional tail index does not vary over the predictor space. Motivated by this observation, new parametric families of copula models are introduced, which allow for non-constant tail indices.

Nous considérons des situations de type régression dans lesquelles la variable réponse est supposée avoir une queue lourde, et dont le comportement de la queue supérieure est caractérisé par une variation régulière avec un paramètre d'indice de queue qui mesure la lourdeur de la queue. Dans ce contexte, nous étudions le comportement de l'indice de queue conditionnel, l'indice de queue de la distribution de la variable de réponse en fonction des variables prédictives, et une nouvelle notion connexe de l'indice de queue conditionnel de la copule. Nos résultats montrent que dans un grand nombre de scénarios, l'indice de queue conditionnel ne varie pas sur l'espace des prédicteurs. Motivés par cette observation, nous introduisons de nouvelles familles paramétriques de modèles de copules, qui permettent des indices de queue non constants.

[10:50-11:20]

Stanislav Volgushev (University of Toronto) **Michaël Lalancette** (Université du Québec à Montréal) **Alexander Ryabchenko** (University of Toronto) **Sebastian Engelke** (University of Geneva)

Learning Hüsler-Reiss Graphical Models Under Connectedness Constraints

Apprentissage de modèles graphiques de Hüsler-Reiss sous contraintes de connexité

Graphical models for Hüsler-Reiss distributions provide parsimonious and interpretable models for extremes in high dimensions. By definition, Hüsler-Reiss graphical models reside on connected graphs. Further, the precision matrix of a Hüsler-Reiss distribution has to satisfy certain constraints in order to be a valid parameter matrix. Yet, all the methods that are available to date for estimating sparse do not automatically respect those constraints necessitating ad-hoc approaches to post-processing estimators. In this talk, we present a penalization procedure that explicitly takes into account

Les modèles graphiques pour les distributions de Hüsler-Reiss fournissent des modèles parcimonieux et interprétables pour les extrêmes en grande dimension. Par définition, les modèles graphiques de Hüsler-Reiss résident sur des graphes connectés. En outre, la matrice de précision d'une distribution de Hüsler-Reiss doit satisfaire à certaines contraintes pour être une matrice de paramètres valide. Pourtant, les méthodes existant à ce jour pour l'estimation de la matrice de précision ne respectent pas automatiquement ces contraintes, si bien qu'il faut des approches ad-hoc pour les estimateurs de post-traitement. Dans cet exposé, nous présentons une procédure de pénalisation qui prend explicitement

the connectedness constraint of a Hüsler-Reiss graphical model and produces sparse estimators of a Hüsler-Reiss precision matrix that are automatically valid parameter matrices. We provide convergence guarantees for our rank-based estimators and discuss the application of our theoretical results to the problem of matrix completion for Hüsler-Reiss precision matrices.

en compte la contrainte de connexité d'un modèle graphique de Hüsler-Reiss et produit des estimateurs épars d'une matrice de précision de Hüsler-Reiss qui sont automatiquement des matrices de paramètres valides. Nous fournissons des garanties de convergence pour nos estimateurs basés sur le rang et discutons de l'application de nos résultats théoriques au problème de la complétion de la matrice pour les matrices de précision de Hüsler-Reiss.

[11:20-11:50]

Michaël Lalancette (Université du Québec à Montréal)

On Pairwise Interaction Multivariate Pareto Models and Score Matching

Sur les modèles de Pareto multivariés à interaction de deuxième ordre et le "Score Matching"

Multivariate Pareto distributions play an important role in multivariate extreme value theory as models for tail dependence. They also play a central role in continuous extremal graphical models. Pairwise interaction models are exponential family statistical models in which the multivariate dependence structure is entirely characterized by a number of pairwise dependence parameters. They are an efficient way to construct non-Gaussian graphical models in which structure learning and constrained parameter inference can be carried out naturally. In this talk, the intersection between multivariate Pareto and pairwise interaction models is discussed, as well as the principle of score matching and its role in fitting such models. Partially based on ongoing work with Frank Röttger.

Les lois de Pareto multivariées jouent un rôle important en théorie des valeurs extrêmes multivariée en tant que modèles de dépendance extrémale. Ils jouent également un rôle central dans la définition des modèles graphiques extrémaux continus. Les modèles à interaction de deuxième ordre sont des familles exponentielles de lois dans lesquelles la structure de dépendance multivariée est entièrement caractérisée par les paramètres de dépendance bivariée entre les paires de variables. Ils représentent une façon efficace de construire des modèles graphiques non normaux dans lesquels l'apprentissage structurel et l'inférence sous contrainte peuvent être réalisés naturellement. Dans cet exposé, l'intersection entre les modèles de Pareto multivariés et à interaction de deuxième ordre est discutée, ainsi que le principe du "score matching" et son rôle dans l'estimation de tels modèles. Basé en partie sur un travail en cours avec Frank Röttger.

Multi-state Modeling for the Analysis of Lifetime Data
Modélisation multi-états pour l'analyse des données de durée de vie

Chair/Président: Yildiz Yilmaz

Organizer/Responsable: Yildiz Yilmaz

Room/Salle: A 2071

Date: Wednesday June 5 / mercredi 5 juin

Time/Heure: 10:20-11:50

Abstract/Résumé

[10:20-10:50]

Yingwei (Paul) Peng (Queen's University)

An Additive Hazards Frailty Model with Semi-varying Coefficients

Un modèle de fragilité à risques additifs avec des coefficients semi-variables

Proportional hazards frailty models have been extensively investigated and used to analyze clustered and recurrent failure times data. However, the proportional hazards assumption in the models may not always hold in practice. In this talk, I will present an additive hazards frailty model with semi-varying coefficients, which allows some covariate effects to be time-invariant while other covariate effects to be time-varying. The time-varying and time-invariant regression coefficients are estimated by a set of estimating equations, whereas the frailty parameter is estimated by the moment method. The large sample properties of the proposed estimators are established. The finite sample performance of the estimators is examined by simulation studies. The proposed model and estimation are illustrated with an analysis of data from a rehospitalization study of colorectal cancer patients.

Les modèles de fragilité à risques proportionnels ont été étudiés de manière approfondie et utilisés pour analyser les données sur les temps de défaillance groupés et récurrents. Cependant, l'hypothèse des risques proportionnels dans les modèles n'est pas toujours valable dans la pratique. Dans cet exposé, je présenterai un modèle de fragilité à risques additifs avec des coefficients semi-variables, qui permet à certains effets de covariables d'être invariants dans le temps tandis que d'autres effets de covariables varient dans le temps. Les coefficients de régression variables dans le temps et invariants dans le temps sont estimés par un ensemble d'équations d'estimation, tandis que le paramètre de fragilité est estimé par la méthode des moments. Les propriétés d'un grand échantillon des estimateurs proposés sont établies. La performance des estimateurs sur un échantillon fini est examinée par des études de simulation. Le modèle et l'estimation proposés sont illustrés par une analyse des données d'une étude de réhospitalisation de patients atteints d'un cancer colorectal.

[10:50-11:20]

Leilei Zeng (University of Waterloo) **Yidan Shi** (University of Pennsylvania)

A Mixture Hidden Markov Model for Multiple Types of Disease

Modèle de Markov caché à mélange pour plusieurs types de maladies

Multistate models are widely used for analyzing longitudinal data on disease progression over time. Many diseases manifest differently and what appears to be a coherent collection of symptoms is often the expression of a variety of distinct disease subtypes, each with a different rate of onset of symptoms and progression. We propose a mixture hidden Markov model (MHMM), where the underlying process is characterized by a finite mixture of multiple Markov chains, one for each

Les modèles multi-états sont largement utilisés pour analyser les données longitudinales sur l'évolution des maladies dans le temps. De nombreuses maladies se manifestent différemment et ce qui semble être un ensemble cohérent de symptômes est souvent l'expression d'une variété de sous-types de maladies distincts, chacun ayant une vitesse différente d'apparition des symptômes et de progression. Nous proposons un modèle de Markov caché à mélange, où le processus sous-jacent est caractérisé par un mélange fini de plusieurs chaînes de Markov, une pour chaque sous-type de

disease subtype, while the observation process contains states corresponding to the common symptomatic stages of these diseases. Information on type of disease is partially available and reflects the pathway through certain hidden states in the corresponding disease process, facilitating the estimation of parameters involved in the proposed models. The method is demonstrated on a dataset to model the development and progression of dementia caused by Alzheimer's disease and non-AD dementia.

maladie, tandis que le processus d'observation contient des états correspondant aux stades symptomatiques communs de ces maladies. Les informations sur le type de maladie sont partiellement disponibles et reflètent le cheminement à travers certains états cachés dans le processus pathologique correspondant, ce qui facilite l'estimation des paramètres impliqués dans les modèles proposés. Nous démontrons la méthode sur un ensemble de données pour modéliser le développement et la progression de la démence causée par la maladie d'Alzheimer et de la démence non Alzheimer.

[11:20-11:50]

Candemir Cigsar (Memorial University of Newfoundland) **Leila Torabi** (Memorial University of Newfoundland) **Zhaozhi Fan** (Memorial University of Newfoundland)

Quantile Regression for Sequentially Observed Bivariate Survival Data

Régression quantile (QR) pour des données de survie bivariées séquentiellement observées

Quantile regression (QR) offers a flexible way to assess the effects of covariates on the quantiles of the conditional distribution of a random variable, given covariates. Since the effects of covariates can be assessed at any quantile, QR provides a better understanding of the effects of covariates comparing with traditional regression models. In this study, we consider a parametric conditional QR model for survival data with time-fixed covariates, and introduce a multi-stage estimation procedure to estimate the effects of covariates on the quantiles of marginal distributions of sequentially observed bivariate survival times. We model the dependency between survival times with a Clayton copula. Our estimation method is based on the martingale estimating equations. We discuss asymptotic and finite sample properties of the estimators obtained from this procedure. Finally, the method is illustrated by analyzing a colon cancer dataset.

La régression quantile (RQ) offre un moyen souple d'évaluer les effets des covariables sur les quantiles de la distribution conditionnelle d'une variable aléatoire, pour les covariables déterminées. Comme les effets des covariables peuvent être évalués sur tout quantile, la régression quantile fournit une meilleure compréhension des effets des covariables, comparativement aux modèles de régression traditionnels. Notre étude s'intéresse au modèle de RQ conditionnelle paramétrique pour des données de survie avec des covariables avec effets fixes dans le temps et présentons une procédure d'estimation multi-étapes afin d'estimer les effets des covariables sur les quantiles de distributions marginales de temps de survie bivariés séquentiellement observés. Nous modélisons la dépendance entre les temps de survie avec une copule de Clayton. Notre méthode d'estimation est basée sur les équations d'estimation de la martingale. Nous discutons les propriétés asymptotiques et d'échantillons finis des estimateurs obtenus par cette procédure. Enfin, la méthode est illustrée par l'analyse des données sur le cancer du côlon.

Chair/Président: Elif Fidan Acar

Organizer/Responsable: Elif Fidan Acar

Room/Salle: A 2065

Date: Wednesday June 5 / mercredi 5 juin

Time/Heure: 10:20-11:50

Abstract/Résumé

[10:20-10:50]

Saman Muthukumarana (University of Manitoba)

Modelling Fish Movements and Determining Vulnerability to Fishing Effort in Lake Winnipeg Using Bayesian State-space Models

Modélisation des déplacements de poissons et détermination de la vulnérabilité de l'effort de pêche au lac Winnipeg au moyen de modèles spatiotemporels bayésiens

Bayesian state-space models coupled with acoustic telemetry can be used as a powerful tool for effective decision-making in fisheries management. In this study, we develop a Bayesian state-space modeling (BSSM) approach and different smoothing methods, including kernel smoothing and cross-validated local polynomial regression to reconstruct fish movement paths using data obtained from a telemetry receiver grid in Lake Winnipeg. Using the BSSM approach, we obtain more realistic movement paths compared to the smoothing methods. We then use the estimated fish movement paths and fishing information from Lake Winnipeg, such as the amount of landings and quota assignments to investigate the potential interactions between fish movements and fishing activities in the lake. The newly developed fishery metrics include the probability of fish presence, probability of encounter and potential fishing pressure, which are useful in making effective fishery management decisions on Lake Winnipeg.

Les modèles spatiotemporels bayésiens accompagnés de la télémétrie acoustique peuvent servir d'outils de taille pour prendre des décisions de façon efficace en gestion des pêches. Dans cette étude, nous développons une approche de modélisation spatiotemporelle bayésienne (BSSM) et plusieurs méthodes de lissage, comprenant le lissage de noyaux et la régression polynomiale locale par validation croisée servant à reconstruire les déplacements de poissons à l'aide de données obtenues d'un récepteur de télémesures au lac Winnipeg. Au moyen de l'approche BSSM, nous obtenons des trajets de déplacement plus réalistes par rapport à ceux obtenus avec les méthodes de lissage. Nous exploitons ensuite les trajets de déplacement des poissons et les renseignements relatifs à la pêche du lac Winnipeg (comme le nombre de débarquements et l'affectation des quotas) pour enquêter sur les interactions potentielles entre le déplacement des poissons et les activités de pêche dans le lac. Cette toute nouvelle mesure de pêche comprend : la probabilité de présence de poisson, la probabilité de rencontre et la pression potentielle de la pêche. Toutes des mesures utiles afin de prendre des décisions efficaces dans la gestion de pêche au lac Winnipeg.

[10:50-11:20]

Théo Michelot (Dalhousie University)

Multiscale Models of Animal Movement With Irreversible Dynamics

Modèles de déplacements d'animaux avec dynamiques irréversibles

Ecologists seek to understand how animal distributions arise from their small-scale movements. Data are collected by attaching tags to animals to record their locations (e.g., GPS collars), and statistical approaches

Les écologues étudient comment les distributions d'animaux émergent de leurs déplacements à petite échelle. Des données sont recueillies avec des balises attachées aux animaux pour enregistrer leur positions (e.g., colliers GPS), et les approches statistiques

fall into two categories: small-scale models of animals' movement decisions, and large-scale models of spatial distributions. The movement decisions give rise, in the long term, to large-scale distributions, but this mechanism is usually ignored, and it has been difficult to describe distributions as emerging properties of habitat-driven movements. I will show that irreversible processes with explicit stationary distributions are promising multiscale models for the movement of animals. This approach can be used to understand how animals' distributions are affected by the environment (e.g., habitat fragmentation), and makes it possible to combine data sources that have previously been analysed separately.

sont de deux types : modèles à petite échelle pour les décisions des animaux, ou modèles à grande échelle pour leur répartition dans l'espace. Les décisions de déplacements donnent lieu, sur le long terme, aux distributions à large échelle, mais ce mécanisme est souvent ignoré, et il est difficile de modéliser ce lien. Je présenterai une classe de processus irréversibles avec distribution stationnaire explicite, comme modèles de déplacements multi-échelles. Cette approche peut être utilisée pour comprendre comment les distributions d'animaux sont affectées par l'environnement (e.g., fragmentation d'habitat), et permet de combiner plusieurs types de données qui doivent habituellement être analysées séparément.

[11:20-11:50]

Alysha Cooper (University of Guelph) **Ayesha Ali** (University of Guelph) **Zeny Feng** (University of Guelph)

Sparse Regression Modeling for Compositional Data: Regularized Dirichlet-Multinomial Regression via Dominating Hyperplane Regularization

Modélisation de la régression parcimonieuse pour les données compositionnelles : régression multinomiale de Dirichlet régularisée par régularisation de l'hyperplan principal

Compositional data, prevalent in many ecological and biological applications—such as ecology, bioinformatics, and microbiology—pose unique challenges for analysis. The Dirichlet-multinomial (DM) regression framework is instrumental in addressing these challenges due to its ability to accommodate multinomial overdispersion. While the MM-algorithm is effective in fitting non-penalized DM regression models for compositional data, it is problematic for fitting doubly sparse DM models due to the non-separability of regression coefficients in the penalty function. Here we introduce dominating hyperplane regularization as a novel framework for efficiently fitting sparse regression models for complex grouped data via the MM algorithm. We demonstrate the performance of our method through simulations and its application to benthic compositions from the Alberta Oil Sands region.

Les données de composition, courantes dans de nombreuses applications écologiques et biologiques, telles que l'écologie, la bioinformatique et la microbiologie, posent des défis uniques en matière d'analyse. Le cadre de régression multinomiale de Dirichlet permet de les relever grâce à sa capacité à prendre en compte la surdispersion multinomiale. Bien que l'algorithme MM soit efficace pour ajuster les modèles de régression multinomiale de Dirichlet non pénalisés pour les données de composition, il pose problème quand il faut ajuster des modèles de régression multinomiale doublement clairsemés en raison de la non-séparabilité des coefficients de régression dans la fonction de pénalisation. Nous présentons la régularisation par l'hyperplan principal comme un nouveau cadre permettant d'ajuster efficacement des modèles de régression peu denses pour des données groupées complexes avec l'algorithme MM. Nous démontrons l'efficacité de notre méthode à l'aide de simulations et de son application aux compositions benthiques de la région des sables bitumineux de l'Alberta.

Chair/Président: Owen G Ward

Organizer/Responsable: Owen G Ward

Room/Salle: ED 2018B

Date: Wednesday June 5 / mercredi 5 juin

Time/Heure: 10:20-11:50

Abstract/Résumé

[10:20-10:50]

Neil A. Spencer (University of Connecticut)

Robust Bayesian Model Selection for Network Data

Sélection d'un modèle bayésien robuste pour les données de réseaux

The asymptotic behavior of Bayesian model selection is well-understood in the context of independent and identically distributed data. As the number of observations increases, the posterior concentrates on whichever candidate model is closest to the truth, with closeness determined by the Kullback-Leibler divergence. In this work, we extend these results to node exchangeable network data. Our contributions include: (1) establishing criteria for posterior concentration, (2) determining an appropriate notion of "closeness" to the truth, (3) identifying scenarios where Bayesian model selection is unstable, and (4) proposing an extension of BayesBag to network data to address this instability.

Le comportement asymptotique de la sélection d'un modèle bayésien est bien compris dans le contexte de données distribuées de façon identique ou indépendante. Au fur et à mesure que les observations augmentent, la distribution a posteriori se concentre à trouver quel modèle se rapproche le plus du réel, dont les critères de rapprochement sont déterminés par la divergence Kullback-Leibler. Dans le cadre de ce travail, nous élargissons l'étendue de ces résultats à des données de réseau à liens échangeables. Notre contribution comprend les points suivants : 1. L'établissement de critères pour la concentration a posteriori. 2. La détermination d'une notion adéquate de ≪ rapprochement ≫ au réel. 3. L'identification de scénarios dans lesquels la sélection d'un modèle bayésien est instable. 4. La proposition d'une extension de BayesBag pour les données de réseaux afin de résoudre cette instabilité.

[10:50-11:20]

Jie Jian (University of Waterloo)

Restricted Tweedie Stochastic Block Models

Modèles de blocs stochastiques Tweedie restreints

The stochastic block model (SBM) is a widely used framework for community detection in networks, where the network structure is typically represented by an adjacency matrix. However, conventional SBMs are not directly applicable to an adjacency matrix that consists of non-negative zero-inflated continuous edge weights. To model the network where edge weights represent financial values between agents, we propose an innovative SBM based on a restricted Tweedie distribution. Additionally, we incorporate edge information and account for its dynamic effect on edge weights. Notably, we show that given a sufficiently large number of

Le modèle à blocs stochastiques (SBM) est un cadre largement utilisé pour la détection de communauté dans les réseaux, lorsque la structure du réseau est généralement représentée par une matrice de contiguïté. Cependant, les SBM conventionnels ne sont pas directement applicables à une matrice de contiguïté composée de poids de bord continus non négatifs à surreprésentation de zéros. Pour modéliser le réseau commercial international dans lequel les poids de bord représentent les valeurs commerciales entre les pays, nous proposons un SBM innovant basé sur une distribution Tweedie restreinte. De plus, nous y incorporons des informations nodales et prenons en compte son effet dynamique sur les poids de bord. Nous montrons notamment que, étant donné un

nodes, estimating this covariate effect becomes independent of community labels of each node when computing the maximum likelihood estimator of parameters in our model. This result enables the development of an efficient two-step algorithm that separates the estimation of covariate effects from other parameters.

nombre suffisamment grand de nœuds, l'estimation de cet effet de covariable devient indépendante des étiquettes de communauté de chaque nœud lors du calcul de l'estimateur du maximum de vraisemblance des paramètres dans notre modèle. Ce résultat permet le développement d'un algorithme efficace en deux étapes qui sépare l'estimation des effets covariables des autres paramètres.

