Société statistique du Canada
Statistical Society of Canada

50th Annual Meeting of the Statistical Society of Canada

50e Congrès annuel de la Société statistique du Canada

May 28 – May 31, 2023
28 mai au 31 mai 2023

Carleton University, Ottawa

Table of Contents • Table des matières

Table of Contents • Table des matières **1**

Welcome to Ottawa and to Carleton University • Bienvenue à Ottawa et à l'Université Carleton **2**

Message from the SSC President • Message de la Presidente de la SSC **4**

Sponsors • Commanditaires **5**

Exhibitors • Exposants **7**

Job Fair • Foire à l'emploi **8**

Organizers • Organisateurs **9**

General Information • Informations générales **12**

The Conference • Le congrès **14**

Social Events • Activités sociales **17**

Committees and Meetings • Comités et réunions **19**

Social Events • Événements sociaux **21**

Workshops • Ateliers **22**

Workshop Descriptions • Descriptifs des ateliers **24**

Scientific Program • Programme scientifique **34**

Abstracts • Résumés **76**

Author List • Liste des auteurs **299**

Welcome to Ottawa and to Carleton University • Bienvenue à Ottawa et à l'Université Carleton

The School of Mathematics and Statistics at Carleton University welcomes you to Ottawa! We are pleased to be hosting the SSC annual meeting for the very first time.

About Ottawa, Canada's national capital

Land Acknowledgement

The Ottawa region is the traditional and unceded territory of the Anishinabeg Algonquin, who have stewarded, travelled and inhabited these lands for millennia. We acknowledge the enduring presence of all First Nations, Métis and Inuit on this land.

Stellar Location

Located between Toronto and Montreal, and just north of the eastern United States, getting to Ottawa is easy. And since the metropolitan region stretches across the Ottawa River to Gatineau, Québec, you can visit two provinces in one destination. Cross one of the bridges, take a boat across the river, or even ride the world's first interprovincial zipline, Interzip Rogers (seasonal)!

As Canada's capital, Ottawa is the perfect place to experience some of the country's best attractions, celebrations and flavours in one welcoming city. With its unique blend of urban and rural beauty, year-round outdoor activities, and vibrant neighbourhoods, the Ottawa region is best explored over three days or more. Check out what our nation's capital region has to offer!

About Carleton University

Situated on unceded Algonquin territory beside the historic Rideau Canal, an official UNESCO World Heritage Site, Carleton University was founded by the community n 1942 to meet the needs of veterans returning from the Second World War.

Carleton University is situated between the Rideau River and the Rideau Canal in Ottawa, the capital city of Canada.

Students, staff, faculty and visitors to Carleton University can take advantage of what Ottawa has to offer all year around: the arts, the outdoors, professional sports, and an exciting, vibrant night life and a strong sense of community. Carleton's proximity to government institutions, libraries, media and a thriving knowledge economy make Carleton a great place to learn and live.

Carleton University is close to Canada's majestic Parliament Buildings, the bustling Byward Market, and the beautiful

L'École de mathématiques et statistique de l'Université Carleton vous souhaite la bienvenue à Ottawa ! Nous sommes heureux d'accueillir le congrès annuel de la SSC pour la première fois.

À propos d'Ottawa, la capitale nationale du Canada

Reconnaissance des terres

La région d'Ottawa est le territoire traditionnel non cédé des Algonquins Anishinabeg, qui ont géré, parcouru et habité ces terres depuis des millénaires. Nous reconnaissons la présence durable sur ces terres de toutes les Premières nations, des Métis et des Inuits.

Une situation exceptionnelle

Située entre Toronto et Montréal, et juste au nord de l'Est des États-Unis, Ottawa est facile d'accès. Et puisque la région métropolitaine s'étend de l'autre côté de la rivière des Outaouais jusqu'à Gatineau (Québec), vous pouvez visiter deux provinces en un seul séjour. Traversez l'un des ponts, prenez un bateau pour traverser la rivière, ou empruntez Interzip Rogers (en saison), la première tyrolienne au monde qui relie deux provinces !

Capitale du Canada, Ottawa est l'endroit idéal où découvrir les meilleures attractions, célébrations et saveurs du pays en une même ville accueillante. Avec son mélange unique de beauté urbaine et rurale, d'activités de plein air en toute saison et de quartiers animés, la région d'Ottawa mérite bien un séjour d'au moins trois jours. Découvrez (vidéo en anglais) ce que la région de la capitale nationale a à offrir !

À propos de l'Université Carleton

Située sur un territoire algonquin non cédé longeant l'historique canal Rideau, consacré patrimoine mondial de l'humanité par l'UNESCO, l'Université Carleton a été fondée par la communauté en 1942 pour répondre aux besoins des anciens combattants revenant de la Seconde Guerre mondiale.

L'Université Carleton est située entre la rivière Rideau et le canal Rideau à Ottawa, capitale du Canada.

Étudiants, personnel, corps enseignant et visiteurs de l'Université Carleton profitent à longueur d'année de tout ce qu'Ottawa a à offrir : arts, plein air, sports professionnels, vie nocturne passionnante et dynamique et un fort sentiment d'appartenance à la communauté. La proximité des institutions gouvernementales, des bibliothèques, des médias et d'une économie du savoir florissante font de Carleton un endroit où il fait bon apprendre et vivre.