[11:20-11:50]

Peter W. MacDonald (McGill University) **Eric Kolaczyk** (McGill University)

Summaries of Markov Models for Evolving Networks - Statistical Properties

Résumés de modèles de Markov pour réseaux évolutifs – propriétés statistiques

In this work, we consider a class of continuous-time Markov chain models for binary network data which evolves over time on an aligned set of nodes. We investigate both continuous and discrete-time observation schemes, where under discrete observation, inference is based on partial observation of the underlying continuous-time process. We study the statistical properties of some commonly used dynamic network summaries (edge density, number of grown or dissolved edges) towards estimation and inference, as well as assessment of stationarity of the observed sequence.

Dans ce travail, nous étudions une classe de modèles de chaînes de Markov en temps continu pour des données de réseaux binaires qui évoluent dans le temps en fonction d'un ensemble de sommets alignés. Nous analysons les schémas d'observation en temps continu et en temps discret. Dans le cas de l'observation en temps discret, l'inférence est basée sur l'observation partielle du processus en temps continu sous-jacent. Nous examinons les propriétés statistiques de certains résumés de réseaux dynamiques souvent utilisés (densité des arêtes, nombre d'arêtes ajoutées ou enlevées) pour effectuer l'estimation et l'inférence, ainsi qu'une évaluation de la stationnarité de la séquence observée.

New Advancements in Formal Privacy Methods and Synthetic Data Generation
Nouvelles avancées en méthodes formelles de protection de la vie privée et génération de données synthétiques

Chair/Président: Bei Jiang

Organizer/Responsable: Bei Jiang, Éric Gagnon

Room/Salle: A 1046

Date: Wednesday June 5 / mercredi 5 juin

Time/Heure: 10:20-11:50

Abstract/Résumé

[10:20-10:50]

Weijie Su (University of Pennsylvania)

Enhancing Privacy Guarantees for the Census Data via Gaussian Differential Privacy

Amélioration des garanties de confidentialité pour les données de recensement grâce à la confidentialité différentielle gaussienne

The U.S. Census Bureau employs differential privacy mechanisms to safeguard the confidentiality of the 2020 Census data. However, the current methodology for privacy accounting, which utilizes zero Concentrated Differential Privacy (zCDP), is not optimal. This approach tends to introduce excessive noise for a specified privacy budget, leading to increased bias and compromised data accuracy. Although f-differential privacy (f-DP) offers a potential pathway for a more efficient privacy budget allocation, its integration into privacy accounting remains an unresolved challenge, as acknowledged by the Bureau. This talk presents an f-DP-based privacy accounting framework that refines the Bureau's existing RDP methodology. Our approach demonstrates the ability to achieve a reduced privacy budget for a comparable level of noise and, conversely, requires the addition of less noise to meet a predetermined privacy budget. Furthermore, our methodology results in noisy measurement files with enhanced accuracy within the same privacy constraints.

Le Bureau du recensement des États-Unis utilise des mécanismes de confidentialité différentielle pour protéger la confidentialité des données du recensement de 2020. Cependant, la méthodologie actuelle de comptabilisation de la confidentialité, qui utilise la confidentialité différentielle zéro concentrée (zCDP), n'est pas optimale. Cette approche tend à introduire un bruit excessif pour un budget de confidentialité donné, ce qui entraîne une augmentation des biais et compromet la précision des données. Bien que la confidentialité différentielle f (f-DP) permette une allocation plus efficace du budget de confidentialité, son intégration dans la comptabilité de confidentialité reste un défi non résolu, comme l'a reconnu le Bureau. Cet exposé présente un cadre de comptabilisation de la protection de la vie privée basé sur la f-DP qui affine la méthodologie RDP existante du Bureau. Notre approche montre qu'il est possible de réduire le budget de protection de la vie privée pour un niveau de bruit comparable et, inversement, qu'il est possible de respecter un budget de protection de la vie privée prédéterminé avec un ajout de bruit moindre. En outre, notre méthodologie permet d'obtenir des fichiers de résultats perturbés avec une précision accrue tout en respectant les mêmes contraintes de confidentialité.

[10:50-11:20]

Jingchen (Monika) Hu (Vassar College) **Terrance Savitsky** (U.S. Bureau of Labor Statistics) **Matthew Williams** (RTI International)

Mechanisms for Global Differential Privacy under Bayesian Data Synthesis

Mécanismes de confidentialité différentielle globale dans le cadre d'une synthèse bayésienne des données

We review, propose, and compare several Bayesian data synthesizers with different differential privacy guaran-

Nous examinons, proposons et comparons plusieurs synthétiseurs de données bayésiens avec différentes garanties différentielles

tees that can be used by data stewards for microdata dissemination with privacy protection. The pseudo posterior mechanism achieves an asymptotic differential privacy guarantee and a variant of it can provide faster convergence. The newly proposed censoring mechanism embedded in the pseudo posterior mechanism censors the pseudo likelihood of every record within [exp(-epsilon/2), exp(epsilon/2)], which provides a stronger, non-asymptotic differential privacy guarantee. Through a series of simulation studies with bounded, univariate data and an application to sample of the Survey of Doctoral Recipients where a beta regression synthesizer is utilized, we demonstrate that the pseudo posterior mechanism creates synthetic data with the highest utility at the price of a weaker, asymptotic privacy guarantee, while the censoring mechanism embedded in the pseudo posterior mechanism produces synthetic data with a stronger, non-asymptotic privacy guarantee at the cost of slightly reduced utility. The perturbed histogram is included for comparison.

de confidentialité qui peuvent être utilisées par les gestionnaires de données pour la diffusion de microdonnées avec protection de la confidentialité. Le mécanisme pseudo-postérieur offre une garantie différentielle asymptotique de protection de la confidentialité et une variante de ce mécanisme permet une convergence plus rapide. Le mécanisme de censure nouvellement proposé, intégré au mécanisme pseudo-postérieur, censure la pseudo-vraisemblance de chaque enregistrement dans la plage [exp(-epsilon/2), exp(epsilon/2)], ce qui permet d'obtenir une garantie différentielle de confidentialité non asymptotique plus forte. Grâce à une série d'études de simulation avec des données univariées bornées et à une application à un échantillon de l'enquête sur les titulaires de doctorat dans laquelle un synthétiseur de régression bêta est utilisé, nous démontrons que le mécanisme pseudo postérieur crée des données synthétiques de la plus grande utilité au prix d'une garantie de confidentialité asymptotique plus faible, tandis que le mécanisme de censure intégré dans le mécanisme pseudo postérieur produit des données synthétiques avec une garantie de confidentialité non asymptotique plus forte au prix d'une légère réduction de l'utilité. Nous incluons l'histogramme perturbé à des fins de comparaison.

[11:20-11:50]

Héloïse Gauvin (Statistique Canada)

Creating a Synthetic version of a Longitudinal and Structured file: challenges and lessons learned

Créer une version synthétique d'un fichier longitudinal et structuré : les défis et les leçons apprises

In recent years, generating synthetic data has increasingly been viewed by statistical agencies as a means of disseminating useful statistical information while fulfilling their obligations to protect the personal data they have been entrusted with. In 2018, Statistics Canada was one of the first national statistical agencies to release smart synthetic data files, where there is a goal of maintaining the analytical value of the original data. Since then, through various initiatives, its expertise has kept growing, establishing the Agency as a leader in the field. This presentation will focus on the production of more elaborate smart synthetic dataset than those previously produced. Intended to safely run open-source micro-simulations for the new Canadian retirement income model (PASSAGES), a massive smart synthetic file including both longitudinal and family components was created. We will show how the synthesis process had to be enhanced to handle the challenges of capturing temporal correlation and life events.

Ces dernières années, la production de données synthétiques est de plus en plus considérée par les organismes statistiques comme un moyen de diffuser des informations statistiques utiles tout en remplissant leurs obligations de protection des données personnelles qui leur sont confiées. En 2018, Statistique Canada a été l'un des premiers organismes statistiques nationaux à publier des fichiers de données synthétiques « smart », dans le but de maintenir la valeur analytique des données originales. Depuis, grâce à diverses initiatives, son expertise n'a cessé de croître, faisant de l'Agence un leader dans le domaine. Cette présentation se concentrera sur la création d'un jeu de données synthétiques « smart » plus élaboré que ceux produits précédemment. Destiné à exécuter en toute sécurité des microsimulations à code source ouvert pour le nouveau modèle canadien des revenus de retraite (PASSAGES), un énorme fichier synthétique « smart » comprenant à la fois des composantes longitudinales et familiales a été créé. Nous montrerons comment le processus de synthèse a dû être amélioré pour relever les défis liés à la capture des corrélations temporelles et des événements de la vie.

Spotlight on CANSSI postdocs
Vitrine des étudiants postdoctoraux de l'INCASS

Chair/Président: Andrea Benedetti

Organizer/Responsable: Andrea Benedetti

Room/Salle: A 1043

Date: Wednesday June 5 / mercredi 5 juin

Time/Heure: 10:20-11:50

Abstract/Résumé

[10:20-10:50]

William Ruth (University of Montreal) **Rado Malalatiana Ramasy** (University of Montreal) **Rowin Alfaro** (University of Montreal) **Ariel Mundo** (University of Montreal) **Bouchra Nasri** (University of Montreal)

Statistical Considerations in Causal Mediation Analysis

Considérations statistiques dans l'analyse de la médiation causale

Causal mediation analysis is a popular tool for studying complicated causal dependence between multiple variables. We investigate the extent to which the effect of an exposure, X, on a response, Y, is mediated by a third variable, M. One common approach involves fitting some regression models and computing straightforward functions of the estimated coefficients. Unfortunately, uncertainty quantification for these "mediation effects" is non-trivial in even the simplest settings. A popular alternative to these analytical standard error formulas is to use the bootstrap, although selecting a single version is challenging in practice. We present a range of implementation for the bootstrap on a complicated model involving non-linearity and mixed-effects, and illustrate our analysis on a dataset investigating the relationship between trustworthiness of peoples' preferred news source and willingness to adhere to pandemic lockdown mandates.

L'analyse de médiation causale constitue une méthode en plein essor pour explorer les relations causales complexes entre plusieurs variables. Cette approche permet d'évaluer dans quelle mesure l'effet d'une exposition, X, sur une issue, Y, est médié par une variable intermédiaire, M. Habituellement, cette analyse s'appuie sur l'ajustement de modèles de régression et le calculer directement les fonctions des coefficients estimés. Toutefois, la quantification de l'incertitude associée à ces « effets de médiation » s'avère compliquée, même dans les scénarios les plus élémentaires. Face à la difficulté d'utiliser les formules analytiques pour estimer l'erreur standard, le recours au bootstrap représente une alternative courante, bien que la sélection d'une méthode spécifique constitue un défi en pratique. Nous proposons ici plusieurs variétés de bootstrap adaptées à un modèle complexe intégrant de la non-linéarité et des effets mixtes. Nous illustrons nos analyses à travers un jeu de données dans le cadre d'une étude sur la relation entre la crédibilité de la source d'information préférée des individus et leur propension à respecter les directives de confinement pendant une pandémie.

[10:50-11:20]

Chi-Kuang Yeh (University of Waterloo) **Peijun Sang** (University of Waterloo) **Qihuang Zhang** (McGill University) **Archer Yi Yang** (McGill University) **Celia Greenwood** (McGill University)

Multivariate Spatial Functional Principal Component Analysis

Analyse en composantes principales multivariée pour données fonctionnelles spatiales

Various forms of data have become available due to recent technological advancements, among which is functional data (FD). FD differs from scalar or vector-valued quantities as it entails observations in the form of functions or curves. This emerging area poses unique chal-

Diverses formes de données sont devenues disponibles grâce à des avancées technologiques récentes, dont les données fonctionnelles (FD). Celles-ci diffèrent des quantités scalaires ou à valeurs vectorielles, car les observations sont sous la forme de fonctions ou de courbes. Ce secteur émergent présente des problèmes uniques

lenges due to its intrinsically infinite-dimensional (IID) nature. In circumstances where multiple stochastic processes are recorded simultaneously for a selection of subjects, such as in neuroscience when objects are indexed by their unique spatial labels (SLs), it yields multivariate spatial FD. These SLs attached to this data play an important role in determining the accounted-for correlation. We propose a new approach to acquiring information from these IID multivariate objects with spatial information. Our method specifically allows multiple curves with possibly different resolutions and without requiring a shared grid. Simulation studies and two real-world examples demonstrate the effectiveness of our methodology.

en raison de sa nature intrinsèquement de dimension infinie (IID). Dans des cas où des processus stochastiques multiples sont enregistrés simultanément pour une sélection de sujets, comme en neuroscience lorsque des objets sont indexés par leur étiquette spatiale unique (SL), cela produit des données fonctionnelles spatiales multivariées. Les étiquettes spatiales (SLs) liées à ces données jouent un rôle important dans la détermination de la corrélation expliquée. Nous proposons une nouvelle approche pour extraire de l'information de ces objets multivariés IID avec information spatiale. Notre méthode permet explicitement les courbes multiples avec des résolutions possiblement différentes, sans nécessiter une grille partagée. Des études de simulation et deux exemples sur des données réelles illustrent l'efficacité de notre méthodologie.

Chair/Président: Andrei Volodin

Organizer/Responsable: Andrei Volodin

Room/Salle: IIC 2001

Date: Wednesday June 5 / mercredi 5 juin

Time/Heure: 10:20-11:50

Abstract/Résumé

[10:20-11:50]

Cong Jiang (University of Montreal) **Michael Wallace** (University of Waterloo) **Mary Thompson** (University of Waterloo)

Dynamic Treatment Regimes and Interference

Régimes de traitement dynamiques et interférences

In the realm of personalized medicine, where treatments are tailored to individual patient characteristics, dynamic treatment regimes (DTRs) have emerged as a crucial framework for making optimal sequential treatment decisions. DTRs are sequences of decision rules leveraging patient-specific data to generate treatment recommendations. Traditionally, DTRs have been formulated under the assumption of no interference, meaning that the treatment of one person does not influence another's outcome. However, this assumption is often questionable in practical scenarios, such as in infectious diseases or social network behaviors. In this presentation, we will discuss recent advancements in DTR estimation with interference. We will present improvements to the established dynamic weighted ordinary least squares (dWOLS) method adapted for dyadic networks, where interference arises between pairs of individuals. Central to our approach are the core components of our proposed interference-aware dWOLS: network propensity functions and network interference balancing weights. These elements are instrumental in addressing interference within networks, thereby ensuring the double robustness property of our estimation method. We will showcase the application of household smoking cessation to elucidate the proposed methods, utilizing the Population Assessment of Tobacco and Health (PATH) data.

Dans le domaine de la médecine personnalisée, où les traitements sont adaptés aux caractéristiques individuelles des patients, les régimes de traitement dynamiques (RTD) sont apparus comme un cadre essentiel pour prendre des décisions séquentielles optimales en matière de traitement. Les RTD sont des séquences de règles de décision qui s'appuient sur des données spécifiques au patient pour générer des recommandations de traitement. Traditionnellement, les RTD ont été formulés en partant de l'hypothèse de l'absence d'interférence, ce qui signifie que le traitement d'une personne n'influe pas sur le résultat d'une autre. Cependant, cette hypothèse est souvent discutable dans des scénarios pratiques, tels que les maladies infectieuses ou les comportements des réseaux sociaux. Dans cette présentation, nous discuterons des avancées récentes dans l'estimation RTD avec interférence. Nous présenterons des améliorations apportées à la méthode des moindres carrés ordinaires pondérés dynamiques (dWOLS) adaptée aux réseaux dyadiques, où une interférence survient entre paires d'individus. Les éléments centraux de notre proposition de dWOLS tenant compte des interférences sont au cœur de notre approche : les fonctions de propension du réseau et les poids d'équilibrage des interférences du réseau. Ces éléments permettent de traiter les interférences au sein des réseaux, garantissant ainsi la double robustesse de notre méthode d'estimation. Nous présenterons l'application de l'arrêt du tabac dans les ménages afin d'élucider les méthodes proposées, en utilisant les données de l'étude Population Assessment of Tobacco and Health (PATH).

Advances in Statistical Models for Single Cell RNA-seq Data
Progrès en modèles statistiques pour les données d'ARN-seq de cellules uniques

Chair/Président: Qihuang Zhang

Organizer/Responsable: Qihuang Zhang

Room/Salle: ED 2018A

Date: Wednesday June 5 / mercredi 5 juin

Time/Heure: 10:20-11:50

Abstract/Résumé

[10:20-10:50]

Qingrun Zhang (University of Calgary) **Sandesh Acharya** (University of Calgary) **Jiami Guo** (University of Calgary)

Stabilized Marker Gene and Pathway Identification in Single-Cell RNA-seq Data

Gène marqueur stabilisé et identification de voie dans des données de séquençage de l'ARN à cellule unique

It is challenging to select informative marker genes that can robustly identify and discriminate cell types while provide meaningful cellular biological functions. Based upon our previous success in developing stability tool for bulk RNA-seq (Stabilized COre Gene identification and Pathway Election, or SCOPE), we developed single-cell version, scSCOPE, to allow robust identification of marker genes as well as meaning functional annotation for cell clusters. scSCOPE employs a dual-tier stabilization strategy: Iterative LASSO Regression, which identifies genes stable across multiple LASSO rounds, and co-expression network analysis, capturing gene-gene interactions. We tested scSCOPE using scRNA-seq data from immune system, identifying stable and functional relevant outcomes. Confrontational to the current trend of massive heterogeneity in scRNA-seq data, scSCOPE will discover more homogeneous results by strengthening the stability of statistical models

Il est complexe de sélectionner des gènes marqueurs informatifs pouvant rigoureusement identifier et distinguer les types de cellules tout en procurant des fonctions biologiques cellulaires pertinentes. En nous basant sur nos réussites antérieures dans le développement d'un outil de stabilité pour un grand nombre de séquençages de l'ARN (identification de gène stabilisée et élection de voie, ou SCOPE), nous avons élaboré une version à cellule unique (scSCOPE) pour rendre possible l'identification rigoureuse des gènes marqueurs ainsi que l'annotation fonctionnelle des regroupements cellulaires. L'outil scSCOPE adopte une stratégie de stabilisation en deux volets : une régression LASSO itérative, qui identifie les gènes stables sur plusieurs rondes LASSO, et l'analyse de réseau de co-expression, servant à déceler les interactions entre gènes. Nous avons testé scSCOPE à l'aide de données de séquençage de l'ARN à cellule unique (scRNA-seq) provenant du système immunitaire, afin d'identifier les résultats fonctionnels pertinents et constants. En s'opposant à la tendance actuelle d'hétérogénéité massive dans les données scRNA-req, scSCOPE découvrira davantage de résultats homogènes en renforçant la stabilité des modèles statistiques.

[10:50-11:20]

Pingzhao Hu (Western University)

Spatial Transcriptomic Profile Prediction from Histology Images using Novel Contrastive Learning

Prédiction du profil transcriptomique spatial à partir d'images histologiques à l'aide d'un nouvel apprentissage contrastif

Progress in spatial transcriptomics allows precise RNA measurement. Predicting gene expression from histology image is vital, but current methods lack capturing 2D visual features and spatial dependencies. We present a contrastive learning model predicting RNA-seq expression from histology images and imputing

Les progrès de la transcriptomique spatiale permettent de mesurer l'ARN avec précision. La prédiction de l'expression génique à partir d'images histologiques est essentielle, mais les méthodes actuelles ne permettent pas de capturer les caractéristiques visuelles 2D ni les dépendances spatiales. Nous présentons un modèle d'apprentissage contrastif prédisant l'expression du séquençage de

gene expression. Sequenced spots' histology images are cropped, processed through convolutional and graph convolutional modules, aligning image and gene expression features into a 256-dimensional space. Projected gene expression undergoes imputation via a Transformer-based autoencoder with RMSE and zero-inflated negative binomial distribution reconstruction loss. The model excels in gene expression prediction and spatial identification than other baseline models. Overall, we provide a robust solution for spatial transcriptomics from histology images, uncovering tissue molecular signatures.

l'ARN à partir d'images histologiques et imputant l'expression génique. Les images histologiques des taches séquencées sont recadrées, traitées par des modules convolutifs et graphiques convolutifs, ce qui permet d'aligner les caractéristiques de l'image et de l'expression génique dans un espace à 256 dimensions. L'expression génique projetée subit une imputation au moyen d'un auto-encodeur basé sur un transformateur avec une perte de racine d'erreur quadratique moyenne et reconstruction de la distribution binomiale négative avec excès de zéros. Le modèle permet une prédiction de l'expression génétique et une identification spatiale supérieures à celles d'autres modèles de base. Nous proposons, à partir d'images histologiques, une solution robuste pour la transcriptomique spatiale, qui permet d'identifier les signatures moléculaires des tissus.

[11:20-11:50]

Xuekui Zhang (University of Victoria) **Li Xing** (University of Saskatchewan)

Does increasing sample size inflate false positive rates?

L'augmentation de la taille de l'échantillon gonfle-t-elle les taux de faux positifs ?

In statistical theory, the notion that enlarging the sample size affects Type I error rates is unfounded. Yet, this concept frequently raises concerns outside the realm of professional statisticians. It is essential to recognize that the gap between theoretical statistics and its application in the real world often stems from the imperfect adherence to model assumptions, adding layers of complexity to the issue. Our stance is that an increased sample size does not inherently introduce problems. However, it can potentially amplify small issues to critical levels. In this talk, we will illustrate this point through simulation examples, reinforcing our viewpoint. Furthermore, we will introduce two solutions to mitigate the impact of increasing sample sizes. The first is a broad-based solution applicable to a range of scenarios, while the second specifically addresses the challenges of differential expression analysis of single-cell genomic data.

En théorie statistique, l'idée que l'augmentation de la taille de l'échantillon affecte les taux d'erreur de Type I est infondée. Pourtant, ce concept soulève fréquemment des préoccupations en dehors du domaine des statisticiens professionnels. Il est essentiel de reconnaître que l'écart entre la statistique théorique et son application dans le monde réel provient souvent du respect imparfait des hypothèses du modèle, ce qui ajoute des couches de complexité à la question. Notre position est qu'une augmentation de la taille de l'échantillon n'introduit pas intrinsèquement de problèmes. Cependant, elle peut potentiellement amplifier des petits problèmes à des niveaux critiques. Dans cette présentation, nous illustrerons ce point à travers des exemples de simulations, renforçant notre point de vue. De plus, nous présenterons deux solutions pour atténuer l'impact de l'augmentation des tailles d'échantillons. La première est une solution générale applicable à une gamme de scénarios, tandis que la seconde aborde spécifiquement les défis de l'analyse d'expression différentielle des données génomiques de cellules uniques.

Chair/Président: Hina Shaheen

Room/Salle: C 4036

Date: Wednesday June 5 / mercredi 5 juin

Time/Heure: 10:20-11:50

Abstract/Résumé

[10:20-10:35]

Jacqueline A. May (University of Waterloo) **Zeny Feng** (University of Guelph) **Sarah J. Adamowicz** (University of Guelph)

Approaches for Handling Missing Values and Their Impacts on Statistical Inferences: A Molecular Rate Case Study

Approches de traitement des valeurs manquantes et de leur impact sur les inférences statistiques : étude de cas sur les taux moléculaires

We investigated the impact of different approaches for handling missing data on inferences using a molecular evolution case study. Using a fish trait dataset, we aimed to identify traits that were significantly associated with molecular rates. Multivariable regressions were performed using: 1) a complete-case dataset, and datasets imputed using 2) trait-only (non-phylogenetic) imputation and 3) phylogenetic imputation methods. Results were compared to assess the impact of the missing data handling approach on the significance level of the association for each trait. Results indicated that the model fitted to the phylogenetic imputed data aligned with complete case, while also revealing new insights about molecular rate correlates in fishes. These results are supported by those of previous studies and suggest that the missing data handling approach can have a considerable impact on biological conclusions.

Nous avons étudié l'impact de différentes approches de traitement des données manquantes sur les inférences via une étude de cas sur l'évolution moléculaire. Sur la base d'un ensemble de données sur les caractères des poissons, nous avons cherché à identifier les caractères qui étaient significativement associés aux taux moléculaires. Nous avons réalisé des régressions multivariables en utilisant : 1) un ensemble de données complet et imputé des ensembles de données à l'aide de 2) méthodes d'imputation de traits uniquement (non phylogénétiques) et de 3) méthodes d'imputation phylogénétiques. Nous avons ensuite comparé les résultats pour évaluer l'impact de la méthode de traitement des données manquantes sur le niveau de signification de l'association pour chaque caractère. Les résultats indiquent que le modèle ajusté aux données imputées phylogénétiques s'aligne sur le cas complet, tout en révélant de nouvelles informations sur les corrélats du taux moléculaire chez les poissons. Nous avons corroboré ces résultats par ceux d'études antérieures qui suggèrent que cette approche du traitement des données manquantes peut avoir un impact considérable sur les conclusions biologiques.

[10:35-10:50]

Yuan Bian (University of Western Ontario) **Grace Y. Yi** (University of Western Ontario) **Wenqing He** (University of Western Ontario)

Boosting Learning in the Presence of Incomplete Data

Apprentissage par boosting en présence de données incomplètes

Boosting techniques have attracted increasing attention in both machine learning and statistical research. While various methods have been developed for different settings, most are designed primarily for complete datasets, limiting their applicability to handle incomplete data such as missing observations and censored data. To address the challenges posed by incomplete data, we in-

Les techniques de boosting ont attiré une attention croissante tant dans l'apprentissage automatique que dans la recherche statistique. Bien que diverses méthodes aient été développées pour différents contextes, la plupart sont principalement conçues pour des ensembles de données complets, limitant ainsi leur applicabilité pour traiter des données incomplètes telles que des observations manquantes et des données censurées. Pour relever les

troduce boosting estimation techniques tailored specifically for such scenarios. By accounting for the missing data effects, we develop an implementation algorithm using a functional gradient descent and evaluate its performance through numerical studies in finite sample settings.

défis posés par les données incomplètes, nous introduisons des techniques d'estimation de boosting spécialement conçues pour de tels scénarios. En tenant compte des effets des données manquantes, nous développons un algorithme d'implémentation utilisant une descente de gradient fonctionnelle et évaluons ses performances à travers des études numériques dans des contextes d'échantillonnage fini.

[10:50-11:05]

Gracia Y. Dong (University of Toronto/University of Victoria) **Jennifer McNichol** (Simon Fraser University) **Laura L.E. Cowen** (University of Victoria)

Population Size Estimation in a Two-Sample Study using Capture-Recapture Techniques

Estimation de la taille de la population dans une étude à deux échantillons à l'aide de techniques de capture-recapture

Two-sample capture-recapture studies, commonly used in epidemiological and wildlife ecology literature, have predominantly relied on the Lincoln-Petersen estimator for analysis. We outline the use of the Lincoln-Petersen estimator and two alternative closed-population methods to analyze data from a two-sample capture-recapture study: Huggins' conditional likelihood method and Pledger's likelihood method with mixtures. These methods offer opportunities to model capture probabilities dependent on time, behaviour, individual heterogeneity, or incorporating individual covariates. An extensive simulation study is performed to quantify the effects of model misspecification on population size estimates, as well as the ability of AIC to perform model selection when capture probabilities vary by time, behavioural response, or hidden group membership. We found that the models are not robust to mispecification, and AIC was not capable of selecting the correct model with two capture occasions.

Les études de capture-recapture à deux échantillons, souvent utilisées dans la littérature sur l'épidémiologie et l'écologie de la faune, reposent principalement sur l'estimateur de Lincoln-Petersen en ce qui concerne l'analyse. Nous décrivons l'utilisation de l'estimateur de Lincoln-Petersen et de deux autres méthodes de population fermée pour analyser les données d'une étude à deux échantillons à l'aide de techniques de capture-recapture : la méthode de vraisemblance conditionnelle de Huggins et la méthode de vraisemblance de Pledger avec des mélanges. Ces méthodes permettent de modéliser les probabilités de capture en fonction du temps, du comportement, de l'hétérogénéité individuelle ou de l'intégration de covariables individuelles. Une étude de simulation approfondie est réalisée pour quantifier les effets d'une erreur de spécification du modèle d'estimation de la taille de la population, ainsi que la capacité du critère d'information d'Akaike (AIC) à effectuer une sélection de modèles lorsque les probabilités de capture varient en fonction du temps, de la réponse comportementale ou de l'appartenance à un groupe caché. Nous avons constaté que les modèles ne sont pas robustes aux erreurs de spécification et que l'AIC ne permet pas de sélectionner le bon modèle avec deux occasions de capture.

[11:05-11:20]

Jonathan Babyn (Dalhousie University)

Evaluating the Feasibility of Juvenile Only CKMR on Grey Seals

Évaluation de la faisabilité de l'application de la méthode de capture-marquage-recapture génétique sur des juvéniles seulement pour les phoques gris

Close-kin mark recapture (CKMR) is a method to estimate the abundance of wildlife populations. It swaps the probability of finding a physically marked individual (from an earlier sample) as used in traditional mark recapture for the probability of finding a close-kin pair (e.g., parent of offspring, half sibling, etc.) in the population. Typically, this involves samples of both adults and juveniles from the population of interest. Here we

La méthode de capture-marquage-recapture génétique permet d'estimer la densité des populations de la faune et de la flore sauvages. Elle remplace la méthode traditionnelle de probabilité de trouver un individu physiquement marqué (à partir d'un échantillon antérieur) par la probabilité de trouver un couple de proches parents (p. ex., le parent d'une progéniture, un demi-frère ou une demi-sœur) dans la population. En général, cette méthode nécessite des échantillons d'adultes et de juvéniles de la popula-

examine the feasibility of applying CKMR with a juvenile only based sampling scheme using an individual based simulation modelled after the Sable Island grey seal colony. We show that it is possible to get reasonably accurate estimates of population abundance despite challenges with estimating average fecundity and utilizing potential Grandparent-Grandchild kin pairs.

tion concernée. Nous examinons la faisabilité de l'application de la méthode de capture-marquage-recapture génétique à un schéma d'échantillonnage reposant uniquement sur des juvéniles, à l'aide d'une simulation individuelle modélisée à partir de la colonie de phoques gris de l'île de Sable. Nous démontrons qu'il est possible d'obtenir des estimations suffisamment précises de l'abondance de la population malgré les difficultés liées à l'estimation de la fécondité moyenne et à l'utilisation des paires potentielles de parents grands-parents-petits-enfants.

[11:20-11:35]

Hoang Nguyen (Fisheries and Marine Institute of Memorial University of Newfoundland)

Accounting for Movement in Spatial Surplus Production Models: A Case Study on 3LN Redfish

Prise en compte du déplacement dans les modèles spatiaux de production excédentaire : étude de cas du sébaste 3LN

Fish movements present computational challenges for spatial surplus production models (SSPMs) and can be confounded with process errors, hindering the identification of movement and SSPM parameters. We propose leveraging a Gaussian random field (GRF) with a Matérn covariance structure to account for process errors and complex movements, thereby circumventing computational and confounding issues. Our approach uses advanced techniques such as converting GRF into Gaussian Markov random field, Laplace approximation for high-dimensional integrations, and automatic differentiation for efficient implementation. Through simulation studies, wherein data is generated explicitly considering movements, our novel random field approach demonstrates superior performance in estimating fish spatial abundance and SSPM parameters compared to existing methods. Our method reveals the spatial distribution of annual stock densities for redfish in NAFO 3LN divisions based on survey and commercial catch data.

Dans les modèles spatiaux de production excédentaire (SSPM), la prise en compte du déplacement des poissons présente des problèmes computationnels qui peuvent être confondus avec des erreurs de traitement, rendant difficile l'identification du déplacement et des paramètres des SSPM. Nous proposons de mettre à profit un champ aléatoire gaussien (GRF) avec une structure de la covariance Matérn afin de prendre en compte les erreurs de traitement et les déplacements complexes, contournant ainsi les problèmes de computation et de confusion. Notre approche se fonde sur l'utilisation de techniques avancées, telles que la conversion du GRF en champ aléatoire de Markov gaussien, l'approximation de Laplace pour les intégrations de grande dimension et la différentiation automatique pour une implémentation efficace. À l'aide d'études de simulation, dans lesquelles les données sont générées en prenant explicitement en compte les déplacements, la performance de notre nouvelle approche de champ aléatoire se révèle supérieure pour l'estimation de l'abondance spatiale des poissons et des paramètres des SSPM, par comparaison à celle des méthodes existantes. À partir des données d'enquête et de la pêche commerciale, notre méthode montre la distribution spatiale des densités des stocks annuels de sébastes dans les divisions 3LN de l'Organisation des pêches de l'Atlantique Nord-Ouest (OPANO).

Chair/Président: Kevin Granville

Room/Salle: C 3053

Date: Wednesday June 5 / mercredi 5 juin

Time/Heure: 10:20-11:50

Abstract/Résumé

[10:20-10:35]

Orla A. Murphy (Dalhousie University) **Jonathan Jalbert** (Polytechnique Montréal)

Predicting Extreme Rainfall in Nova Scotia Using a Spatial Bayesian Hierarchical Model

Prévision des précipitations extrêmes en Nouvelle-Écosse à l'aide d'un modèle hiérarchique bayésien spatial

The peaks-over-threshold (POT) approach to defining extreme observations is intuitive and less wasteful than other methods such as the block maxima approach. However, POT requires the selection of an appropriate high threshold to define an extreme observation. In spatial modeling problems, POT requires threshold selection at each location in the study. Threshold selection is usually performed using visual tools and therefore can be subjective and tedious when modeling data from many locations. Extended generalized Pareto distributions (EGPDs), extensions of the standard POT model, are more robust to threshold selection and appear suitable for modeling extremes using low thresholds, thereby incorporating more data in the analysis and potentially eliminating careful threshold selection by location. This work will extend an EGPD spatially using a Bayesian hierarchical model to predict rainfall return levels across Nova Scotia.

La méthode des excès pour définir les observations extrêmes est intuitive et moins coûteuse que d'autres méthodes telles que la méthode des maxima en blocs. Cependant, la méthode des excès nécessite la sélection d'un seuil élevé approprié pour définir une observation extrême. Dans les problèmes de modélisation spatiale, la méthode des excès nécessite la sélection d'un seuil à chaque endroit de l'étude. La sélection des seuils est généralement effectuée à l'aide d'outils visuels et peut donc être subjective et fastidieuse lors de la modélisation de données provenant de nombreux endroits. Les distributions de Pareto généralisées étendues (EGPD), extensions du modèle des excès standard, sont plus robustes à la sélection des seuils et semblent convenir à la modélisation des extrêmes à l'aide de seuils bas, ce qui permet d'incorporer davantage de données dans l'analyse et d'éliminer potentiellement la sélection minutieuse des seuils à chaque endroit. Ce travail étendra un EGPD dans l'espace en utilisant un modèle hiérarchique bayésien pour prédire les niveaux de retour des précipitations en Nouvelle-Écosse.

[10:35-10:50]

Henrik Stryhn (University of Prince Edward Island)

Design and Analysis for Ranking of Machine-Rated Applications to a Professional Program

Conception et analyse pour le classement d'applications à un programme de formation professionnelle évaluées par ordinateur

Recent advances in artificial intelligence (AI) learning algorithms have made it possible to use machine (AI) rating to evaluate applications. We use data from two admissions cycles to a veterinary program, where 778 and 552 applicants were rated by both a panel of human raters and AI rating, to discuss new challenges for statistical design and analysis posed by AI rating. Statistical methodology relies on replication, in the context of rating the use of multiple raters, and methods exist

De récentes percées en algorithmes d'apprentissage dans le domaine de l'intelligence artificielle (IA) ont rendu possible l'utilisation de l'ordinateur pour l'évaluation des applications. À l'aide de données tirées de deux cycles d'admission à un programme en médecine vétérinaire dans lesquels 778 et 552 postulants ont été notés par des évaluateurs humains et par l'IA, nous abordons les nouveaux problèmes de conception et d'analyse statistiques que pose l'évaluation par intelligence artificielle. La méthodologie statistique se base sur la réplication, dans un contexte d'évaluation

to adjust for heterogeneity between raters. For AI rating, adjudicating applications together in sets is attractive to keep computing costs down. We explore how rating each application in multiple sets (of 5) may serve as replication and compare ways of constructing adjudication sets. Methods for analyzing the resulting data, with ratings on an integer 1-10 scale, are contrasted with those for human rating. Our study leads to recommendations for design and analysis of AI rating applications.

de l'utilisation d'évaluateurs multiples, et il existe des méthodes pour tenir compte de l'hétérogénéité entre les évaluateurs. Dans le cas de l'évaluation par IA, il est intéressant de juger les demandes ensemble, par séries, afin de réduire les coûts de calcul. Nous examinons comment l'évaluation de chaque application dans des ensembles multiples (de 5) peut servir comme réplication et comparons des moyens de construire des ensembles d'évaluation. Les méthodes d'analyse des données qui en résultent, avec des notes sur une échelle de nombres entiers de 1 à 10 sont comparées à celles de l'évaluation par des humains. Notre étude mène à des recommandations en matière de conception et d'analyse des applications d'évaluation par IA.

[10:50-11:05]

Syeda Fateha Akter (Memorial University of Newfoundland)

Parameter Estimation of Poisson Autoregressive Moving Average Model

Estimation de paramètres d'un modèle de moyennes mobiles autorégressif de Poisson

When equally spaced time series of counts is observed along with covariate information at each time point several authors have discussed the analysis of such data with Poisson Autoregressive or Integer valued Autoregressive (INAR) models. The basic properties and estimation of these INAR models are well documented in the literature. There are however some count time series data that may not be adequately model with INAR models. Consequently, we consider Poisson autoregressive moving average model of order (1,1) for count time series with covariate information. We derive the basic properties of the model and discuss the estimation of the model parameters. The performance of our proposed methods are examined through simulation studies. Keywords: Poisson autoregressive model, autoregressive moving average model, estimation.

Dans le cas de séries temporelles de dénombrement espacées de façon égale avec de l'information sur les covariables pour chaque point dans le temps, plusieurs auteurs ont abordé l'analyse avec des modèles autorégressifs à valeur entière (INAR) ou de Poisson. Les propriétés de base et l'estimation de ces modèles INAR sont bien documentées dans la littérature. Cependant, il existe des données de séries temporelles de dénombrement ne pouvant pas être modélisées correctement avec des modèles INAR. Par conséquent, nous examinons un modèle de moyennes mobiles autorégressif de Poisson d'ordre (1,1) pour les séries temporelles de dénombrement avec information sur les covariables. Nous dérivons les propriétés de base du modèle et discutons de l'estimation des paramètres du modèle. Nous évaluons la performance de la méthode proposée grâce à des études desimulation. Mots clés : Modèle autorégressif de Poisson, modèle de moyennes mobiles autorégressif, estimation.