L'Université Carleton se trouve à proximité des majestueux

Gatineau Park.

About the School of Mathematics and Statistics

The School of Mathematics and Statistics is home to 15 faculty members in Statistics and Probability, with an additional two cross-appointed faculty coming from Health Science and Neuroscience. The School of Mathematics and Statistics specializes in the following fields:

- Applied Mathematics
- Pure Mathematics
- Probability and Statistics
- Bioinformatics
- Biostatistics

Our Programs

- Bachelor of Mathematics
- Master of Science: Mathematics and Statistics
- Ph.D.: Mathematics and Statistics

Recent new educational programs include the graduate Masters and PhD programs in Data Science, Analytics, and Artificial Intelligence.

For more information to help you plan your trip, please visit:

- SSC 2023 https://ssc.ca/en/meetings/annual/2023-ssc-annual-meeting-ottawa
- Ottawa Tourism / Tourisme Ottawa https://ottawatourism.ca/en/about-ottawa

édifices du Parlement canadien, de l'animation du marché By et du magnifique parc de la Gatineau.

À propos de l'École de mathématiques et statistique

L'École de mathématiques et statistique compte 15 professeurs en Statistique et Probabilité, auxquels s'ajoutent deux professeurs nommés conjointement par les Sciences de la santé et les Neurosciences. L'École de mathématiques et statistique est spécialisée dans les domaines suivants :

- Mathématiques appliquées
- Mathématiques pures
- Probabilité et statistique
- Bioinformatique
- Biostatistique

Nos programmes

- Baccalauréat en mathématiques
- Maîtrise en science : Mathématiques et statistique
- Doctorat : Mathématiques et statistique

Parmi les nouveaux programmes éducatifs récents figurent les programmes de master et de doctorat en Science des données, analyse et intelligence artificielle.

Pour vous aider à préparer votre voyage, veuillez consulter :

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

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, the first in-person meeting since 2019. I believe that everyone is excited and thrilled to finally reconnect with colleagues and friends, 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 85 sessions and more than 300 speakers, in addition to 7 workshops, 2 case studies, with a record number of participants, a job fair, and social events, culminating with the banquet and the awards ceremony. As usual, we will have plenary talk by the Presidential Invited Addressee and many award winners. On behalf on 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.

Bruno N. Rémillard
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, la première qui se tient en personne depuis 2019. Je suis certain que tout le monde est excité et ravi de renouer enfin 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 85 séances et plus de 300 conférenciers, en plus de 7 ateliers, 2 études de cas, avec un nombre record de participants, un salon de l'emploi et des événements sociaux, culminant avec le banquet et la cérémonie de remise des prix. Comme d'habitude, nous aurons une conférence plénière par l'invité du président 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.

Bruno N. Rémillard
Président de la SSC

Sponsors • Commanditaires

The Statistical Society of Canada would like to thank each of the sponsors, whose generous contributions have made this 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

- Carleton University Office of the Vice President (Research and International)

Carleton University
Research and International

- Carleton University Office of the Provost and Vice-President (Academic)

Carleton University
Office of the Provost and Vice-President (Academic)

- Carleton University Faculty of Science

Carleton University
Faculty of Science

- Canadian Statistical Sciences Institute • Institut canadien des sciences statistiques

CANSSI INCASS

Gold Sponsors • Commanditaires d'or

- Bank of Canada / Banque du Canada

BANK OF CANADA
BANQUE DU CANADA

- Springer

Springer

Silver Sponsors • Commanditaires d'argent

- Communications Security Establishment

- National Research Council Canada • Conseil national de recherches Canada

National Research Council Canada
Conseil national de recherches Canada

Bronze Sponsors • Commanditaires de bronze

- Alberta Health Services Provincial Research Data Services

- Calgary Emergency Medicine

Exhibitors • Exposants

- Springer
- Carleton University Faculty of Science
- Canadian Statistical Sciences Institute (CANSSI) • Institut canadien des sciences statistiques (INCASS)

Job Fair • Foire à l'emploi

In order to assist job seekers, we have arranged a Job Fair during SSC2023.

Afin d'aider les demandeurs d'emploi, nous avons organisé une foire de l'emploi pendant SSC2023.

- National Research Council Canada • Conseil national de recherches Canada

National Research Council Canada | Conseil national de recherches Canada

- Communications Security Establishment

Organizers • Organisateurs

Scientific Program Committee • Comité du programme scientifique

- Sévérien Nkurunziza (Chair • Président) University of Windsor
- Fabrice Larribe Université du Québec à Montréal
- Steve Ferreira Guerra McGill University
- Jun Cai University of Waterloo
- Joseph Beyene McMaster University
- Reza Ramezan University of Waterloo
- Tiffany Timbers University of British Columbia
- Shui Feng McMaster University
- Wesley Burr Trent University
- Wilson Lu Acadia University

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

- Shirley Mills (Co-Chair • Co-Présidente)
- Yiqiang Zhao (Co-Chair • Co-Président)
- Ahmed Almaskut
- Song Cai
- Dave Campbell
- Sanjeena Dang
- Utkarsh Dang
- Paul Villeneuve
- Yunran Wei

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:

- AV and IT support from Wade Hong and Julio Aguilar-Hernandez of the Faculty of Science
- Alex Pilkington of CUES
- Richard Cronin of Facilities Management at Carleton
- Linda Cruz and Kendra Reay of Carleton Conference Services
- William Baeta, Patrick Jesse Skeen, Jean Berbari, and Frederic Filliodeau of Aramark
- Phil Genest of Baldwin A/V
- Martin Lessard and Jarrett Bartlett of the Museum of History
- Sarah Kelly, Chantal Mousseau, and Paul Yamashita of Myatering Group
- Jacqueline Boutet of Leduc Bus Lines
- WHOVA representatives Cameron Fenley and Roddy Marshall
- Our many student volunteers who answer questions, provide directions to attendees, set up signs and booths, assist with a/v in rooms, assist at the registration desk, etc.
- At the SSC Office Larysa Valachko, Clayton Forrest and Michelle Benoit have provided crucial support for many aspects, including registration, website and WHOVA app.
- We also are grateful for the assistance from Margaret Tannahill-Wade, (School Administrator), Tracie Grism (Administrative Assistant), Gary Bazdell (Undergraduate Advisor), Vicky Tian (Computer Systems and Network Administrator) and Dr. Robert Burk (Interim Director), School of Mathematics and Statistics.
- SSC Meetings Coordinator Nadia Ghazzali, other SSC executive members and previous local arrangements chairs all shared their experience, offered useful advice, and answered our numerous questions.
- And finally, but most importantly, Angelo Canty managed all electronic services related to the meeting and put together the PDF version of the conference program.

Il est impossible d'organiser un événement de la taille du congrès annuel de la SSC sans l'aide de plusieurs personnes et organisations. Le comité des arrangements locaux tient à remercier toutes celles et tous ceux qui ont contribué à l'organisation de cet événement :

- Wade Hong et Julio Aguilar-Hernandez, de la Faculté des sciences, qui ont apporté leur soutien en matière d'audiovisuel et d'informatique
- Alex Pilkington du CUES
- Richard Cronin, de la gestion des installations à Carleton
- Linda Cruz et Kendra Reay des services de congrès de Carleton
- William Baeta, Patrick Jesse Skeen, Jean Berbari et Frédéric Filliodeau d'Aramark
- Phil Genest de Baldwin A/V
- Martin Lessard et Jarrett Bartlett du Musée canadien de l'histoire
- Sarah Kelly, Chantal Mousseau et Paul Yamashita de My Catering Group
- Jacqueline Boutet de Leduc Bus Lines
- Cameron Fenley et Roddy Marshall, représentants de WHOVA
- Nos nombreux étudiants bénévoles qui répondent aux questions, donnent des indications aux participants, installent les panneaux et kiosques, aident à l'installation des équipements audiovisuels dans les salles, aident au bureau d'inscription, etc.
- Au bureau de la SSC, Larysa Valachko, Clayton Forrest et Michelle Benoit, qui ont apporté un soutien crucial à de nombreux aspects, notamment les inscriptions, le site web et l'application WHOVA.
- Nous remercions également Margaret Tannahill-Wade (administratrice de l'école), Tracie Grism (assistante administrative), Gary Bazdell (conseiller des étudiants de premier cycle), Vicky Tian (administratrice des systèmes informatiques et des réseaux) et Robert Burk (directeur par intérim) de l'École de mathématiques et statistique.
- Nadia Ghazzali, coordinatrice des réunions de la SSC, les autres membres de l'exécutif de la SSC et les responsables des arrangements locaux des congrès précédents, qui ont tous partagé leur expérience, offert d'utiles conseils et répondu à nos nombreuses questions.
- Enfin, mais surtout, Angelo Canty, qui a géré tous les services électroniques liés au congrès et réalisé la version PDF du programme du congrès.

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 SSC2023.

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

How to get to Carleton University • Comment se rendre à l'Université Carleton https://carleton.ca/campus/directions/

Interactive Carleton University campus map • Carte interactive du campus de l'Université Carleton https://carleton.ca/campus/map/

Pdf map of Carleton Campus • Carte en pdf du campus de Carleton https://carleton.ca/campus/wp-content/uploads/map.pdf

Note that some campus buildings have recently been renamed so maps may show either name:

- Residence Commons (RC) is now called Teraanga pronounced (pronounced tare-ang-gah) Commons and houses The Caf.
- University Centre (UC) is now called Nideyinàn (pronounced nih-DAY-in-nahn) and houses Ollie's.
- Robertson Hall (RH) is now called Pigiarvik (pronounced pee-ghee-awe-vik).

Notez que certains édifices du campus ont récemment été renommés et que les cartes peuvent donc indiquer l'un ou l'autre nom :

- Residence Commons (RC) s'appelle désormais Teraanga (prononcé tare-ang-gah) Commons et abrite The Caf.
- Le centre universitaire (UC) s'appelle désormais Nideyinàn (prononcé nih-DAY-in-nahn) et abrite Ollie's.
- Robertson Hall (RH) s'appelle désormais Pigiarvik (prononcé pee-ghee-awe-vik).