[11:05-11:20]

Archer Gong Zhang (University of Toronto) **Nancy Reid** (University of Toronto) **Qiang Sun** (University of Toronto)

A Semiparametric Approach to Data-Integrated Causal Inference

Approche semi-paramétrique pour une inférence causale avec données intégrées

In causal inference, data may be collected from both experimental and observational studies. Experimental studies often suffer from the lack of external validity due to limitations inherent in the studies. Observational studies are usually broad enough to be representative of the target populations but often lack internal validity due to inevitable confounders. Recently, there has been a lot of discussions on integrating these data. In this talk, we introduce a semiparametric approach based on the density ratio model (DRM) to utilize the comple-

En inférence causale, les données peuvent être recueillies à partir d'études observationnelles et expérimentales. D'une part, les études expérimentales manquent souvent de validité externe à cause des limites inhérentes dans les études. Par contre, les études observationnelles sont généralement assez larges pour être représentatives de populations cibles, mais manquent souvent de validité interne en raison des inévitables facteurs de confusion. Un grand nombre de discussions a récemment porté sur l'intégration de ces données. Dans le cadre de cet exposé, nous présentons une approche semi-paramétrique basée sur le modèle de taux

mentary features of the two types of studies. DRM can efficiently address the latent structures among multiple interconnected populations. If the related studies share common measurements for the same causal effect, the collected datasets are naturally expected to be from connected populations, and therefore the DRM may work well. We study several estimators of the causal effect, considering not only the mean but also distributional perspectives.

de densité (DRM) dans le but de se servir des caractéristiques complémentaires des deux types d'étude. Le DRM réussit de façon efficace à gérer les structures latentes parmi les populations interconnectées. Si les études partagent des mesures communes pour le même effet causal, on s'attend à ce que les ensembles de données recueillies proviennent naturellement de populations connectées, et donc le DRM pourrait bien fonctionner. Nous étudions plusieurs estimateurs de l'effet causal, en tenant compte non seulement de la moyenne, mais aussi des perspectives distributionnelles.

[11:20-11:35]

Zixuan Yang (The University of Western Ontario) **Douglas Woolford** (Statistical & Actuarial Sciences, University of Western Ontario)

Predict the Wildfire Occurrence in Ontario, Canada: An Errors in Variable Modelling Approach

Approche de modélisation des erreurs dans les variables pour prévoir la fréquence des feux de forêt en Ontario, au Canada

Wildfire occurrence prediction (FOP) models are used as decision support tools by fire management agencies. However, most of the research in this area has focused on daily predictions based on currently observed fire-weather conditions. Less research has considered forecasting FOP beyond the current day. Here, we consider short-term forecasting of FOP one to four days into the future using weather forecast data from the Fort Frances and Dryden Fire Management Districts in the Province of Ontario, Canada. In this initial study, we show that a key predictor for human-caused fire occurrence is subject to forecasting errors, and ignoring these leads to biased or incorrect estimators. It can be shown that the measurement error distribution can be modeled by normal mixtures. Thus, we employ a modified simulation-extrapolation (SIMEX) algorithm that assumes the forecast error in the predictor to be a mixture of normal distributions. Since we are interested in predicting both the presence and the number of fires, logistic and count based models are all to be considered. Through simulation, we demonstrate that ignoring the measurement errors in a predictor result in biased forecasts, while the SIMEX estimators lead to improvements.

Les agences de gestion des incendies utilisent des modèles de prévision des feux de forêt comme outils d'aide à la décision. Cependant, la plupart des recherches dans ce domaine ont porté sur les prévisions quotidiennes en fonction des conditions météorologiques actuellement observées. Peu de recherches ont porté sur les prévisions de feux de forêt au-delà de la journée en cours. Nous examinons les prévisions à court terme (d'un à quatre jours) des feux de forêt à l'aide des données des prévisions météorologiques des zones de gestion des incendies de Fort Frances et de Dryden, en Ontario, au Canada. Dans cette première étude, nous montrons qu'un prédicteur clé de l'occurrence des incendies d'origine humaine fait l'objet d'erreurs de prévision, et que le fait de ne pas en tenir compte produit des estimateurs biaisés ou incorrects. Nous démontrons que la distribution des erreurs de mesure peut être modélisée par des mélanges normaux. Nous utilisons donc un algorithme de simulation-extrapolation (SIMEX) modifié selon lequel l'erreur de prévision dans le prédicteur est un mélange de distributions normales. Nous prenons en compte les modèles logistiques et les modèles de dénombrement puisque nous souhaitons prédire à la fois la présence et le nombre d'incendies. Nous démontrons à l'aide de simulations que le fait de ne pas tenir compte des erreurs de mesure dans un prédicteur entraîne des prévisions biaisées, tandis que les estimateurs SIMEX permettent d'obtenir des améliorations.

[11:35-11:50]

Matthias Schonlau (Department of Statistics and Actuarial Science, University of Waterloo)

Hammock Plots: Visualizing Categorical and Numerical Variables

Graphiques de hamac : visualisation des variables catégorielles et numériques

I discuss the hammock plot for visualizing categorical or mixed categorical/numeric data. Hammock plots can be viewed as a generalization of parallel coordinate plots where the lines are replaced by boxes (or plotting ele-

Je discute le graphique de hamac pour visualiser les données catégorielles ou mixtes catégorielles / numériques. Les graphiques de hamac peuvent être considérés comme une généralisation des graphiques de coordonnées parallèles où les lignes sont rem-

ments) and the width of the boxes is proportional to the number of observations they represent. I also introduce a modification to the hammock plot to avoid what Hoffman et al. termed the reverse line width illusion. Further, I give an historical overview over hammock-type plots such as common angle, GPCP, parsets, and alluvial plots and discuss the type of boxes used to connect adjacent variables.

placées par des boîtes (ou des éléments de traçage) et la largeur des boîtes est proportionnelle au nombre d'observations qu'elles représentent. J'introduis également une modification au graphique du hamac pour éviter ce que Hoffman et al. ont appelé l'illusion de largeur de ligne inverse (≪ reverse line width illusion ≫). De plus, je donne un aperçu historique des graphiques de type hamac tels que l'angle commun, le GPCP, les ensembles parallèles et les graphiques alluviaux et je discute du type de boîtes utilisées pour connecter les variables adjacentes.

Recent Advances in Event History Analysis

Chair/Président: Candemir Cigsar

Organizer/Responsable: Candemir Cigsar

Room/Salle: A 1046

Date: Wednesday June 5 / mercredi 5 juin

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-14:00]

Jerald F. Lawless (University of Waterloo)

Assessing Interventions in Observational Event History Studies

Évaluation des interventions dans les études observationnelles de l'historique des événements

Observational cohort studies are a major source of information about disease and other life history processes. For disease processes interest often lies in adverse events that may occur over time. Interventions involving medication, surgery or adjustments to patient management are often prescribed for cohort members and there is naturally a desire to assess their relative effectiveness in reducing event occurrence. This is complicated, however, by the fact that interventions are prescribed based on an individual's current condition and disease history. To target estimands of average causal effects like those studied in randomized control trials, researchers have proposed using observational data to emulate a randomized intervention study. This has numerous challenges for complex dynamic processes with many time-varying factors. Average causal effects, moreover, have little bearing in real world clinical practice, where treatment decisions are based on the disease course of individual patients. We propose methods based on observable event history data, with adjustment for process history and confounders that may be time-varying. Illustrations for cohorts of persons with rheumatic disease will be discussed. This talk is based on joint work with Richard Cook.

Les études de cohortes observationnelles constituent une source majeure d'informations sur les maladies et d'autres processus du cycle de vie. Pour les processus pathologiques, l'intérêt réside souvent dans les événements indésirables qui peuvent survenir au fil du temps. Des interventions impliquant des médicaments, des interventions chirurgicales ou des ajustements de la prise en charge du patient sont souvent prescrites aux membres de la cohorte et il est naturellement intéressant d'évaluer leur efficacité relative en termes de réduction de l'occurrence d'événements. Cette démarche est toutefois compliquée par le fait que les interventions sont prescrites en fonction de l'état actuel et de l'historique de la maladie d'un individu. Pour cibler les estimations des effets causaux moyens comme ceux étudiés dans les essais contrôlés avec assignation aléatoire, les chercheurs ont proposé d'utiliser des données d'observation pour imiter une étude d'intervention avec assignation aléatoire. Cela pose de nombreux problèmes dans le cas de processus dynamiques complexes comportant de nombreux facteurs variables dans le temps. De plus, les effets causaux moyens sont peu pertinent dans la pratique clinique réelle, car les décisions de traitement sont basées sur l'évolution de la maladie de chaque patient. Nous proposons des méthodes basées sur les données observables de l'historique des événements, avec un ajustement pour l'historique du processus et des facteurs de confusion qui varient dans le temps. Nous discuterons d'illustrations pour des cohortes de personnes atteintes de maladies rhumatismales. Cet exposé est basé sur un travail conjoint avec Richard Cook.

[14:00-14:30]

Joan X. Hu (Simon Fraser University) **Ken Peng** (Simon Fraser University) **Tim B. Swartz** (Simon Fraser University)

An Extended Hawkes Process Model for Recurrent Events

Modèle de processus de Hawkes étendu pour les événements récurrents

Recent Advances in Event History Analysis

Event history data in sports often exhibit special features, including self-exciting, and potential dependence within clusters. The conventional models based on gap times between events may struggle to capture all the complexities. In this project, we propose a novel stochastic process model inspired by the dynamic occurrence of corner kicks in soccer matches. Notice that some corner kicks can lead to the follow-up ones quickly, but it is not always the case. On the other hand, that self-exciting phenomena may potentially only present within a short time frame. The proposed model can accommodate the two features as a promising alternative to the well-known Hawkes model for self-exciting processes. The data from the 2019 regular season of the Chinese Super League are used to illustrate the proposed model and the associated analysis. This is joint work with Ken Peng and Tim Swartz.

Les données d'historique d'événements dans les sports affichent souvent des caractéristiques particulières, y compris une dépendance auto-excitante et potentielle dans les grappes. Il peut être difficile avec les modèles conventionnels basés sur les délais d'attente entre les événements de saisir toutes les complexités. Avec notre projet, nous proposons un nouveau modèle de processus stochastique inspiré par l'occurrence dynamique des corners dans les matchs de soccer. À noter que certains corners peuvent entraîner rapidement d'autres corners, même si ce n'est pas toujours le cas. Par ailleurs, ce phénomène auto-excitant peut potentiellement se manifester pendant un court laps de temps seulement. Le modèle proposé peut prendre en compte les deux caractéristiques comme solution de rechange prometteuse au modèle bien connu de Hawkes pour les processus auto-excitants. Les données de la saison régulière de 2019 de la Chinese Super League illustrent le modèle proposé et l'analyse afférente. En collaboration avec Ken Peng et Tim Swartz.

[14:30-15:00]

Yi Xiong (State University of New York at Buffalo) **Gary Chan** (University of Washington) **Malka Gorfine** (Tel Aviv University) **Li Hsu** (Fred Hutchinson Cancer Center)

Causal Inference in Cost-Effectiveness Analysis with Semi-competing Risks Data

Inférence causale pour les analyses de coût-efficacité avec des données de risques semi-concurrents

Health-care policy makers are often interested in the cost-effectiveness of an intervention. The effectiveness is usually measured by quality adjusted life years, which is subject to informative censoring, and the costs, both of which are often assessed from large-scale observational studies and databases (e.g., claims data, large cohort studies) and are thus susceptible to confounding. There is considerably rich literature available to accommodate censoring and adjust for confounding factors. However, most cost-effectiveness studies are primarily concerned with the terminal event rather than the entire disease progression. This paper is motivated by informing optimal initial screening age for colorectal cancer (CRC) through cost-effectiveness analysis. We provide a unified measure of cost-effectiveness with semi-competing risks and multistate modeling, which allows us to gain insights on benefit and cost at each stage of cancer progression. Unlike most existing causal inference works focusing on static interventions, we develop a causal framework and estimation procedure to evaluate cost-effectiveness as a function of time-varying screening strategy. These methods are justified theoretically and numerically using both simulation and the CRC data from the Women's Health Initiative observational study

Les responsables des politiques en soins de santé s'intéressent souvent au rapport coût-efficacité d'une intervention. Celui-ci est généralement mesurée en utilisant les années de vie ajustées pour tenir compte de la qualité de vie, lesquelles sont soumises à une censure informative, et les coûts. Ces deux valeurs sont fréquemment évalués à partir de bases de données et d'études observationnelles de grande envergure (p. ex. des données de réclamations et de grandes études de cohortes) et donc susceptibles à des variables confondantes. Il existe une littérature très riche permettant de prendre en compte la censure et d'ajuster pour les variables de confusion. Cependant, la plupart des études coût-efficacité s'intéressent principalement à l'événement terminal plutôt qu'à l'ensemble de la progression de la maladie. Le présent article vise à déterminer l'âge optimal du dépistage initial du cancer colorectal (CCR) par le biais d'une analyse coût-efficacité. Nous proposons une mesure unifiée du rapport coût-efficacité avec des risques semi-concurents et une modélisation multi-états, ce qui nous permet d'obtenir des informations sur les avantages et les coûts à chaque étape de la progression du cancer. Contrairement à la plupart des travaux d'inférence causale existants, qui se concentrent sur les interventions statiques, nous développons un cadre causal et une procédure d'estimation pour évaluer le rapport coût-efficacité en fonction d'une stratégie de dépistage variable dans le temps. Ces méthodes sont justifiées théoriquement et numériquement en utilisant à la fois des simu-

Recent Advances in Event History Analysis

lations et les données sur le cancer colorectal de l'étude observationnelle Women's Health Initiative.

Navigating Academic Sabbatical (Panel)
Gérer son congé sabbatique (Table ronde)

Chair/Président: Kuan Liu

Organizer/Responsable: Kevin McGregor

Room/Salle: A 1045

Date: Wednesday June 5 / mercredi 5 juin

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-15:00]

Bei Jiang (University of Alberta) **Olli Saarela** (University of Toronto) **Paul Gustafson** (University of British Columbia) **Matthias Schonlau** (University of Waterloo)

Navigating Academic Sabbatical

Gérer son congé sabbatique

The New Investigator Committee is excited to present a panel session for early-career academics considering or preparing for a sabbatical. Sabbaticals are an important aspect of academic life, offering opportunities for research, collaboration, and professional development. This 90-minute panel session aims to provide valuable insights into planning, completing, and maximizing the benefits of an academic sabbatical. This session will help demystify the process of planning and applying for a sabbatical and cover important topics on setting goals and objectives for a sabbatical, work-life balance, and maintaining lab and current projects during a sabbatical. Our diverse panellists will share tips and tricks for a successful, effective, and rejuvenating academic sabbatical and share their experience and advice on overcoming various challenges, such as travelling and travelling with family. Each panellist will have approximately 15 minutes to share their experiences and insights, followed by a 30-minute interactive Q&A session.

Le comité des nouveaux chercheurs est heureux de présenter une table ronde destinée aux universitaires en début de carrière qui envisagent ou se préparent à prendre un congé sabbatique. Les congés sabbatiques sont un aspect important de la vie universitaire, car ils offrent des possibilités de recherche, de collaboration et de développement professionnel. Cette session de 90 minutes fournira des informations précieuses sur la planification, la réalisation et l'optimisation des avantages d'un congé sabbatique universitaire. Elle aidera à démystifier le processus de planification et de demande d'un congé sabbatique et couvrira des sujets importants sur la définition des objectifs d'un congé sabbatique, l'équilibre entre vie professionnelle et vie privée, et le maintien des projets en cours pendant le congé. Nos différents intervenants partageront conseils et astuces pour un congé sabbatique réussi, efficace et revigorant et partageront leur expérience et leurs conseils pour surmonter les différents défis, tels que les voyages et les déplacements en famille. Chaque panéliste disposera d'environ 15 minutes pour partager son expérience et ses idées, suivies d'une session interactive de questions-réponses de 30 minutes.

Bridging the Gap with Large Language Models (Panel)
Combler le fossé avec de grands modèles linguistiques (table ronde)

Chair/Président: Vahid Partovi Nia

Organizer/Responsable: Vahid Partovi Nia

Room/Salle: A 1043

Date: Wednesday June 5 / mercredi 5 juin

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-15:00]

Alejandro Murua (Université de Montréal) **Pascal Poupart** (Vector Institute, University of Waterloo) **Pierre-Jérôme Bergeron** (Google) **Martin Lysy** (University of Waterloo)

Bridging the Gap with Large Language Models

Combler le fossé avec de grands modèles de langage

We explore the dynamic relationship between statistics and large language models, uncovering the transformative potential of combining statistical methodologies with advanced natural language processing techniques. Data scientists face new challenges and opportunities with the explosion of data in the digital era. New generative models, such as GPT-3.5 deployed in ChatGPT, and a more recent variant deployed in GPT4o, are transforming human society by providing new productivity tools. These models have emerged as powerful data analysis, interpretation, and decision-making tools. This session aims to bridge the gap between traditional statistical methods and the innovative capabilities offered by large language models.

Nous explorons la relation dynamique entre la statistique et les grands modèles de langage, en découvrant le potentiel transformateur de la combinaison des méthodes statistiques avec les techniques avancées de traitement du langage naturel. Avec l'explosion des données à l'ère numérique, les scientifiques des données ont été confrontés à de nouveaux défis et de nouvelles opportunités. De nouveaux modèles génératifs, tels que GPT-3.5 déployé dans ChatGPT, et une variante plus récente déployée dans GPT4o, transforment la société humaine en fournissant de nouveaux outils de productivité. Ces modèles sont devenus de puissants outils d'analyse de données, d'interprétation et de prise de décision. Cette session vise à combler le fossé entre les méthodes statistiques traditionnelles et les capacités innovantes qu'offrent les grands modèles de langage.

Q & A with the Director and Deputy Director of CANSSI
Questions-réponses avec le directeur et la directrice adjointe de l'INCASS

Chair/Président: Donald Estep

Organizer/Responsable: Donald Estep

Room/Salle: ED 2018B

Date: Wednesday June 5 / mercredi 5 juin

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-15:00]

Donald Estep (Simon Fraser University/CANSSI) **Andrea Benedetti** (McGill University)

CANSSI Programs and Plans

Programmes et plans de l'INCASS

The session will start with a presentation about CANSSI programs, including updates on application processes for the CRT and GSES programs. We will also discuss plans for the National Retreat in the Fall. As always, we will answer any questions about CANSSI.

Nous présenterons d'abord les programmes de l'Institut canadien des sciences statistiques (INCASS), ainsi que les nouveautés sur les processus de candidature dans le cadre des programmes des équipes de recherche collaborative et de bourses d'enrichissement des étudiants diplômés. Nous discuterons également des préparatifs de la retraite nationale qui aura lieu en automne. Comme toujours, nous répondrons à toutes les questions concernant l'INCASS.

Pierre Robillard Invited Address
Allocution du recipiendaire du Prix Pierre-Robillard

Chair/Président: Christian Léger

Organizer/Responsable: Christian Léger

Room/Salle: IIC 2001

Date: Wednesday June 5 / mercredi 5 juin

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-15:00]

Qiuqi Wang (Georgia State University)

Standard and comparative e-backtest based on elicitability

Backtests-e standard et comparatifs basés sur l'elicitabilité

Backtesting risk measures is important for financial regulators to evaluate risk forecasts reported by financial institutions. As a natural extension to standard/traditional backtests, comparative backtests are introduced to compare different forecasting methods. Based on recently developed concepts of e-values and e-processes, we design a model-free method for standard backtests of identifiable risk measures. In addition, we develop model-free comparative backtests for elicitable risk measures by constructing e-processes. Our method can be applied to common risk measures including mean, variance, Value-at-Risk, Expected Shortfall, and expectiles. Simulation and real data analysis will be demonstrated as an illustration.

Le backtesting des mesures de risque est important pour les régulateurs financiers chargés d'évaluer les prévisions de risque communiquées par les institutions financières. Extension naturelle des backtests standards/traditionnels, nous introduisons des backtests comparatifs qui permettent de comparer plusieurs méthodes de prévision. Nous nous appuyons sur les concepts récemment développés de valeurs-e et de processus-e pour concevoir une méthode sans modèle pour les backtests standard de mesures de risque identifiables. Nous développons ensuite des backtests comparatifs sans modèle pour les mesures de risque élicitables en construisant des processus-e. Notre méthode peut être appliquée à des mesures de risque courantes, notamment la moyenne, la variance, la valeur à risque, le déficit attendu et les expectiles. Nous présentons à titre d'illustration des simulations et des analyses de données réelles.

Advancing Precision Medicine through Innovative Statistical Methods
Faire progresser la médecine de précision par des méthodes statistiques innovantes

Chair/Président: Cong Jiang

Organizer/Responsable: Cong Jiang

Room/Salle: ED 2018A

Date: Wednesday June 5 / mercredi 5 juin

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-14:00]

Hengrui Cai (University of California, Irvine)

Doubly Robust Interval Estimation for Optimal Policy Evaluation in Online Learning

Estimation d'intervalles doublement robuste pour l'évaluation optimale des politiques dans l'apprentissage en ligne

Evaluating the performance of an ongoing policy plays a vital role in many areas such as precision medicine, to provide crucial instruction on the early-stop of the online experiment and timely feedback from the environment. Policy evaluation in online learning thus attracts increasing attention by inferring the mean outcome of the optimal policy (i.e., the value) in real-time. Yet, such a problem is particularly challenging due to the dependent data generated in the online environment, the unknown optimal policy, and the complex exploration and exploitation trade-off in the adaptive experiment. In this paper, we aim to overcome these difficulties in policy evaluation for online learning. We explicitly derive the probability of exploration that quantifies the probability of exploring the non-optimal actions under commonly used bandit algorithms. We use this probability to conduct valid inference on the online conditional mean estimator under each action and develop the doubly robust interval estimation (DREAM) method to infer the value under the estimated optimal policy in online learning. The proposed value estimator provides double protection on the consistency and is asymptotically normal with a Wald-type confidence interval provided. Extensive simulations and real data applications are conducted to demonstrate the empirical validity of the proposed DREAM method.

Il est essentiel dans bien des domaines de pouvoir évaluer les résultats d'une politique en cours, comme en médecine de précision, afin de comprendre l'effet d'un arrêt précoce d'une expérience en ligne et obtenir un retour d'information opportun de l'environnement. On s'intéresse donc de plus en plus à l'évaluation des politiques dans l'apprentissage en ligne ; il s'agit alors de déduire le résultat moyen de la politique optimale (sa valeur) en temps réel. Cependant, ce problème est particulièrement difficile à résoudre, puisque l'environnement en ligne génère des données dépendantes, la politique optimale est inconnue et les expériences adaptatives nécessitent un compromis complexe entre exploration et exploitation. Dans cet article, nous visons à surmonter ces difficultés. Nous dérivons explicitement la probabilité d'exploration, à savoir la probabilité d'explorer les actions non optimales dans les algorithmes de bandit couramment utilisés. Nous utilisons cette probabilité pour effectuer une inférence valide sur l'estimateur de la moyenne conditionnelle en ligne pour chaque action et développons la méthode d'estimation par intervalle doublement robuste (DREAM) pour déduire la valeur de la politique optimale estimée dans le cadre de l'apprentissage en ligne. L'estimateur de valeur proposé offre une double protection sur la convergence et est asymptotiquement normal avec un intervalle de confiance de type Wald. Nous menons des simulations approfondies et des applications sur données réelles pour démontrer la validité empirique de la méthode DREAM proposée.

[14:00-14:30]

Dylan Spicker (University of New Brunswick)

Infinite and Irregular: Developments for Dynamic Treatment Regimes with Stochastic Decision Points

Infini et irrégulier : Développements pour les régimes de traitement dynamiques avec des points de décision stochastiques

Dynamic treatment regimes (DTRs) can be used to formalize precision medicine in the longitudinal setting. DTRs are sequences of functions that take in patient information and produce treatment recommendations. When estimating DTRs, typically the goal is to optimize an outcome of interest, in expectation. These sequences of decisions allow for the longitudinal treatment of complex conditions in a personalized manner. Historically, the estimation of optimal DTRs has relied upon several assumptions that limit the utility of DTRs for certain applications. In this work, I am concerned with a priori assumptions regarding the finite, deterministic schedule of treatments. I will discuss developments in optimal DTR estimation that relax these assumptions allowing for stochastic numbers of treatments with covariate-driven observation times. The talk will highlight the importance of these procedures, challenges with their implementation, and methods that have shown promise for their resolution.

Les régimes de traitement dynamiques peuvent être utilisés pour concrétiser la médecine de précision dans un cadre longitudinal. Les régimes de traitement dynamiques sont des séquences de fonctions qui prennent en compte les renseignements sur le patient et produisent des recommandations de traitement. Lors de l'estimation des régimes de traitement dynamiques, on cherche en général à optimiser l'espérance d'un résultat souhaité. Ces séquences de décisions permettent le traitement longitudinal de conditions complexes de manière personnalisée. Traditionnellement, l'estimation des régimes de traitement dynamiques optimaux repose sur plusieurs hypothèses qui limitent le potentiel des régimes de traitement dynamiques pour certaines applications. Dans le cadre de ces travaux, je me penche sur les hypothèses a priori concernant le calendrier fini et déterministe des traitements. Je présenterai les avancées en matière d'estimation optimale des régimes de traitement dynamiques, qui assouplissent ces hypothèses en tenant compte des nombres stochastiques de traitements avec des temps d'observation déterminés par des covariables. Je parlerai de l'importance de ces procédures, des problèmes liés à la mise en œuvre de celles-ci et des méthodes qui se sont révélées prometteuses pour les résoudre.

Restoring Survey Response Rates and Mitigating the Nonresponse Error: Current Approaches and Recent Findings
Rétablir les taux de réponse aux enquêtes et atténuer l'erreur de non-réponse : approches actuelles et résultats récents

Chair/Président: Peter G. Wright

Organizer/Responsable: Peter G. Wright

Room/Salle: C 2045

Date: Wednesday June 5 / mercredi 5 juin

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-14:00]

Emma Troughton (Statistics Canada) **Peter Wright** (Statistics Canada)

Recent Initiatives to Assess the Potential Nonresponse Error in Social Surveys

Initiatives récentes pour évaluer l'erreur potentielle due à la non-réponse des enquêtes sociales

Declining response rates to social surveys increase the potential of nonresponse error. To respond to this challenge, Statistics Canada is undertaking a comprehensive assessment using a variety of approaches that have so far provided considerable insight into this kind of error. These approaches have included the use of supplementary data to assess the impact on estimates of the Labour Force Survey, as well as the implementation of a follow-up survey of nonrespondents using multi-mode collection and minimal respondent burden to encourage participation. These studies form part of Project to study citizen participation, a multi-year initiative to improve the response rates and to mitigate the risk and impact of nonresponse error throughout the survey process.

Les taux de réponse décroissants aux enquêtes sociales augmentent le potentiel d'erreur due à la non-réponse. Pour faire face à ce défi, Statistique Canada entreprend une évaluation complète en utilisant une variété d'approches qui ont jusqu'à présent fourni un aperçu considérable de ce type d'erreur. Ces approches comprennent l'utilisation de données supplémentaires afin d'évaluer l'impact sur les estimations de l'Enquête sur la population active, ainsi que la mise en oeuvre d'une enquête de suivi des non-répondants qui utilise une collecte multi-modale et minimise le fardeau de réponse pour encourager la participation. Ces études font partie du Projet d'étude de la participation citoyenne, une initiative pluriannuelle pour améliorer les taux de réponse et pour atténuer les risques et l'impact de l'erreur due à la non-réponse à travers le processus d'enquête.

[14:00-14:30]

France Lapointe (Institut de la Statistique du Québec) **Éric Gagnon** (Institut de la Statistique du Québec)

Are Satistical Surveys of Individuals (Once Again) at a Crossroads?

Les enquêtes statistiques auprès des individus (encore) à la croisée des chemins ?

The Institut de la statistique du Québec typically uses the Régie de l'assurance maladie du Québec's Fichier d'inscription des personnes assurées [insured persons enrollment file] as the sampling frame for its surveys of individuals. In 2021, new legislative provisions made it possible for the Institut to obtain information from public bodies to carry out its mission. In this presentation, we will provide examples of statistical surveys that significantly improved their response rates by leveraging these newly accessible administrative sources. We will also discuss the Institut's methodological work to better understand the limitations that stem from its growing

Le plus souvent, l'Institut de la statistique du Québec utilise le Fichier d'inscription des personnes assurées (FIPA) de la Régie de l'assurance maladie du Québec (RAMQ) comme base de sondage pour ses enquêtes auprès d'individus. Depuis 2021, de nouvelles dispositions législatives permettent à l'Institut d'obtenir des renseignements d'un organisme public afin de les utiliser pour la réalisation de sa mission. Dans la présentation, on donnera des exemples d'enquêtes statistiques qui ont pu profiter des sources administratives maintenant accessibles afin d'améliorer notablement leurs taux de réponse. On abordera également les travaux méthodologiques effectués par l'Institut pour mieux comprendre les limites découlant des difficultés croissantes qu'il éprouve à

challenges to survey certain subpopulations. Finally, we will present the results of tests conducted in winter 2024 to optimize our strategies for contacting and following-up with individuals.

mener des enquêtes auprès de certaines sous-populations. Enfin, on présentera les résultats de tests menés à l'hiver 2024 pour améliorer les stratégies de contact et de relance des individus.

[14:30-15:00]

Hélène Chaput (Insee (France))

Mixed-mode collection of household surveys in France: difficulties and opportunities

Mixed-mode collection of household surveys in France : difficulties and opportunities

Since 2010, INSEE has been moving its household surveys from an almost exclusively face-to-face approach to a mixed-mode approach. The first aim is to modernise survey collection by improving efficiency. The second one is to encourage households to respond by making it easier for them through a wider range of protocols. Mixed-mode data collection obviously raises new issues in terms of survey logistics, questionnaire design, statistical processing and even sampling if we want to have the means to control mode effects. These new developments therefore represent both difficulties specific to the use of mixed-mode surveys and opportunities. This paper looks back over several years of implementation and software development aimed at collecting mixed-mode household surveys.

Depuis 2010, l'Insee a entrepris de passer ses enquêtes auprès des ménages d'une collecte quasi exclusivement face-à-face à une collecte multimode. L'objectif est de moderniser la collecte des enquêtes en améliorant son efficience, et d'inciter les enquêtés à répondre en leur facilitant cette réponse par une palette accrue de modalités. La collecte multimode pose évidemment des questions nouvelles de logistique d'enquête, de conception de questionnaire, de traitement statistique, et même d'échantillonnage si on veut se doter des moyens de contrôle des effets de mode. Ces nouveautés constituent donc autant des difficultés propres à l'exploitation des enquêtes multimodes que des opportunités. Ce papier propose un retour sur plusieurs années de mise en oeuvre et de développements logiciels destinés à collecter des enquêtes ménages par multimode.

Convergence of MCMC Algorithms
Convergence des algorithmes MCMC

Chair/Président: Jeffrey Negrea

Organizer/Responsable: Jeffrey S. Rosenthal

Room/Salle: A 2071

Date: Wednesday June 5 / mercredi 5 juin

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-14:00]

Trevor Campbell (The University of British Columbia) **Nikola Surjanovic** (University of British Columbia) **Saifuddin Syed** (Oxford University) **Alexandre Bouchard-Côté** (University of British Columbia)

An Exploration-agnostic Characterization of the Ergodicity of Parallel Tempering

Une caractérisation exploration-agnostique de l'ergodicité de l'atténuation parallèle

Non-reversible parallel tempering (NRPT) is an effective algorithm for sampling from target distributions with complex geometry, such as those arising from posterior distributions of weakly identifiable and high-dimensional Bayesian models. In this talk I will establish the uniform geometric ergodicity of NRPT under an efficient local exploration hypothesis, which avoids the intricacies of dealing with kernel-specific properties. The rates that we obtain are bounded in terms of an easily-estimable divergence, the global communication barrier (GCB), that was recently introduced in the literature. We obtain analogous ergodicity results for classical reversible parallel tempering, providing new evidence that NRPT dominates its reversible counterpart. I will conclude the talk with simulations that validate the new theoretical analysis.

L'atténuation parallèle irréversible (NRPT) est un algorithme efficace pour l'échantillonnage de distributions cibles comportant une géométrie complexe, comme celles générées par des distributions a posteriori de modèles bayésiens de haute dimension et faiblement identifiables. Dans le cadre de cette présentation, j'établirai l'ergodicité géométrique uniforme de la NRPT selon une hypothèse d'exploration locale efficace, tout en évitant la gestion complexe des propriétés spécifiques aux noyaux. Les taux obtenus sont limités en termes de divergence facilement estimable, c'est-à-dire l'obstacle à la communication globale (GCB) qui a récemment été présentée dans la documentation. Nous obtenons des résultats d'ergodicité analogues pour l'atténuation parallèle réversible classique, ce qui est une nouvelle preuve que la NRPT est supérieure à sa version réversible. Je conclurai la présentation avec des simulations qui valident la nouvelle analyse théorique.

[14:00-14:30]

Gareth O. Roberts (University of Warwick) **Jeffrey S. Rosenthal** (University of Toronto) **Nick Tawn** (University of Warwick)

Parallel Tempering Schemes and Robustness to Dimensionality

Schémas d'atténuation parallèle et robustesse à la dimensionalité

This talk will overview recently work on parallel tempering algorithms and present results on their performance in high-dimensional settings in stylised examples where limit results can be obtained. From a methodological point of view, the presentation will present and analyse the ALPS algorithm and give theoretical justification for its good theoretical properties in high-dimensions.

Cet exposé donnera un aperçu des travaux récents sur les algorithmes d'atténuation parallèle et présentera des résultats sur leurs performances dans des contextes à haute dimension, dans des exemples stylisés où des résultats limites peuvent être obtenus. D'un point de vue méthodologique, l'exposé présentera et analysera l'algorithme ALPS et donnera une justification théorique de ses bonnes propriétés théoriques en haute dimension.

[14:30-15:00]

Jeffrey S. Rosenthal (University of Toronto)

Experiments with MCMC Tempering Options

Expérimentations avec options de tempérage par chaîne de Markov Monte-Carlo (MCMC)

Simulated and parallel tempering MCMC algorithms are widely used to move between separated modes of the target distribution. Various options are available, including dynamically adjusted temperature spacings (Atchadé, R., and Roberts, 2011; Roberts and R., 2014); the even/odd momentum-inducing scheme (Okabe et al., 2001; Syed et al., 2021); the QuanTA mode-mapping adjustment (Tawn and Roberts, 2019; Roberts, R., and Tawn 2022); the ALPS cold-temperature variant (Tawn et al., 2021; Roberts, R., and Tawn 2022); and what assumptions are made about the within-chain mixing rates. In this talk, we will review these options, and present recent simulation experiments comparing their efficiency in various examples.

Les algorithmes MCMC de tempérage simulé et parallèle sont largement utilisés pour passer entre des modes séparés de la distribution cible. Diverses options sont disponibles, y compris les espacements de température adaptés de manière dynamique (Atchadé, R.et Roberts, 2011 ; Roberts et R., 2014) ; le système pair-impair inductif d'impulsion (Okabe et al., 2001 ; Syed et al., 2021) ; l'ajustement de mappage de modes QuanTA (Tawn et Roberts, 2019 ; Roberts, R. et Tawn 2022) ; la variante de température froide ALPES (Tawn et al., 2021 ; Roberts, R. et Tawn 2022) ; et les hypothèses posées au sujet des taux de mélange au sein de la chaîne. Nous passons en revue ces options et présentons des expériences de simulation pour en comparer l'efficacité dans divers exemples.

Prediction Using Different Models
Prédiction à l'aide de différents modèles

Chair/Président: Julie Carreau

Room/Salle: A 2065

Date: Wednesday June 5 / mercredi 5 juin

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-13:45]

Michael Guerzhoy (University of Toronto) **Max Piasevoli** (Princeton University, Microsoft Corporation) **Tracy Qian** (University of Toronto)

Automatic Model Selection using Wasserstein Generative Adversarial Networks

Sélection automatique de modèles à l'aide de réseaux antagonistes génératifs de Wasserstein

We propose a novel approach for automatic model selection for hierarchical models using Wasserstein Generative Adversarial Networks (WGANs). Model checking and selection can be performed by graphically comparing fake data generated by the proposed model to the actual data. The aim is to select a model that generates fake data with a similar distribution to that of the actual data. The critic component of a WGAN is trained to discriminate data generated by the generator component from the real data. We propose using the critic components of WGANs trained on data simulated from candidate models. If the critic component of a WGAN for a candidate model cannot successfully discriminate between synthetic data generated from that model and the real data, that indicates better model fit. We describe an algorithm for model selection using this intuition. We demonstrate that our approach can be used to select appropriate models for synthetic and real social science datasets.

Nous proposons une nouvelle approche pour la sélection automatique de modèles hiérarchiques à l'aide de réseaux antagonistes génératifs de Wasserstein (WGAN). La vérification et la sélection des modèles peuvent être effectuées en comparant graphiquement les données fictives générées par le modèle proposé aux données réelles. L'objectif est de sélectionner un modèle qui génère des données fictives dont la distribution est similaire à celle des données réelles. La composante critique d'un WGAN est entraînée à distinguer les données générées par la composante génératrice des données réelles. Nous nous proposons d'utiliser les composantes critiques des WGAN entraînés sur des données simulées à partir de modèles candidats. Si la composante critique d'un WGAN pour un modèle candidat ne peut pas discriminer avec succès les données synthétiques générées par ce modèle et les données réelles, cela indique une meilleure adéquation du modèle. Nous décrivons un algorithme de sélection de modèle fondé sur cette intuition. Nous démontrons que notre approche peut être utilisée pour sélectionner des modèles appropriés pour des ensembles de données synthétiques et réelles en sciences sociales.

[13:45-14:00]

Funmilola Mary Taiwo (University of Manitoba)

Bayesian multiclass approach for predicting student dropout

Approche bayésienne multiclasse pour prédire le décrochage scolaire

This study investigates the application of Bayesian models in predicting student dropout across multiple classes. Traditional methods often struggle with the complexity of multiclass scenarios and fail to capture uncertainty adequately. Leveraging Bayesian frameworks, our proposed model offers a robust solution by incorporating prior knowledge, handling uncertainty, and providing probabilistic predictions. Through comparative analysis

Dans cette étude, nous examinons l'application de modèles bayésiens pour prédire le décrochage d'étudiants dans plusieurs classes. Il est souvent difficile pour les méthodes traditionnelles de gérer la complexité des scénarios multiclasses et de tenir compte de l'incertitude de manière adéquate. Le modèle que nous proposons, qui s'appuie sur des cadres bayésiens, offre une solution robuste puisqu'il intègre des connaissances préalables, gère l'incertitude et fournit des prédictions probabilistes. Grâce à une ana-

with existing methods, including logistic regression and support vector machines, we demonstrate the superior predictive performance of our Bayesian approach. The findings underscore the efficacy of Bayesian models in multiclass dropout prediction.

lyse comparative avec les méthodes actuelles, dont la régression logistique et les machines à vecteurs de support, nous démontrons que notre approche bayésienne est plus efficace sur le plan de la prédiction. Les résultats mettent en évidence l'efficacité des modèles bayésiens dans la prédiction du décrochage scolaire multiclasse.