Registration • Inscription

Registered participants must pick up their badges at the Registration Desk at the times indicated below. Walk-ins must register online and present their receipt at the registration desk. The Registration Desk is located in Richcraft Hall 2nd floor, top of stairs (or at the elevators) at the entrance to the Atrium. Hours are:

Sunday, May 28: 8:00 a.m.-4:30 p.m.

Monday May 29: 7:45 a.m.-5:00 p.m.

Tuesday May 30: 8:00 a.m. – 5 p.m.

Wednesday May 31: 8:15 a.m. – 4:00 p.m.

Les participants inscrits doivent retirer leur badge au bureau des inscriptions aux heures indiquées ci-dessous. Les personnes qui se présentent sans inscription devront s'inscrire en ligne et présenter leur reçu au bureau des inscriptions. Le bureau des inscriptions est situé à Richcraft Hall, au 2e étage, en haut des escaliers (ou des ascenseurs) à l'entrée de l'atrium. Les heures d'ouverture sont les suivantes :

Dimanche 28 mai : 8h00-16h30

Lundi 29 mai : 7h45-17h00

Mardi 30 mai : 8h00 - 17h00

Mercredi 31 mai : 8h15 - 16h00

Parking and Transportation • Stationnement sur le campus et transports

We strongly urge that you plan to arrive early. On Sunday May 28 there will be road closures due to the Tamarack Ottawa Race Weekend. At any time, Ottawa traffic is congested to please plan accordingly.

The LRT on campus is under construction so it is not available.

There is bus service to the Carleton campus but it will be running on a summer schedule and may be unreliable so allow yourself plenty of time. There is a stop by Richcraft Hall.

You may wish to carpool, take a taxi, Uber or Lyft.

If driving, please note the parking locations marked with **P** on the campus map. The closest one to the conference is the parking garage marked **P** and opposite the Health Science Bldg. (**HS** on the map).

Please check out https://carleton.ca/campus/wp-content/uploads/map.pdf for visitor parking information.

Nous vous conseillons vivement d'arriver tôt. Dimanche 28 mai, il y aura des fermetures de routes en raison de la fin de semaine des courses d'Ottawa Tamarack. En tout temps, la circulation à Ottawa est congestionnée, alors prévoyez vos déplacements en conséquence.

Le TLR sur le campus est en cours de construction et n'est donc pas disponible.

Le campus de Carleton est desservi par un service de bus, mais celui-ci fonctionne selon un horaire d'été et risque d'être peu fiable ; prévoyez donc suffisamment de temps. Il y a un arrêt près de Richcraft Hall.

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

Si vous venez en voiture, veuillez noter les emplacements de stationnement marqués d'un **P** sur le plan du campus. Le plus proche de la conférence est le stationnement marqué **P**, en face du Health Science Bldg (**HS** sur le plan).

Veuillez consulter le site https://carleton.ca/campus/wp-content/uploads/map.pdf pour plus d'informations sur le stationnement visiteurs.

Campus Security • Sécurité sur le campus

- Fire/Police/Ambulance: 911
- Campus Safety Services:
  - EMERGENCIES: 613-520-4444 (ext. 4444 from any campus phone)
  - General Inquiries: 613-520-3612 (ext. 3612 from

- Pompiers/Police/Ambulance : 911
- Services de sécurité du campus :
  - URGENCES : 613-520-4444 (poste 4444 à partir d'un téléphone du campus)
  - Renseignements généraux : 613-520-3612 (poste 3612

any campus phone)
- Inside elevators press the assistance button
- Assistance phones located throughout the campus

à partir d'un téléphone du campus)
- Dans les ascenseurs, appuyez sur le bouton d'assistance
- Téléphones d'assistance situés partout sur le campus

Internet Access • Accès internet

Carleton University is a member of eduroam, 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 Carleton University from eduroam participating institutions can access basic wireless services without having to obtain a Carleton University Guest account. For those visitors who do not come from eduroam participating institutions, you can access a wireless network as follows:

Wi-Fi Network Name: SSC2023

Password: carleton2023

Location: Richcraft, Tory, Azreli and Herzberg

Activation date: May 26th, 2023

Deactivation date: June 1st, 2023

L'Université Carleton est membre d'eduroam, 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 d'établissements participant à eduroam peuvent accéder aux services sans fil de base sans avoir à obtenir un compte d'invité. Pour les visiteurs qui ne viennent pas d'établissements participant à eduroam, vous pouvez accéder à un réseau sans fil de la manière suivante :

Nom du réseau Wi-Fi : SSC2023

Mot de passe : carleton2023

Sites : Richcraft, Tory, Azreli et Herzberg

Date d'activation : 26 mai 2023

Date de désactivation : 1er juin 2023

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

Due to the scarcity of restaurants near the Carleton U campus, lunch is provided daily as part of your registration - with *daily* vouchers @ The Caf in Residence Commons; boxed lunches will be available in RB2211 only for those attending lunchtime committee meetings.

En raison de la rareté des restaurants à proximité du campus de l'Université Carleton, le repas de midi est inclus dans l'inscription - avec des bons *quotidiens* pour le café The Caf aux Residence Commons ; des boîtes à lunch seront disponibles dans la salle RB2211 uniquement pour ceux qui assistent aux réunions des comités à l'heure du repas.