[14:00-14:15]

Yuxuan Zhao (University of Waterloo)

Inference for time-delay differential equations

Inférence pour les équations différentielles à retard

Modeling dynamic systems often involves feedbacks among components. However, there is time delay for these systems to sense and respond to feedbacks. Delay differential equations (DDEs) are commonly for these purposes. Normally, the model equations and noisy observations from the dynamic systems are provided while parameters are unknown and required for estimation. We extend manifold-constrained Gaussian process inference (MAGI) to conduct parameter inference in DDEs using noisy and sparse observations. This method imposes a Gaussian process model over a time series data conditional on the manifold constraint that the DDEs must be satisfied under a Bayesian framework. To get a computational-efficiency algorithm, linear interpolation is applied to approximate the values of the lagged state variables. In this work, we also obtain some error bounds for the derivatives of the state variables along with relevant simulation results to justify the approximation method is valid. Moreover, we present two simulation examples, including the Hutchinson equation and the lac operon system, and a real-world application using Ontario COVID data, to illustrate the efficiency of our model.

La modélisation de systèmes dynamiques nécessite souvent des rétroactions entre les composants. Cependant, ces systèmes détectent les rétroactions et y réagissent avec un certain retard. Les équations différentielles à retard sont souvent utilisées à cette fin. Normalement, les équations du modèle et les observations bruitées des systèmes dynamiques sont fournies alors que les paramètres sont inconnus et doivent être estimés. Nous étendons l'inférence de processus gaussien à variété contrainte à l'inférence de paramètres dans les équations différentielles à retard au moyen d'observations bruyantes et éparses. Cette méthode applique un modèle de processus gaussien aux données d'une série temporelle sous réserve de la contrainte de la variété selon laquelle les équations différentielles à retard doivent être satisfaites dans un cadre bayésien. Pour obtenir un algorithme computationnel efficace, une interpolation linéaire est appliquée pour approximer les valeurs des variables d'état retardées. Dans le cadre de ces travaux, nous obtenons également des limites d'erreur pour les dérivées des variables d'état, ainsi que des résultats de simulation pertinents pour justifier la validité de la méthode d'approximation. Enfin, nous présentons deux exemples de simulation, y compris l'équation de Hutchinson et le système de l'opéron lac, ainsi qu'une application réelle avec des données de COVID de l'Ontario afin d'illustrer l'efficacité de notre modèle.

[14:15-14:30]

Megan French (Department of Statistics and Actuarial Science, University of Waterloo) **Ryan Browne** (University of Waterloo)

Block Diagonal Gaussian Mixture Models

Modèles de mélange gaussien en blocs diagonaux

We investigate methods of estimating a block diagonal structure of Gaussian covariance matrices. The space of all possible partitions of variables is large and infeasible to search in large dimensions. Through hierarchical clustering we are able to restrict the search space and obtain a sequence of nested partitions. We propose a Ward procedure where each partition is chosen based on the maximal log-likelihood. We compare the performance

Nous étudions les méthodes qui permettent d'estimer une structure en blocs diagonaux de matrices de covariance gaussiennes. L'espace de toutes les partitions possibles des variables est vaste et il est impossible d'y effectuer une recherche en grandes dimensions. Grâce au regroupement hiérarchique, nous sommes en mesure de restreindre l'espace de recherche et d'obtenir une séquence de partitions imbriquées. Nous proposons une procédure de Ward où chaque partition est choisie sur la base de la log-vraisemblance

of this method to single, average, and complete linkage hierarchical clustering based on the sample correlation matrix in a simulation study. We then extend this methodology to Gaussian mixture models and demonstrate its performance through a simulation study and data analysis.

maximale. Dans une étude de simulation, nous comparons les performances de cette méthode à celles du regroupement hiérarchique à lien unique, moyen et complet basé sur la matrice de corrélation de l'échantillon. Nous étendons ensuite cette méthodologie aux modèles de mélange gaussien et en démontrons les performances par une étude de simulation et une analyse de données.

[14:30-14:45]

Robert Zimmerman (University of Toronto) **David A. van Dyk** (Imperial College London) **Vinay L. Kashyap** (Harvard & Smithsonian) **Aneta Siemiginowska** (Harvard & Smithsonian)

Separating Flaring and Quiescent States in Active Coronae using State-Space Models

Séparation des états en éruption et en quiescence des couronnes actives à l'aide de modèles d'espace d'états

Astronomers are often interested in classifying the states of active coronae as either flaring or quiescent. Because these states are not directly observable, they must be inferred only from the characteristics of emitted photons captured by detectors. It would be useful to develop a model that can estimate parameters related to the timing of a star's flaring activity, and thus help to elucidate the underlying physical processes that drive stellar flares. We take a state-space modelling approach to this problem by considering the underlying physical process within the star as a latent Markov chain on a continuous state-space; the distribution of the observed photon counts at any time depends implicitly on the state of the underlying chain at that same time. The state predictions from this model are then dichotomized to produce flaring classifications. We demonstrate these techniques using data from the active dMe flare star EV Lac.

Les astronomes s'intéressent souvent à la classification des états (en éruption et en quiescence) des couronnes actives. Ces états n'étant pas directement observables, ils ne peuvent être déterminés qu'à partir des caractéristiques des photons émis captés par des détecteurs. C'est pourquoi il serait utile de développer un modèle capable d'estimer les paramètres liés à la chronologie de l'activité éruptive d'une étoile, et donc de contribuer à élucider les processus physiques sous-jacents qui sont à l'origine des éruptions stellaires. Ainsi, nous adoptons une approche de modélisation d'espace d'état pour ce problème en prenant en compte le processus physique sous-jacent de l'étoile, comme une chaîne de Markov latente sur un espace d'état continu (la distribution des comptes de photons observés à tout moment dépend implicitement de l'état de la chaîne sous-jacente à ce même moment). Les prédictions d'état de ce modèle sont ensuite dichotomisées pour produire des classifications d'éruption. Nous démontrons ces techniques à l'aide de données provenant de l'étoile EV Lacertae (étoile éruptive de type spectral M).

[14:45-15:00]

Owen G. Ward (Simon Fraser University)

Statistical Network Analysis with Aggregated Relational Data

Analyse statistique de réseaux avec des données relationnelles agrégées

Statistical network analysis has been a fruitful direction of research in recent years. However, in many settings, the exact edges of such a network may not be available. In such scenarios it may instead be possible to collect aggregate relational data, detailing the number of edges between a node and specific subpopulations present in a network. Existing probabilistic models for network data can then be extended to this setting. In this talk we will provide an introduction to statistical models for aggregate relational data before considering new results for community detection for data of this form. In particular, we will examine how to perform community detection for large networks along with examining the em-

L'analyse statistique des réseaux a été un axe de recherche prometteur au cours des dernières années. Toutefois, dans de nombreux contextes, il se peut que les arêtes exactes d'un tel réseau ne soient pas disponibles. Dans de tels scénarios, il peut être possible de collecter des données relationnelles agrégées, détaillant le nombre d'arêtes entre un nœud et des sous-populations spécifiques présentes dans un réseau. Il est alors possible d'étendre à ce contexte les modèles probabilistes actuels pour les données de réseau. Dans cette présentation, nous aborderons les modèles statistiques pour les données relationnelles agrégées avant d'examiner de nouveaux résultats pour la détection de communautés pour des données de cette forme. Nous examinerons en particulier la façon de détecter les communautés pour les grands réseaux, ainsi

pirical properties required for community detection in these models.

que les propriétés empiriques requises pour la détection des communautés dans ces modèles.

Statistical Models for Clinical and Healthcare Data
Modèles statistiques pour les données cliniques et de santé

Chair/Président: Marie-Pierre Sylvestre

Room/Salle: C 2033

Date: Wednesday June 5 / mercredi 5 juin

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-13:45]

Kehinde I. Olobatuyi (University of Victoria) **Laura L.E. Cowen** (University of Victoria) **Patrick Brown** (University of Toronto) **Matthew Parker** (Simon Fraser University)

Multi-state Dynamic Capture-Recapture model for Big Data: Estimating undetected COVID-19 Cases in British Columbia, Canada

Modèle dynamique de capture et de recapture multi-états pour données volumineuses : estimation des cas de COVID-19 non détectés en Colombie-Britannique, Canada

The accurate quantification of the impact of the COVID-19 pandemic on both public health and the economy is essential for informed policy-making. However, the true scope of the pandemic remains challenging to ascertain due to undetected cases, particularly when relying on reported cases, which are contingent on test availability and strategies. Previous approaches, such as early computation of infection fatality ratios based on confirmed cases and deaths, proved to be flawed, given the dynamic nature of COVID-19 over time. In this study, we develop a multi-state model to capture the dynamics of COVID-19. Using individual-level information from the popdataBC database, we estimate the case detection probability, infection probability, and recovery probability. The analysis of this extensive dataset prompts a discussion on the computational challenges encountered and the methodologies employed to address them. Our application provides an estimate of the total COVID-19 burden in year 2020.

La quantification précise de l'impact de la pandémie de COVID-19 sur la santé publique et l'économie est essentielle pour une prise de décision éclairée. Toutefois, il reste difficile de déterminer l'ampleur réelle de la pandémie en raison des cas non détectés, en particulier lorsque l'on s'appuie sur les seuls cas signalés, ce qui dépend de la disponibilité des tests et des stratégies. Les approches précédentes, telles que le calcul précoce des taux de létalité de l'infection sur la base des cas confirmés et des décès, se sont révélées erronées, compte tenu de la nature dynamique du virus. Dans cette étude, nous développons un modèle multi-états pour capturer la dynamique du COVID-19. En utilisant les informations individuelles de la base de données popdataBC, nous estimons la probabilité de détection des cas, la probabilité d'infection et la probabilité de guérison. L'analyse de ce vaste ensemble de données suscite une discussion sur les défis informatiques rencontrés et les méthodologies employées pour les relever. Notre application fournit une estimation de la charge totale de COVID-19 en 2020.

[13:45-14:00]

Razvan G. Romanescu (University of Manitoba)

Epidemic Spread Dynamics on Associative Networks

Dynamique de propagation des épidémies sur les réseaux associatifs

Compartmental models of disease spread have been well studied on networks built according to the Configuration Model, i.e., where the degree distribution of individual nodes is specified, but where connections are made randomly. Dynamics of spread on such 'first order' networks were shown to be profoundly different compared

Les modèles compartimentaux de propagation des maladies ont été bien étudiés sur des réseaux construits selon le modèle de configuration, c'est-à-dire où la distribution des degrés des nœuds individuels est spécifiée, mais où les connexions sont faites de manière aléatoire. La dynamique de la propagation sur ces réseaux de « premier ordre » s'est révélée profondément différente de

to epidemics under the traditional mass action assumption. Assortativity, i.e., the preferential mixing of nodes according to degree, is a second order property that is thought to impact epidemic trajectory. We first show how assortative mixing can come about from individual preferences to connect with others of lower or higher degree, and propose an algorithm for constructing such a network. We then investigate via simulation how this network structure favors or inhibits diffusion processes, such as the spread of an infectious disease. Finally, we propose some analytic results to characterize the mean behavior of diffusion processes over this network type.

celle des épidémies sous l'hypothèse traditionnelle de l'action de masse. L'assortativité, c'est-à-dire le mélange préférentiel des nœuds en fonction de leur degré, est une propriété de second ordre dont on pense qu'elle a un impact sur la trajectoire de l'épidémie. Nous montrons d'abord comment le mélange assortatif peut résulter des préférences individuelles à se connecter avec d'autres personnes de degré inférieur ou supérieur, et nous proposons un algorithme pour construire un tel réseau. Nous étudions ensuite par simulation comment cette structure de réseau favorise ou inhibe les processus de diffusion, tels que la propagation d'une maladie infectieuse. Enfin, nous proposons des résultats analytiques pour caractériser le comportement moyen des processus de diffusion sur ce type de réseau.

[14:00-14:15]

Jianchu Chen (University of Waterloo) **Richard J. Cook** (University of Waterloo)

Cost-effective design of Survival Studies Involving Intermittent Observation of Time-Dependent Covariates

Conception économique d'études de survie comportant une observation intermittente de covariables dépendantes du temps

Electronic medical record databases offer an unprecedented opportunity to study chronic diseases. In survival analysis interest may lie in studying the effects of time-dependent biomarkers on death through Cox regression models. However, collecting and cleaning data on all covariates at all times can be labor-intensive, and in such settings it's common to select a single clinic visit at which covariates are measured. Using large sample theory involving misspecified models, we investigate the large sample bias of estimators based on ad hoc strategies for selecting the assessment time including using a) the last visit time before failure or censoring, b) the last visit time under left-truncation, and c) the first visit time. Alternative selection schemes are discussed for efficient selection of individuals and time-points for measurement of covariates to enable joint modeling and consistent parameter estimation - methods range from simple random sampling to outcome dependent sampling.

Les bases de données de dossiers médicaux électroniques offrent une possibilité sans précédent d'étudier les maladies chroniques. Dans l'analyse de la survie, l'intérêt peut être d'étudier les effets des biomarqueurs dépendants du temps sur la mortalité en utilisant des modèles de régression de Cox. Cependant, comme il peut être laborieux de collecter et nettoyer les données de toutes les covariables en tout temps, il est courant dans un tel contexte de sélectionner une seule clinique pour laquelle les covariables sont mesurées. En utilisant la théorie de grand échantillonnage comportant des modèles mal spécifiés, nous étudions le biais des estimateurs de grand échantillon, basé sur des stratégies ad hoc de sélection du temps d'évaluation, y compris a) le dernier temps de visite avant la défaillance ou la censure ; b) le dernier temps de visite sous troncature gauche ; et c) le premier temps de visite. Nous abordons d'autres modes pour la sélection efficace d'individus et de points temporels afin de mesurer les covariables pour permettre une modélisation conjointe et une estimation cohérente des paramètres – les méthodes allant de l'échantillonnage aléatoire simple à l'échantillonnage dépendant des résultats.

[14:15-14:30]

Renny Doig (Simon Fraser University) **Liangliang Wang** (Simon Fraser University)

Auxiliary-try Metropolis: Incorporating Auxiliary Variables into Multiple-try Metropolis

Metropolis essai-auxiliaire : intégration de variables auxiliaires à un Metropolis à essai multiple

The multiple-try Metropolis (MTM) algorithm extends the Metropolis-Hastings algorithm by considering multiple candidate proposal draws at each iteration. This can improve exploration of the state space and reduce convergence times. However, single MTM chains struggle to explore spaces with multimodal surfaces. A pre-

L'algorithme Metropolis à essai multiple (MTM) élargit l'algorithme Metropolis-Hastings en tenant compte de plusieurs candidatures tirées à chaque itération. Cette façon peut améliorer l'exploration spatio-temporelle et réduire le temps de convergence. Par contre, les chaînes MTM simples peinent à explorer les espaces avec des surfaces multimodales. D'ailleurs, une approche

viously developed approach to tackle the problem of multimodality incorporated multiple interacting MTM chains. In our work, we propose a general framework called auxiliary-try Metropolis, in which candidate proposals can be generated conditional on variables auxiliary to the chain. Within this framework one can augment MTM with information independent of the current chain. We consider multiple auxiliary variable frameworks and compare them with existing interacting-chain MTM as well as other sampling methods which have been shown to perform well when sampling from multimodal distributions.

développée antérieurement pour résoudre le problème multimodal a justement intégré des chaînes d'interaction MTM. Dans le cadre de notre travail, nous proposons un cadre général appelé « Metropolis Essai-Auxiliaire », dans lequel les candidatures peuvent être générées de façon conditionnelle aux variables auxiliaires à la chaîne. Dans ce cadre, il est possible d'accroître la MTM avec de l'information indépendante de la chaîne actuelle. Nous évaluons plusieurs cadres de variable auxiliaire et les comparons avec la MTM à chaîne d'interaction actuelle ainsi qu'avec d'autres méthodes d'échantillonnage ayant été démontrées comme efficaces pour l'échantillonnage à partir de distributions multimodales.

[14:30-14:45]

Audrey Béliveau (University of Waterloo) **Xiangshan Kong** (University of Waterloo)

Generalized Fused Lasso for Treatment Pooling in Network Meta-Analysis

Fused lasso généralisé pour la mise en commun des traitements dans la méta-analyse en réseau

This work develops a generalized fused lasso (GFL) approach to fitting contrast-based network meta-analysis (NMA) models. The GFL method penalizes all pairwise differences between treatments, resulting in the pooling of treatments that are not sufficiently different. This approach offers an intriguing avenue for potentially mitigating biases in treatment rankings and multiple comparison issues. A simulation study confirms the ability of the GFL approach to pool treatments that have the same (or similar) effects while also revealing when incorrect pooling may occur and its potential benefits. Finally, the novel GFL-NMA method is applied to real-world datasets on Parkinson's and diabetes. In both cases, the full (standard) NMA model was not favored compared to the best-fitting GFL-NMA model with AICc selection of the tuning parameter (difference in AICc larger than 9).

Ce travail développe une approche de fused lasso généralisé (GFL) pour l'ajustement des modèles de méta-analyse en réseau (NMA) basés sur les contrastes. La méthode GFL pénalise toutes les différences par paire entre traitements, ce qui permet de regrouper les traitements qui ne sont pas suffisamment différents. Cette approche offre une possibilité intéressante d'atténuer les biais dans le classement des traitements et les problèmes de comparaisons multiples. Nous confirmons par une étude de simulation la capacité de l'approche GFL à regrouper les traitements qui ont des effets identiques (ou similaires), tout en révélant les cas où un regroupement incorrect peut se produire et ses avantages potentiels. Enfin, nous appliquons la nouvelle méthode GFL-NMA à des ensembles de données réelles sur la maladie de Parkinson et le diabète. Dans les deux cas, le modèle NMA complet (standard) est inférieur au modèle GFL-NMA le mieux ajusté avec la sélection AICc du paramètre d'accord (différence d'AICc supérieure à 9).

Statistical Challenges: Privacy, High-Dimensionality, and Real-World Applications
Défis statistiques : confidentialité, haute dimensionnalité et applications concrètes

Chair/Président: Zeinab Mashreghi

Organizer/Responsable: Zeinab Mashreghi

Room/Salle: A 1043

Date: Wednesday June 5 / mercredi 5 juin

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-16:00]

Anne-Sophie Charest (Université Laval) **Mamadou Mbodj** (Université Laval) **Sébastien Gambs** (Université du Québec à Montréal)

Exploit Membership Inference attacks and Imputation strategies for Attribute disclosure from Estimated models

Exploiter les attaques d'inférence d'appartenance et les stratégies d'imputation pour la divulgation d'attributs à partir de modèles estimés

Remote-access servers allow analysts to obtain statistics, model parameters and other outputs estimated from confidential data, but without access to the actual dataset. While this strategy reduces disclosure risk, they remain sensitive to other attacks, such as membership inference attacks, which can predict from the released model whether a specific observation was part of the training dataset. This attack may be harmful on its own, for example if all individuals in the dataset share a sensitive attribute such as a disease, but it can also enable other attacks. Here, we explain how to perform attribute disclosure by using such membership attacks. We illustrate the success of the approach with experimental results on random forests trained to solve a classification task on the confidential data. We also explore the relationship between such attacks and data imputation and combine imputation strategies with the proposed attack to improve our attribute disclosure success rate.

Les serveurs d'accès à distance permettent aux analystes d'obtenir des statistiques, paramètres de modèles et autres résultats sans accès direct au jeu de données. Cette stratégie réduit le risque de divulgation, mais reste sensible aux attaques d'appartenance, qui prédisent à partir du modèle obtenu si une observation faisait partie ou non des données utilisées. Ceci peut être préjudiciable en soi si tous les individus partagent un attribut sensible tel qu'une maladie, et peut être utilisé pour permettre d'autres attaques. Nous montrons ici comment faire une divulgation d'attributs avec une attaque d'appartenance. Nous illustrons le succès de la stratégie à l'aide de résultats expérimentaux sur des forêts aléatoires entrainées pour résoudre une tâche de classification. Nous explorons également la relation entre ces attaques et l'imputation, et combinons des stratégies d'imputation avec l'attaque proposée pour améliorer notre taux de divulgation.