Athletics Facilities • Installations sportives

Carleton University offers state-of-the art athletics facilities. Information on pool swim schedules can be found at https://athletics.carleton.ca/cu-facilities/swimming-pool/; information on our fitness centre is available at https://athletics.carleton.ca/cu-facilities/fitness-centre/.

L'Université Carleton dispose d'installations sportives de pointe. Des informations sur les horaires de natation en piscine sont disponibles à l'adresse https://athletics.carleton.ca/cu-facilities/swimming-pool/ ; vous trouverez plus de renseignements sur notre centre de fitness à https://athletics.carleton.ca/cu-facilities/fitness-centre/.

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 have used 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 présentation 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

Rooms • Salles

All SSC2023 talks are planned to be in Richcraft Hall (RB since it formerly was named River Bldg.), Tory Building (TB) and Azrieli Theatres (AZ). Note that all buildings are connected by a tunnel system which you may wish to use if inclement weather.

Les sessions se tiendront à Richcraft Hall (RB, anciennement River Building), Tory Building (TB) et Azrieli Theatres (AT). Notez que tous les édifices sont reliés par un système de tunnel 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 Richcraft Hall RB2220-2224 with Case Study posters being displayed Monday May 29 afternoon and contributed research posters appear on Tuesday May 30 afternoon. In both cases, posters will be displayed between noon and 4:00 p.m., the authors being with their posters from 1:30 p.m. until 3:00 p.m.

Les affiches des études de cas et libres seront exposées à Richcraft Hall RB2220-2224 (études de cas lundi 29 mai en après-midi et recherches libres mardi 30 mai en après-midi). Dans les deux cas, les affiches seront exposées entre midi et 16h00, les auteurs étant présents de 13h30 à 15h00.

Workshops • Ateliers

Workshops organized by the Sections will be held on Sunday May 28 in rooms in Richcraft Hall. Refreshment breaks will take place in Richcraft Hall Atrium and Patio at 10:00-10:20 a.m. and 2:30-2:50 p.m. . Lunch for Workshops will take place in The Caf in Residence Commons between noon and 1:00 p.m.

Les ateliers organisés par les groupes se tiendront dimanche 28 mai dans les salles de Richcraft Hall. Des pauses collation auront lieu dans l'atrium et le patio de Richcraft Hall à 10h00-10h20 et à 14h30-14h50. Le repas de midi pour les ateliers aura lieu au Caf, aux Residence Commons, entre midi et 13 heures.

Other Meetings • Autres réunions

NSERC Discovery Grant Application Assistance: Monday May 29 10:20-11:50 a.m. in room TB 202
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.

NSERC Updates and 2022 Competition Results: Tuesday May 30 5:00-6:00 p.m. in room RB 2200

Atelier d'assistance à la demande de subvention à la découverte du CRSNG : lundi 29 mai 10h20 - 11h50, salle TB 202
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.

CRSNG – Mise à jour et résultats du concours de 2022 : mardi 30 mai 17h - 18h, salle Science RB 2200.

Social Events • Activités sociales

Welcome Reception • Réception de bienvenue

Sunday, May 28, 5:00 - 7:00 pm • Dimanche 28 mai, 17h00 - 19h00

The Welcome Reception will be held in Richcraft Hall Atrium and Patio, 2nd level of RB at the top of the stairs (also at the elevators on the 2nd floor). 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 two minute walk from the workshop meeting rooms.

La réception de bienvenue se tiendra dans l'atrium et le patio de Richcraft Hall, au 2e étage, en haut des escaliers (ascenseurs). Tous les participants au congrès 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 trousse d'inscription. Pour ceux qui arrivent directement des ateliers, la réception se trouve à deux minutes à pied des salles de réunion.

Student Barbeque • Barbecue des étudiants

Monday, May 29 5:00 - 7:00 pm • Lundi 29 mai, 17h00 - 19h00 ..... Quad

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. In case of bad weather the barbeque will take place in the Tory Building

Le BBQ étudiant est gratuit pour les étudiants de premier et deuxième cycles ; une inscription préalable est nécessaire. Les personnes inscrites au BBQ recevront un billet avec leur trousse d'inscription. En cas de mauvais temps, le barbecue aura lieu dans le Tory Building

Banquet

Tuesday, May 30, 6:30pm – 10:30pm • Mardi 30 mai, 18h30 – 22h30 . . . Museum of History Grand Hall & Haida Gwai Salon

The banquet will be held at the Museum of History, 100 Laurier Street, Gatineau, Quebec with a cash bar/cocktails at 6:30 p.m. and dinner service beginning promptly at 7:30 p.m. 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.

If you have a banquet ticket but do not wish to attend the banquet, please return it to the Registration Desk by noon on Tuesday, May 30 noon - this will help us to better plan for attendance at the banquet.

Shuttle buses will transport conference attendees between Richcraft Hall at Carleton University and the Museum of History in Gatineau both at the start and end of the banquet. Buses will be available as of 6 p.m. in front of Richcraft Hall; return buses will be available as of 10:30 p.m. in front of the Museum of History.

Le banquet aura lieu au Musée canadien de l'histoire, 100 rue Laurier, Gatineau, Québec, avec un bar/cocktail à 18h30, le service du souper commençant à 19h30. Tous les participants au congrès qui ont choisi un repas pour le banquet trouveront un ticket de banquet avec le repas choisi dans leur trousse d'inscription. Vous devez placer le ticket de banquet sur la table devant vous lorsque votre table est servie. Veuillez informer votre serveur de toute allergie alimentaire.

Si vous avez un billet pour le banquet mais que vous ne souhaitez pas y participer, veuillez le retourner au bureau des inscriptions avant mardi 30 mai à midi - cela nous aidera à mieux planifier la participation au banquet.

Des navettes transporteront les participants au congrès entre Richcraft Hall sur le campus et le Musée canadien de l'histoire de Gatineau, au début et à la fin du banquet. Les bus seront disponibles à partir de 18 heures devant Richcraft Hall ; les bus de retour seront disponibles à partir de 22h30 devant le Musée.

Other Social Events • Autres évènements sociaux

Accreditation AGM and Social Gathering
Monday May 29th, 5:30pm – 7:30pm, at Ollie's on the main floor of University Centre, Carleton University.

New Investigators Social Gathering
Monday May 29th, 6pm – 8pm, at Ollie's on the main floor of University Centre, Carleton University.

AGA et réunion sociale de l'accréditation
Lundi 29 mai 17h30 – 19h30, chez Ollie's, au rez-de-chaussée du University Centre, à l'Université Carleton.

Rencontre sociale des nouveaux chercheurs
Lundi 29 mai 18h00 – 20h00, chez Ollie's, au rez-de-chaussée du University Centre, à l'Université Carleton.

Committees and Meetings • Comités et réunions

Sunday May 28 — dimanche 28 mai

09:00-10:30 — **RB 3220**
SSC Executive Committee meeting (tentative)/Réunion du comité éxécutif de la SSC (provisoire)

09:00-16:00 — **RB 3228**
CANSSI CRT meeting/Réunion INCASS PRC

11:00-17:00 — **RB 3220**
SSC Board Meeting/Conseil d'administration de la SSC

Monday May 29 — lundi 29 mai

12:00-13:30 — **TB 238**
Student Travel Grants Committee Meeting/Réunion du comité des subventions allouées aux étudiants

12:00-13:30 — **RB 1201**
Publications Committee Meeting/Réunion du comité des publications

12:00-13:30 — **TB 202**
Executive Committee BISS Meeting/Réunion du comité éxécutif du Groupe de statistique industrielle et de gestion

12:00-13:30 — **RB 1200**
Executive Committee Statistical Education Meeting/Réunion du comité éxécutif du Groupe d'éducation en statistique

12:00-13:30 — **TB 236**
Executive Committee Biostatistics Meeting/Réunion du comité éxécutif du Groupe de biostatistique

Tuesday May 30 — mardi 30 mai

12:00-13:30 — **TB 240**
Women in Statistics Committee Meeting/Réunion du comité des femmes en statistique

12:00-13:30 — **TB 236**
Student and Recent Graduate Committee Meeting/Réunion du comité des étudiants et diplômés récents

12:00-13:30 — **TB 202**
Research Committee Meeting/Réunion du comité de la recherche

12:00-13:30 — **RB 1201**
Office Committee Meeting/Réunion du comité du bureau

12:00-13:30 **RB 1200**

Accreditation Committee meeting/Réunion du comité d'accréditation

Wednesday May 31 mercredi 31 mai

12:00-13:30 **TB 202**

CANSSI CRT meeting/Réunion INCASS PRC

12:00-13:30 **RB 1201**

Public Relations Committee Meeting/Réunion du comité des relations publiques

12:00-13:30 **TB 236**

Committee on Equity, Diversity, and Inclusion (EDI) Meeting/Réunion du comité pour l'équité, la diversité et l'inclusion (EDI)

12:00-13:30 **RB 1200**

CJS Editors Meeting/Réunion des rédacteurs de La revue canadienne de statistique

17:00-18:00 **RB 1200**

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

Social Events • Événements sociaux

Sunday May 28 — dimanche 28 mai

17:00-19:00 — **RB Atrium & Patio**

Welcoming Reception/Réception de bienvenue

Monday May 29 — lundi 29 mai

13:30-16:30 — **RB 2228**

Job Fair: Day 1/Foire de recrutement: 1r jour

17:00-19:00 — **Quad**

Student BBQ/Barbecue d'étudiants

17:30-19:30 — **Ollie's Pub & Patio**

Accreditation AGM and social event/Assemblée générale annuelle du comité d'accréditation et l'activité sociale

---

18:00-20:00 — **Ollie's Pub & Patio**

New Investigators' Social Event/Activité sociale des Nouveaux chercheurs

Tuesday May 30 — mardi 30 mai

13:30-16:30 — **RB 2228**

Job Fair: Day 2/Foire de recrutement: 2e jour

18:30-22:30 — **Museum of History Grand Hall & Haida Gwai Salon**

Banquet/Banquet

Wednesday May 31 — mercredi 31 mai

13:30-16:30 — **RB 2228**

Job Fair: Day 3/Foire de recrutement: 3e jour

Workshops • Ateliers

Sunday May 28 — dimanche 28 mai

09:00-12:00 Workshop / Atelier **RB 1201**

Biostatistics Workshop
Atelier biostatistique

09:00-12:00 **Lehana Thabane** (McMaster University) **Kevin Thorpe** (University of Toronto)
Design, analysis, conduct, and reporting of Pragmatic Trials including Controversies and Confusions about them / Conception, analyse, conduite et rapport des essais pragmatiques, y compris les controverses et les confusions à leur sujet