[16:00-16:30]

Mehdi Dagdoug (McGill University) **David Haziza** (University of Ottawa) **Esther Eustache** (Université de Neuchâtel)

High-dimensional Variance Estimation for Linear Model-assisted Estimation and Linear Imputation

Estimation de la variance en grande dimension pour l'estimation assistée par modèle linéaire et l'imputation linéaire

In surveys, auxiliary variables are often used at the estimation stage through the use of a predictive model. In this work, we consider the problem of variance estimation for both model-assisted and imputed estimators based on linear regression when the number of covariates is non-negligible with respect to the sample size.

Dans les enquêtes, on utilise souvent des variables auxiliaires au stade de l'estimation via un modèle prédictif. Dans ce travail, nous examinons le problème de l'estimation de la variance pour les estimateurs assistés par un modèle et les estimateurs imputés basés sur la régression linéaire lorsque le nombre de covariables n'est pas négligeable par rapport à la taille de l'échantillon. Nous

We show that, in both scenarios, customary variance estimators such as Taylor linearization and its g-weighted version are biased negatively, while the Jackknife variance estimator is biased positively. These biases are shown to be substantial, even asymptotically, leading to inconsistent variance estimators. Under suitable conditions, we obtain a closed-form expression for these biases with quantities that can be estimated from the observed data. We present simulation studies showing the ill behavior of these traditional variance estimators in high-dimensional settings and that their bias-corrected versions seem to behave well.

montrons que, dans les deux cas, les estimateurs de variance habituels tels que la linéarisation de Taylor et sa version pondérée g sont biaisés négativement, tandis que l'estimateur de variance Jackknife est biaisé positivement. Ces biais se révèlent substantiels, même asymptotiquement, et conduisent à des estimateurs de variance non convergents. Dans des conditions appropriées, nous obtenons une expression de forme compacte pour ces biais avec des quantités qui peuvent être estimées à partir des données observées. Nous présentons des études de simulation qui montrent le mauvais comportement des estimateurs de variance traditionnels dans des contextes de haute dimension et leurs versions corrigées des biais, qui semblent bien fonctionner.

[16:30-17:00]

Augustine Wigle (University of Waterloo) **Audrey Béliveau** (University of Waterloo)

Estimating Provincial Methane Emissions from Complex Survey Data using a Multi-Stage Framework

Estimation des émissions de méthane à l'échelle provinciale à partir de données d'enquête complexes en utilisant un cadre à plusieurs phases

Measurement-based methane inventories, where oil and gas facilities are surveyed and these data are compiled to estimate total methane emissions, are becoming the gold standard for quantifying emissions. There is a current lack of statistical guidance for the design and analysis of such surveys. The only existing method is a Monte Carlo (MC) procedure which is difficult to interpret, computationally intensive, and open-source code for its implementation is not available. We provide an alternative method by showing that a methane survey corresponds to a multi-stage sampling design. We present estimators of the total emissions and its variance which do not require simulation. We show that the variance contribution from each stage of sampling can be estimated and can inform the design of future surveys. We also introduce a modification of the estimator which is more efficient. Finally, we propose combining the multi-stage approach with a simple MC procedure to model measurement error.

Les inventaires du méthane basés sur la mesure, lorsque les installations pétrolières et gazières sont étudiées et que les données sont compilées pour l'estimation des émissions totales de méthane, deviennent la méthode de référence pour la quantification des émissions. Il manque actuellement de lignes directrices statistiques pour la conception et l'analyse de telles enquêtes. La seule méthode est une procédure de Monte-Carlo (MC) difficile à interpréter, intensive sur le plan computationnel et sans aucun code source ouvert disponible pour son implémentation. Nous fournissons une méthode alternative en montrant qu'une étude portant sur le méthane correspond à un plan d'échantillonnage à plusieurs phases. Nous présentons des estimateurs de la quantité totale d'émissions et de sa variance qui ne requiert pas de simulation. Nous montrons que la contribution de la variance à chaque phase de l'échantillonnage peut être estimée et servir à la conception d'enquêtes subséquentes. Nous présentons aussi une modification de l'estimateur qui est ainsi plus efficace. Nous proposons enfin de combiner l'approche à plusieurs phases avec une procédure MC simple pour modéliser l'erreur de mesure.

Recent Advances in Methods for Incomplete and Complex Survey Data
Progrès récents sur les méthodes des données d'enquête incomplètes et complexes

Chair/Président: Wendy Lou

Organizer/Responsable: Wendy Lou

Room/Salle: C 2045

Date: Wednesday June 5 / mercredi 5 juin

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-16:00]

Trevor James Thomson (Fred Hutchinson Cancer Center) **Joan X. Hu** (Simon Fraser University)

On Developing a Predictive Survival Model with Internal Time-varying Covariates

Développement d'un modèle de survie prédictif avec covariables internes variant avec le temps

Many health-related studies aim to explore how time-to-death is associated with changes in health status. Noting that health status is an internal time-varying covariate, likelihood-based procedures are thus inapplicable in this application. With the objective of providing risk predictions based on one's current health status, this presentation discusses strategies to obtain such predictions. We summarize the incomplete internal covariate process with latent variable(s), and adopt a joint modelling framework that links the survival and health outcome processes together. The proposed estimation procedure extends the conditional score approach of Tsiatis and Davidian (2001) by allowing the longitudinal sub-model(s) to accommodate correlated successive observations. We motivate and illustrate the proposed modelling with a dataset from administrative health records, and demonstrate adequate predictive performance through a simulation study. This is joint work with X. Joan Hu (SFU).

Bon nombre d'études en matière de santé visent à explorer la façon dont le délai jusqu'au décès (TTD) est associé aux changements dans l'état de santé. Compte tenu que l'état de santé est une covariable interne variant avec le temps, conséquemment les procédures basées sur la vraisemblance ne peuvent pas servir dans cette application. L'objectif de cette présentation étant de fournir des prédictions de risque basées sur l'état de santé actuel, nous discutons des stratégies pour obtenir de telles prédictions. Nous résumons le traitement des covariables internes incomplètes avec variable(s) latente(s) et adoptons un cadre de modélisation conjointe qui lie le traitement des résultats de survie et de ceux de santé. La procédure d'estimation proposée étend la méthode du score conditionnel de Tsiatis et Davidian (2001) en permettant un ou des sous-modèle(s) longitudinaux pour prendre en compte les observations successives corrélées. Nous motivons et illustrons la modélisation proposée avec un ensemble de données provenant de dossiers sur l'administration de la santé et en montrons la performance prédictive adéquate à l'aide d'une étude en simulation. En collaboration avec X. Joan Hu (SFU).

[16:00-16:30]

Zilin Wang (Wilfrid Laurier University) **Mary Thompson** (University of Waterloo)

Modelling Missing-not-at-random for Mental Health Data from Complex Surveys

Modélisation de données manquantes de manière non aléatoire pour données d'enquêtes complexes sur la santé mentale

In the analysis of incomplete data in complex surveys, the probability of being non-missing (propensity score) could depend on the study variable, a missing mechanism called missing-not-at-random (MNAR). In this paper, we model the propensity score with the MNAR mechanism by assuming that the probability of missingness is not homogeneous across the support of the study

Lors de l'analyse de données incomplètes dans le cadre d'enquêtes complexes, la probabilité de non-réponse (score de propension) peut dépendre de la variable à l'étude, un mécanisme de non-réponse appelé « manquant de manière non aléatoire » (MNAR). Dans cet article, nous modélisons le score de propension avec le mécanisme MNAR en supposant que la probabilité de non-réponse n'est pas homogène sur l'ensemble du support de la

variable. We assign different probabilities of missingness conditional on observed values in subsets of support of the study variable. Using the observed weighted likelihood function and the estimating equation methods, we estimate and make inferences on the parameters of the distribution of the study variable, the probability of missingness within each subset, and the propensity scores of the study variables. We apply this method to analyze mental health data from the National Nutrition Examination Survey.

variable à l'étude. Nous attribuons différentes probabilités de non-réponse en fonction des valeurs observées dans des sous-ensembles du support de la variable d'étude. En utilisant la fonction de vraisemblance pondérée observée et les méthodes d'équations d'estimation, nous estimons et faisons de l'inférence sur les paramètres de la distribution de la variable à l'étude, la probabilité de non-réponse dans chaque sous-ensemble et les scores de propension des variables de l'étude. Nous appliquons cette méthode aux données sur la santé mentale de la National Nutrition Examination Survey.

[16:30-17:00]

Changbao Wu (University of Waterloo)

Analysis of Complex Data through Combining Information from Multiple Sources

Analyse de données complexes par la combinaison d'informations tirées de plusieurs sources

We discuss issues with combining information from multiple sources under settings involving a probability sample, a non-probability sample, known population controls from a census, or estimated population controls from other surveys. Two main principles, namely, validity and efficiency, are discussed under different scenarios. Inferential techniques are presented under the general framework of calibration and empirical likelihood methods.

Nous étudions les problèmes liés à la combinaison d'informations tirées de plusieurs sources dans un cadre comprenant un échantillon probabiliste, un échantillon non probabiliste, les comptes démographiques connu d'un recensement, ou des estimations de comptes démographiques provenant d'autres enquêtes. Nous abordons deux principes centraux selon différents scénarios : la validité et l'efficacité. Nous présentons des techniques d'inférence selon le cadre général de calibration et des méthodes de vraisemblance empirique.

Innovations in Statistical Modeling for Complex Data Structures
Innovations en modélisation statistique des structures de données complexes

Chair/Président: Cindy Feng

Organizer/Responsable: Cindy Feng

Room/Salle: ED 2018A

Date: Wednesday June 5 / mercredi 5 juin

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-16:00]

Jiguo Cao (Simon Fraser University) **Barinder Thind** (Simon Fraser University) **Kevin Multani** (Stanford University)

Functional Neural Networks

Réseaux neuronaux fonctionnels

We present a methodology for integrating functional data into deep neural networks. The model is defined for scalar responses with multiple functional and scalar covariates. A by-product of the method is a set of dynamic functional weights that can be visualized during the optimization process. This visualization leads to greater interpretability of the relationship between the covariates and the response relative to conventional neural networks. The model is shown to perform well in several contexts, including prediction of new data and recovery of the true underlying relationship between the functional covariate and scalar response; these results were confirmed through real data applications and simulation studies.

Nous présentons une méthodologie pour intégrer des données fonctionnelles dans des réseaux neuronaux profonds. Le modèle est défini pour des réponses scalaires avec plusieurs covariables fonctionnelles et scalaires. Un sous-produit de la méthode est un ensemble de poids fonctionnels dynamiques qui peuvent être visualisés pendant le processus d'optimisation. Cette visualisation conduit à une meilleure interprétation de la relation entre les covariables et la réponse par rapport aux réseaux neuronaux conventionnels. Le modèle est démontré comme performant dans plusieurs contextes, y compris la prédiction de nouvelles données et la récupération de la véritable relation sous-jacente entre la covariable fonctionnelle et la réponse scalaire ; ces résultats ont été confirmés par des applications de données réelles et des études de simulation.

[16:00-16:30]

Muye Nanshan (Simon Fraser University) **Nan Zhang** (Fudan University) **Jiguo Cao** (Simon Fraser University)

Online Functional Principal Component Analysis on a Multidimensional Domain with Dynamic Tuning

Analyse en composantes principales fonctionnelles en ligne sur un domaine multidimensionnel avec réglage dynamique

Functional Principal Component Analysis (FPCA) is an essential dimension reduction tool for functional data. The emergence of streaming or large-scale multidimensional functional datasets has highlighted the demand for an online FPCA approach. This work leverages Riemannian Stochastic Gradient Descent (RSGD) for an efficient online update of the principle components with minimal computational effort. Furthermore, we adjust the tuning parameter dynamically during the online estimation process using a novel evaluation metric, the Averaged Block Validation (ABV) score, and an innovative beam search technique. Theoretical backing

L'analyse en composantes principales fonctionnelles (ACPF) est un outil essentiel de réduction des dimensions pour les données fonctionnelles. L'émergence d'ensembles de données fonctionnelles multidimensionnelles en continu ou à grande échelle a mis en évidence la nécessité d'une approche ACPF en ligne. Ce travail s'appuie sur la descente stochastique du gradient de Riemann (RSGD) pour une mise à jour en ligne efficace des composantes principales avec un effort de calcul minimal. En outre, nous ajustons le paramètre de réglage de manière dynamique pendant le processus d'estimation en ligne à l'aide d'une nouvelle mesure d'évaluation, le score ABV (Averaged Block Validation), et d'une technique innovante de recherche par faisceau. La conver-

for the convergence of the RSGD algorithm with dynamic tuning is provided. Simulation studies and applications to two datasets reveal our method's effectiveness in quickly processing large datasets and accurately estimating FPCs.

gence de l'algorithme RSGD avec réglage dynamique est étayée théoriquement. Des études de simulation et des applications à deux ensembles de données révèlent l'efficacité de notre méthode dans le traitement rapide de grands ensembles de données et dans l'estimation précise des ACP.

[16:30-17:00]

Lam Ho (Dalhousie University)

Modelling and Inferring Phenotypic Trait Evolution on Large Phylogenetic Trees

Modélisation et inférence de l'évolution des caractères phénotypiques sur de grands arbres phylogénétiques

Inferring covariation between multiple biological traits sampled across numerous related species is at the heart of evolutionary biology. To adjust for the shared evolutionary history among organisms, trait evolution models assume phenotypes evolve along the phylogenetic tree according to stochastic processes. In this talk, we will examine general Gaussian models of trait evolution. Making inferences under these models on large trees is extremely challenging due to the strong correlation between observations. An additional challenge arises as missing data are prevalent with a large number of species. We will discuss a highly efficient computational technique that scales linearly with the number of species and illustrates its application using real-life data.

L'inférence de la covariation entre de multiples traits biologiques échantillonnés sur de nombreuses espèces apparentées est au cœur de la biologie évolutive. Pour tenir compte de l'histoire évolutive commune des organismes, les modèles d'évolution des traits supposent que les phénotypes évoluent le long de l'arbre phylogénétique selon des processus stochastiques. Dans cet exposé, nous examinerons les modèles gaussiens généraux de l'évolution des caractères. Faire des inférences avec ces modèles sur de grands arbres est extrêmement difficile en raison de la forte corrélation entre les observations. Un défi supplémentaire se pose lorsque les données manquantes sont fréquentes pour un grand nombre d'espèces. Nous discuterons d'une technique de calcul très efficace qui s'étend linéairement avec le nombre d'espèces et nous illustrerons son application sur des données réelles.

Approaches to Teaching the Analysis of Large-Scale and Complex Data
Approches de l'enseignement de l'analyse de données complexes et à grande échelle

Chair/Président: Alexander Shestopaloff

Organizer/Responsable: Alexander Shestopaloff

Room/Salle: A 1046

Date: Wednesday June 5 / mercredi 5 juin

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-15:52]

Xu (Sunny) Wang (Wilfrid Laurier University) **Sukhjit Sehra** (Wilfrid Laurier University) **Devan G. Becker** (Wilfrid Laurier University)

A Six-Year Journey - Overview of Laurier Data Science Program - Collaboration and Creativity

Un parcours de six ans : Aperçu du programme de science des données à la Wilfrid Laurier University (WLU) – collaboration et créativité

The BSc. Honours in Data Science program was inaugurated in Fall 2018, marking the commencement of a six-year journey. Throughout this period, as instructors, we have fostered collaboration, creatively delivering lectures and labs even amidst the challenges posed by the COVID-19 pandemic. In our upcoming presentation, we will delve into the following key aspects: (1) Providing an overarching view of the Laurier Data Science Program. (2) Exploring the collaborative approach adopted, emphasizing the integration of lectures and labs. (3) Showcasing our innovative use of online resources to enhance the learning experience. (4) Reflecting on the valuable lessons garnered throughout this journey. (5) Delving into the envisioned future trajectory of the data science program. We look forward to sharing insights and engaging in fruitful discussions on the evolution and prospects of the Data Science Program at Laurier.

Le baccalauréat ès sciences spécialisé (BSc. Honours) au programme de science des données a été inauguré à l'automne 2018, marquant ainsi le début d'un parcours de six ans. Pendant toute cette période, à titre d'enseignants, nous avons favorisé la collaboration, présentant de façon créative des cours et des laboratoires, et ce, malgré tous les problèmes que posait la pandémie de la COVID-19. Notre exposé permet de voir de plus près certains des aspects principaux du programme, soit : (1) fournir une vue d'ensemble du programme de science des données Laurier ; (2) explorer l'approche collaborative adoptée, en nous attardant à l'intégration des cours et des laboratoires ; (3) mettre en lumière l'utilisation innovante des ressources en ligne pour améliorer l'apprentissage ; (4) réfléchir sur les leçons précieuses tirées de ce parcours ; et (5) envisager la trajectoire pour l'avenir du programme de science des données. Nous sommes impatients de partager des points de vue et d'engager des discussions fructueuses sur l'évolution et les perspectives d'avenir de ce programme à la WLU.

[15:52-16:15]

Pierre Miasnikof (University of Toronto) **Cristian Bravo** (University of Western Ontario) **Yuri Lawryshyn** (University of Toronto)

Statistics for networks and networks for statistics – why statisticians should know graphs

Statistique pour les réseaux et les réseaux pour la statistique - pourquoi les statisticiens devraient être familiers avec les graphes

The aim of this presentation is to introduce the statistics community to network data and its importance. With recent advances in data collection technologies and the ubiquitous presence of networks around us, new and

L'objectif de cette présentation est d'initier la communauté des statisticiens aux données de réseaux et à leur importance. Avec les récents progrès des technologies de collecte de données et l'omniprésence des réseaux autour de nous, de nouveaux ensembles

typically very large data sets are becoming available and require analysis. Unfortunately, many traditional statistical tools are not applicable to data collected from networks. It is important that modern statisticians become familiar with network data and its challenges, in order to develop the tools to analyze it. We will begin with a brief overview of network science topics. I will then introduce two statistical problems that lie at the core of network analysis. NB. This talk is intended for a wide audience from the mathematical sciences. I will not assume any prior knowledge of graph theory or network science.

de données, généralement très volumineux, se présentent à l'analyse. Malheureusement, la plupart des outils statistiques traditionnels ne sont pas adaptés à ce type de données. Il est important que les statisticiens modernes se familiarisent avec les données de réseaux et les défis qu'elles posent, afin de développer des outils en d'analyse. Nous commencerons par un bref aperçu de la problématique de la science des réseaux. Je présenterai ensuite deux problèmes statistiques qui sont au cœur de l'analyse de réseaux. NB. Cette présentation s'adresse à un large public issu des sciences mathématiques. Je ne supposerai aucune connaissance préalable de la théorie des graphes ou de la science des réseaux.

[16:15-16:37]

G. Alexi Rodríguez-Arelis (The University of British Columbia)

Multiclass Prediction and Inference: A Practical Approach

Prédiction et inférence multi-classes : une approche pratique

Regression modelling is a vast statistical field comprising various approaches that might suit different inferential and predictive practical cases. In this context, when teaching data analysis in an accelerated data science graduate program, it is crucial to establish an efficient and homogeneous workflow that can cater to a wide range of regression approaches using data science-based reproducible tools (such as Jupyter notebooks) along with engaging datasets. This talk will explain this analysis workflow while providing crucial insights on its application in a regression graduate course beyond ordinary least squares. Finally, under an inferential and predictive scenario, this teaching approach will be exemplified via a specific model targeted to multiclass nominal responses: multinomial regression.