---

09:00-16:30 Workshop / Atelier **RB 3224**

Actuarial Science Workshop
Atelier du groupe de science actuarielle

09:00-16:30 **Emiliano Valdez** (University of Connecticut) **Xiaofei Shi** (University of Toronto) **Arthur Charpentier** (Université du Québec à Montréal -UQAM) **Ben Feng** (University of Waterloo)
Machine Learning and Data Mining in Insurance and Finance / Apprentissage automatique et exploration de données dans l'assurance et la finance

---

09:00-16:30 Workshop / Atelier **RB 2308**

BISS Workshop
Atelier du GSIG

09:00-16:30 **Jason Hou-Liu** (University of Waterloo)
High Performance Computing with R / Calculs de haute performance avec R

09:00-16:30 Workshop / Atelier **RB 1200**

Statistics Education Workshop
Atelier d'éducation en statistique

09:00-16:30 **Omidali Aghababaei Jazi** (University of Toronto Mississauga) **Katherine Daignault** (University of Toronto) **Chelsea Chelsea Uggenti** (University of Waterloo)
Developing Students' Communication Skills in Statistics / Développer les compétences de communication en statistique des étudiants

---

09:00-16:30 Workshop / Atelier **RB 2311**

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

09:00-16:30 **Varada Kolhatkar** (University of British Columbia) **Joel Ostblom** (University of British Columbia)
Introduction to Machine Learning in Python with scikit-learn / Introduction à l'apprentissage automatique en Python avec scikit-learn

---

09:00-16:30 Workshop / Atelier **RB 3202**

Survey Methods Workshop
Atelier du Groupe des méthodes d'enquêtes

09:00-16:30 **Jae-Kwang Kim** (Iowa State University)
Statistical Methods for handling missing data / Méthodes statistiques pour le traitement des données manquantes Ⓔ E

09:00-14:50 Workshop / Atelier **RB 3201**

Probability Workshop
Atelier du probabilité

09:00-14:50 **Anatoliy Swishchuk** (University of Calgary)
Stochastic Modelling of Financial and Energy Markets Data / Modélisation stochastique des données des marchés financiers et énergétiques Ⓔ E

Workshop Descriptions • Descriptifs des ateliers

Biostatistics Workshop • Atelier biostatistique

Room/Salle: RB 1201

Date: Sunday May 28 / dimanche 28 mai

Time/Heure: 09:00-12:00

Lehana Thabane (McMaster University) **Kevin Thorpe** (University of Toronto)

Design, analysis, conduct, and reporting of Pragmatic Trials including Controversies and Confusions about them

Conception, analyse, conduite et rapport des essais pragmatiques, y compris les controverses et les confusions à leur sujet

The term "pragmatic attitude," was introduced by Daniel Schwartz and Joseph Lellouch [J Chron Dis 1967;20:637-48] to designate a trial architecture designed to answer the practical question of whether offering an intervention in routine healthcare setting does more good than harm. Pragmatic trials have been receiving increasing attention as patients, clinicians, organizers and payers for health services are increasingly searching for answers (e.g., the burgeoning programs of "Comparative Effectiveness" research). However, this increasing attention has exposed considerable confusion and disagreement about whether pragmatic trials are "less perfect experiments" than efficacy [explanatory] trials, whether permitting non-compliance and crossovers is a "problem," whether knowing one's treatment and telling examiners how one feels creates "bias," and whether other departures from the explanatory architecture are good use of limited resources to answer burning questions about the benefits and harms of interventions. This 3-hour course will cover the following objectives:

1. To describe what a pragmatic trial is, and how it is different from an explanatory trial;
2. To know the pros and cons of pragmatic trials;
3. To be familiar with the PRECIS tool for the design of pragmatic trials;
4. To learn how to analyze and report pragmatic trials using the CONSORT extension to pragmatic trials; and
5. To understand the common controversies and criticisms of pragmatic trials, and how to address them in design and analysis.

Schedule:
Lehana Thabane (McMaster University) (9:00-10:20)
Coffee break: (10:20-10:40)
Kevin Thorpe (University of Toronto) (10:40-12:00)
lunch break: (12:00-13:30)

Le terme « attitude pragmatique » a été introduit par Daniel Schwartz et Joseph Lellouch [J Chron Dis 1967 ;20 :637-48] pour désigner une architecture d'essai conçue pour répondre à la question pratique de savoir si l'offre d'une intervention dans un contexte de soins de santé de routine fait plus de bien que de mal. Les essais pragmatiques font l'objet d'une attention croissante car les patients, les cliniciens, les organisateurs et les payeurs de services de santé sont de plus en plus à la recherche de réponses (par exemple, programmes florissants de recherche sur « l'efficacité comparative »). Cependant, cette attention croissante a mis en évidence une confusion et un désaccord considérables sur la question de savoir si les essais pragmatiques sont des « expériences moins parfaites » que les essais d'efficacité [explicatifs], si l'autorisation de la non-conformité et des croisements est un « problème », si le fait de connaître son traitement et de dire aux examinateurs comment on se sent crée un « biais », et si d'autres écarts par rapport à l'architecture explicative constituent une bonne utilisation de ressources limitées pour répondre aux questions brûlantes sur les avantages et les inconvénients des interventions. Ce cours de 3 heures couvrira les objectifs suivants :