La régression est un vaste champ d'étude statistique englobant diverses approches qui peuvent convenir à différents cas pratiques de prédiction et d'inférence. Pour l'enseignement de l'analyse de données dans le cadre d'un programme d'études accéléré en science des données, il est essentiel dans ce contexte d'établir un flux de travail efficace et homogène qui peut traiter une vaste gamme de régressions à l'aide d'outils reproductibles basés sur la science des données (comme les carnets Jupyter) et d'ensembles de données intéressants. Nous expliquons la séquence des tâches d'analyse, en plus de fournir des points de vue éclairés sur son application dans un cours universitaire sur la régression au-delà de la méthode ordinaire des moindres carrés. Enfin, selon un scénario inférentiel et prédictif, une régression multi-nominale, soit un modèle spécifique ciblé sur des réponses nominales multi-classes, servira à illustrer cette approche d'enseignement.

[16:37-17:00]

Varada Kolhatkar (The University of British Columbia)

Making Machine Learning Approachable in Data Science Education

Rendre l'apprentissage automatique accessible dans l'enseignement de la science des données

Ever heard a student say: "I'm eager to dive into machine learning, but after skimming through the CPSC 340 course slides, I feel overwhelmed. Am I really cut out for this?" or "I understand the theory behind Convolutional Neural Networks, but when I open my laptop to code, it feels like I'm trying to write a book in a language I don't speak." In today's world, machine learning (ML) permeates various aspects of daily life, from spam filtering and voice recognition to advanced recommendation systems and large language models (LLMs). Its widespread application sparks a growing curiosity

Avez-vous déjà entendu un étudiant dire : « Je suis impatient de me plonger dans l'apprentissage automatique, mais après avoir parcouru les diapositives du cours CPSC 340, je me sens dépassé. Suis-je vraiment fait pour cela ? » ou « Je comprends la théorie derrière les réseaux neuronaux convolutifs, mais lorsque j'ouvre mon ordinateur portable pour coder, j'ai l'impression d'essayer d'écrire un livre dans une langue que je ne parle pas ». Dans le monde d'aujourd'hui, l'apprentissage automatique imprègne divers aspects de la vie quotidienne, du filtrage des pourriels et de la reconnaissance vocale aux systèmes de recommandation avancés et aux grands modèles de langage (LLM). Son applica-

among a diverse audience eager to understand the fundamentals of ML and its underlying mechanisms. I'll share my approach to demystifying complex machine learning concepts and transforming them from abstract theories into tangible skills within our Master of Data Science (MDS) program. By focusing on making these intricate topics accessible, I aim to equip students with the ability not just to comprehend, but also competently apply these concepts in real-world scenarios, ensuring that when they next sit down to apply these concepts, it feels like reading their favourite book in their native language.

tion généralisée suscite de plus en plus de curiosité de la part d'un public diversifié désireux d'en comprendre les principes fondamentaux et ses mécanismes sous-jacents. Je présenterai une approche pour démystifier les concepts complexes de l'apprentissage automatique et les transformer en compétences tangibles dans le cadre de notre programme de maîtrise en science des données. En tâchant de rendre ces sujets complexes plus accessibles, je vise à aider les étudiants non seulement à comprendre, mais aussi à appliquer avec compétence ces concepts dans des scénarios du monde réel, en veillant à ce que, lorsqu'ils s'assoient ensuite pour appliquer ces concepts, ils aient l'impression de lire leur livre préféré dans leur langue maternelle.

Chair/Président: Haixu Wang

Room/Salle: A 1049

Date: Wednesday June 5 / mercredi 5 juin

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-15:45]

Jairo Diaz-Rodriguez (York University) **Kelly Ramsay** (York University)

Differentially Private Boxplots

Boîtes à moustaches différentiellement confidentielles

Despite the potential of differentially private data visualization to harmonize data analysis and privacy, research in this area remains relatively underdeveloped. Boxplots are a valuable tool for summarizing dataset's distribution, aiding in the comparison of multiple datasets and the identification of statistical variations or outliers. We introduce a novel differentially private boxplot that facilitates the analysis of confidential data retaining the characteristics of boxplot visualization. We evaluate its effectiveness using several proposed boxplot distance metrics. Theoretical and simulation properties, along with real-world data analysis, validate the performance of our methodology, showing comparable analysis can be achieved through differentially private boxplot visualization. We provide practical guidance on privacy budget and limitations. We suggest our methodology could lay the groundwork for future research in the field of differentially private data visualization.

Malgré le potentiel de la visualisation des données différentiellement confidentielles dans le but d'harmoniser l'analyse et la confidentialité des données, la recherche dans ce domaine est relativement sous-développée. Les boîtes à moustaches sont un outil précieux pour résumer la distribution des ensembles de données, en aidant à la comparaison d'ensembles de données multiples et à l'identification de variations statistiques ou de valeurs aberrantes. Nous présentons une nouvelle boîte à moustaches différentiellement confidentielle qui facilite l'analyse des données confidentielles retenant les caractéristiques de la visualisation de la boîte à moustaches. Nous évaluons son efficacité à l'aide de plusieurs mesures de distance de la boîte à moustaches proposée. Des propriétés théoriques et simulées et une analyse de données du monde réel valident la performance de notre méthodologie, montrant qu'une analyse comparable peut être faite par la visualisation d'une boîte à moustaches différentiellement confidentielle. Un guide pratique sur le budget et les limites de la confidentialité est fourni. Nous avançons que notre méthodologie peut ouvrir la voie à de futures recherches en matière de visualisation de données différentiellement confidentielles.

[15:45-16:00]

Amanjot Bhullar (University of Guelph) **Khurram Nadeem** (University of Guelph) **Ayesha Ali** (University of Guelph) **Evan D. G. Fraser** (University of Guelph)

DeepS^3: A Deep Learning Framework for Predicting Multi-Crop Land Suitability with Satellite Data

DeepS^3 : un cadre d'apprentissage profond pour prédire la qualité de la terre pour plusieurs récoltes avec des données satellites

Assessing land suitability is key to agriculture management, as it helps identify the capability of different lands to cultivate different crops. Here, we present DeepS^3, a multilayer perceptron framework that simultaneously predicts the land suitability for multiple crops based on satellite imagery, crop-specific farm location data, and

L'évaluation de la qualité de la terre est essentielle à la gestion d'agriculture, car elle aide à identifier la capacité de différentes terres à cultiver différentes récoltes. Nous présentons ici DeepS^3, un cadre de perceptron multicouche qui prédit simultanément la qualité de la terre pour plusieurs récoltes en fonction d'une imagerie satellite, de données spécifiques à une récolte dans une région

district level crop production census data. Our method can be viewed as a multi-task learning method that enables a multivariate model in the presence of unobserved responses. We exploit the multi-crop response during training and capture the interdependencies among different crops, thereby facilitating extrapolation. This framework easily generalizes to other spatial problems that incorporate data obtained at diverse spatial resolutions. Applying DeepS^3 to predict crop land suitability for Canada under climate change projects diminishing suitability for canola, peas, wheat, and soy in the Prairie Provinces, mainly driven by increased heat stress.

agricole, et de données de recensement de la production de récolte à l'échelle régionale. Notre méthode peut être perçue comme une méthode d'apprentissage multitâche qui favorise un modèle multivarié en présence de réponses non observées. Nous exploitons la réponse à récolte multiple durant l'apprentissage et capturons les interdépendances entre les différentes récoltes, facilitant ainsi l'extrapolation. Ce cadre se généralise facilement pour résoudre d'autres problèmes spatiaux qui comprennent des données obtenues à diverses résolutions spatiales. En utilisant DeepS^3 pour prédire la qualité des terres agricoles au Canada avec les changements climatiques, on constate une réduction de la qualité pour le canola, les pois, le blé et le soya dans les provinces des Prairies, principalement en raison de la chaleur.

[16:00-16:15]

Neve Loewen (University of Manitoba) **Mohammad Jafari Jozani** (University of Manitoba)

Robust Regression Analysis with Nomination Sampling

Analyse de régression robuste avec échantillonnage nominatif

We introduce a novel methodology for robust regression analysis for when traditional mean regression falls short due to the presence of heteroscedasticity or outliers. Unlike conventional approaches that rely on simple random sampling (SRS), our methodology leverages median nomination sampling (MedNS), using readily available rank information to obtain training data that more accurately captures the central tendency of the underlying population, enhancing the representativeness of the sample in the presence of outliers. We propose a new loss function that integrates the rank information of MedNS data during the training phase of model fitting, offering a form of robust regression. Through simulation studies and a real data application, we evaluate the efficacy of our proposed approach by comparing the integrated mean squared error of regression estimates. We observe that our proposed method provides higher relative efficiency compared to its SRS counterparts when outliers are present.

Nous présentons une nouvelle méthodologie pour l'analyse de régression robuste lorsque la régression moyenne traditionnelle ne suffit pas en raison de la présence d'hétéroscédasticité ou de valeurs aberrantes. Contrairement aux approches conventionnelles qui reposent sur l'échantillonnage aléatoire simple (EAS), notre méthodologie s'appuie sur l'échantillonnage nominatif de la médiane (MedNS), en utilisant les informations de rang facilement disponibles pour obtenir des données d'entraînement qui capturent plus précisément la tendance centrale de la population sous-jacente, améliorant ainsi la représentativité de l'échantillon en présence de valeurs aberrantes. Nous proposons une nouvelle fonction de perte qui intègre les informations de rang des données MedNS pendant la phase d'apprentissage de l'ajustement du modèle, offrant ainsi une forme de régression robuste. Par des études de simulation et à une application de données réelles, nous évaluons l'efficacité de l'approche proposée en comparant l'erreur quadratique moyenne intégrée des estimations de régression. Nous observons que la méthode proposée offre une efficacité relative supérieure à celle de ses homologues EAS en présence de valeurs aberrantes.

Prediction and Learning
Prédiction et apprentissage

Chair/Président: Shirin Golchi

Room/Salle: A 2071

Date: Wednesday June 5 / mercredi 5 juin

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-15:45]

Ajmery Jaman (McGill University)

UPoSI Approach to Valid Post-Selection Inference for Penalized G-Estimation

Approche UPoSI de l'inférence post-sélection valide pour l'estimation G pénalisée

Effect modifier selection is important in the modeling of heterogeneous treatment effects. With data-driven selection of effect modifiers, the quantification of statistical uncertainty is complicated by post-selection inference. Since the selection is performed based on a single sample dataset, the naive inference procedures suffer from overfitting and usually produce an inflated type I error rate. We propose a Universal Post Selection Inference (UPoSI) procedure for the recently developed penalized G-estimator that performs doubly robust estimation of the causal effect of a time-varying exposure with a simultaneous data-adaptive selection of effect modifiers using longitudinal observational data. We perform a simulation study to compare the empirical false coverage rates produced by the proposed method with that obtained from the naive inference procedure based on sandwich variance estimates. We illustrate our method using the data arising from a study of hemodiafiltration.

La sélection des modificateurs d'effet est importante dans la modélisation des effets hétérogènes du traitement. Lors de cette sélection, la quantification de l'incertitude statistique est compliquée en raison de l'inférence post-sélection. La sélection étant effectuée à partir d'un seul échantillon de données, les procédures d'inférence naïve présentent un surajustement et produisent généralement un taux d'erreur de type I élevé. Nous proposons une procédure d'inférence post-sélection universelle (UPoSI) pour le G-estimateur pénalisé récemment mis au point, qui réalise une estimation doublement robuste de l'effet causal d'une exposition variable dans le temps avec une sélection simultanée de modificateurs d'effet adaptée aux données, en utilisant des données d'observation longitudinales. Nous réalisons une étude de simulation pour comparer les taux de fausse couverture empiriques produits par la méthode proposée avec ceux obtenus par la procédure d'inférence naïve basée sur les estimations de la variance sandwich. Nous illustrons notre méthode avec les données d'une étude sur l'hémodiafiltration.

[15:45-16:00]

Minzee Kim (University of Waterloo) **Joel A. Dubin** (University of Waterloo)

A Method for Improving Dynamic Prediction of Joint Models Using a Similarity-Based Approach

Une méthode pour améliorer la prédiction dynamique de modèles conjoints au moyen d'une approche basée sur la similarité

Longitudinal and time-to-event data are often analyzed in biomarker research to study the association between the longitudinal biomarker measurements and the event-time outcome, in which the longitudinal information contributes to the probability of the outcome of interest. An attractive nature of fitting a joint model on this type of data is that we can dynamically predict the survival probability as additional longitudinal information becomes available. We propose a new similarity-based

Les données longitudinales et de durée de vie sont souvent analysées en recherche sur les biomarqueurs afin d'étudier l'association entre les mesures de biomarqueur longitudinales et le résultat de type temps d'événement, où l'information longitudinale contribue à la probabilité du résultat recherché. L'un des attraits d'ajuster un modèle conjoint à ce type de données est que l'on peut prédire de façon dynamique la probabilité de survie au fur et à mesure que de nouvelles informations longitudinales s'ajoutent. Nous proposons une nouvelle méthode basée sur la similarité pour

method in the dynamic prediction of joint models where we consider training the model on only a targeted subset of the data to obtain an improved outcome prediction. Through a comprehensive simulation study, we demonstrate that the predictive performance of the dynamic prediction of joint models can be improved with our proposed similarity-based approach.

la prédiction dynamique de modèles conjoints où le modèle apprend à partir d'un sous-ensemble de données cibles seulement dans le but d'obtenir une prédiction supérieure. Par l'entremise d'une étude de simulation approfondie, nous démontrons que la performance prédictive de la prédiction dynamique des modèles conjoints peut être améliorée grâce à notre approche basée sur la similarité.

[16:00-16:15]

Tatiana Krikella (University of Waterloo) **Joel A. Dubin** (University of Waterloo)

Determination of Subpopulation Size for Similarity-Based Personalized Prediction Models that Jointly Optimize Discrimination and Calibration

Détermination de la taille de sous-population pour les modèles de prédiction personnalisés basés sur la similarité qui optimisent conjointement la discrimination et la calibration

Precision medicine is accelerating rapidly in the health research field. We previously proposed an algorithm which fits a predictive model for an individual patient based on patient similarity using the size of a similar subpopulation that optimizes both model discrimination and calibration. The optimal size of subpopulation is currently tuned through grid search, where the grid values are defined through trial and error to ensure sufficient sample size in the validation step of our algorithm. We propose using simulation-based sample size determination methods to define the lower bound on the grid. These methods determine the minimum sample size needed to precisely estimate the desired performance measures in the validation step. We investigate the relationship between the size of subpopulation and the performance measures under different outcome models, and quantify the uncertainty in the estimation. A comprehensive set of simulation studies are presented to demonstrate our findings.

La médecine de précision croît de façon fulgurante dans le domaine de recherche en santé. Précédemment, nous avons proposé un algorithme qui s'ajuste à un modèle prédictif pour un patient basé sur la similarité du patient au moyen d'une taille d'une sous-population similaire qui optimise la discrimination et la calibration du modèle. La taille optimale d'une sous-population est présentement réglée par l'entremise de recherche par quadrillage, dans laquelle les valeurs du quadrillage sont définies par essai-erreur pour assurer une taille d'échantillon suffisante pour l'étape de validation de notre algorithme. Nous proposons l'adoption de méthodes de détermination de la taille d'échantillon basée sur la simulation dans le but de définir la limite inférieure du quadrillage. Ces méthodes déterminent la taille d'échantillon minimum requise pour précisément estimer les mesures de performance voulues à l'étape de validation. Nous examinons le lien entre la taille de la sous-population et les mesures de performance selon différents modèles de résultat, puis quantifions l'incertitude dans l'estimation. Nous présentons un ensemble approfondi d'études en simulations pour démontrer nos découvertes.

[16:15-16:30]

Li-Pang Chen (National Chengchi University)

Variable Selection and Estimation for Length-Biased and Partly Interval-Censored Survival Data with Mismeasured Covariates

Sélection de variables et estimation pour les données de survie avec censure par intervalle et biais de longueur en présence de covariables mal mesurées

In this talk, we analyze the length-biased and partly interval-censored data, whose challenges primarily come from biased sampling and interfere induced by interval censoring. Moreover, the variables in the dataset are high-dimension and are subject to measurement error. To address those challenges, we explore a valid inference method for handling high-dimensional length-biased and interval-censored survival data with measurement error in covariates under the accelerated fail-

Lors de cet exposé, nous analysons les données avec censure partielle par intervalle et biais de longueur dont la difficulté principale est représentée par l'échantillonnage biaisé et l'interférence causée par la censure par intervalle. En outre, les variables dans l'ensemble de données sont de grande dimension et subissent des erreurs de mesure. Pour aborder ces défis, nous explorons une méthode d'inférence valide dans le but de traiter les biais de longueur de grande dimension et les données de survie censurées par intervalle accompagnées d'erreurs de mesure dans les covariables

ure time model. We primarily employ the SIMEX method to correct for measurement error effects and propose the boosting procedure to do variable selection and estimation. The proposed method is able to handle the case that the dimension of covariates is larger than the sample size and enjoys appealing features that the distributions of the covariates are left unspecified.

selon un modèle de temps d'échec accéléré. Nous employons principalement la méthode SIMEX pour corriger les effets d'erreur de mesure et proposons la procédure de boosting pour sélectionner et estimer les variables. La méthode proposée est capable de traiter le cas où la dimension des covariables est supérieure à la taille d'échantillon et elle profite de caractéristiques avantageuses faisant en sorte que les distributions des covariables demeurent non spécifiées.

[16:30-16:45]

Weifan Yan (McGill University) **Jiayi Geng** (McGill Univerisity) **Hui Shen** (McGill Univerisity) **W. Evan Johnson** (Rutgers University) **Eric Kolaczyk** (McGill Univerisity)

Estimation of Bipartite Motif Frequencies in Ligand-Receptor Interaction Networks

Estimation des fréquences des motifs bipartis dans les réseaux d'interactions ligand-récepteur

Receptor-ligand (L-R) interactions can be used to infer intercellular communication, which is crucial for many biological functions. The single-cell RNA sequencing (scRNA-seq) technologies decode intercellular communication networks, enabling retrieving L–R interactions and making inferences of these pairings. L-R interactions can be represented by bipartite graphs, which can be characterized by frequencies of motifs, which might be associated with disease presence. However, natural variability in the data propagates to the inference of L-R pairs and, hence, nominal network motif frequencies. In this talk, we offer a method to estimate frequencies of motifs up to six vertices. We use method-of-moments to produce bias-corrected estimates accompanied by uncertainty intervals. To construct confidence intervals, we follow the delta method and use a nonstandard bootstrap method to compute asymptotic variances. We illustrate our method on both synthetic data and real single-cell data.

Les interactions ligand-récepteur peuvent servir à analyser la communication intercellulaire, qui est cruciale pour de nombreuses fonctions biologiques. Les technologies de séquençage de cellule unique de l'ARN décodent les réseaux de communication intercellulaire, ce qui permet d'extraire les interactions ligand-récepteur et d'en faire de l'inférence. Les interactions ligand-récepteur peuvent être représentées par des graphes bipartis, qui peuvent être caractérisés par des fréquences de motifs, lesquels sont susceptibles d'être associés à la présence d'une maladie. Cependant, la variabilité naturelle des données se propage à l'inférence des paires ligand-récepteur et, par conséquent, aux fréquences nominales des motifs du réseau. Dans cette présentation, nous proposons une méthode d'estimation des fréquences de motifs jusqu'à six sommets. Nous utilisons la méthode des moments pour produire des estimations dont le biais est corrigé et qui sont accompagnées d'intervalles d'incertitude. Nous suivons la méthode delta pour créer des intervalles de confiance et nous utilisons une méthode de bootstrap non standard pour calculer les variances asymptotiques. Nous illustrons notre méthode à la fois avec des données synthétiques et des données unicellulaires réelles.

Author List • Liste des auteurs