1. Décrire ce qu'est un essai pragmatique, et en quoi il diffère d'un essai explicatif ;
2. Connaître les avantages et les inconvénients des essais pragmatiques ;
3. Se familiariser avec l'outil PRECIS pour la conception d'essais pragmatiques ;
4. Apprendre à analyser et à rendre compte des essais pragmatiques en utilisant l'extension CONSORT aux essais pragmatiques ; et
5. Comprendre les controverses et les critiques courantes des essais pragmatiques, et comment y répondre dans la conception et l'analyse.

Horraire :
Lehana Thabane (Université de McMaster) (09h00-10h20)
Pause café : (10h20-10h40)
Kevin Thorpe (Université de Toronto)(10h40-12h00)
Pause déjeuner : (12h00-13h30).

Actuarial Science Workshop • Atelier du groupe de science actuarielle

Room/Salle: RB 3224

Date: Sunday May 28 / dimanche 28 mai

Time/Heure: 09:00-16:30

Emiliano Valdez (University of Connecticut) **Xiaofei Shi** (University of Toronto) **Arthur Charpentier** (Université du Québec à Montréal -UQAM) **Ben Feng** (University of Waterloo)

Machine Learning and Data Mining in Insurance and Finance

Apprentissage automatique et exploration de données dans l'assurance et la finance

Machine learning and data mining are hot research topics in insurance and finance and are heavily used in the industry of insurance and finance as well. In this workshop, the four invited speakers will introduce methods of machine learning and data mining and discuss their applications in insurance and finance. Each of the four speakers has 80 minutes for the presentation. There will be a twenty-minute break after each presentation and a 90 minute lunch break. The total duration of the workshop is seven and half hours.

Note:
Session 1: (9:00-10:20)
Coffee break: (10:20-10:40)
Session 2: (10:40-12:00)
Lunch break: (12:00-13:30)
Session 3: (13:30-14:50)
Coffee break: (14:50-15:10).
Session 4: (15:10-16:30).

L'apprentissage automatique et l'exploration de données sont des sujets de recherche brûlants dans le domaine de l'assurance et de la finance ; ils sont également très utilisés dans ce secteur. Dans cet atelier, les quatre orateurs invités présenteront des méthodes d'apprentissage automatique et d'exploration de données et discuteront de leurs applications au domaine de l'assurance et de la finance. Chacun des quatre orateurs dispose de 80 minutes pour sa présentation. Une pause de vingt minutes est prévue après chaque présentation et il y aura une pause déjeuner de 90 minutes. La durée totale de l'atelier est de sept heures et demie.

Note :
Session 1 : (09h00-10h20)
Pause café : (10h20-10h40)
Session 2 : (10h40-12h00)
Pause déjeuner : (12h00-13h30)
Session 3 : (13h30-14h50)
Pause café : (14h50-15h10)
Session 4 : (15h10-15h50).

BISS Workshop • Atelier du GSIG

Room/Salle: RB 2308

Date: Sunday May 28 / dimanche 28 mai

Time/Heure: 09:00-16:30

Jason Hou-Liu (University of Waterloo)

High Performance Computing with R

Calculs de haute performance avec R

Complicated models, bigger simulation studies, and larger datasets mean that R code can take a long time to run. This workshop will cover speeding up R code and scripts by using parallelism on high performance compute systems and improving runtime efficiency by leveraging Rcpp and other expeditious C++ based packages. No prior C++ experience is required! An understanding of the concepts behind parallel computing and interoperating with compiled code, along with hands-on practice will be the mainstay of the workshop.

We will cover writing parallelized code for local and remote machines, along with tips and tricks to make the most out of your compute time. Key topics include parallelizing using multithreaded linear algebra libraries, parallelizing within R itself, efficient memory usage on large core-count machines, with a light dusting of hardware intuition. We will emphasize re-using existing code and scripts by making minimal modifications to allow for parallelism. As well, we will learn to profile long-running code to identify hot loops that are candidates for acceleration using compiled code. Packages such as Rfast, Rcpp, and RcppArmadillo will be used to improve the compute efficiency of repeatedly executed code.

To smooth out the setup process, please have RStudio and R installed, along with Rtools (on Windows) or Xcode (on Mac) or gcc (on Linux). As well, we will be relying on the Rcpp, RcppArmadillo, Rfast, Rfast2, pbapply, profvis, and bench packages; please make sure these packages are installed as well. If you don't already have access to a cluster, Compute Canada Database (CCDB) offers free access to a wide range of academic and non-academic researchers and students. Please see https://ccdb.computecanada.ca/security/login for more information and to apply for an account. It may take a couple of weeks to get approved and activate an account, so please create your account prior to the workshop if you are interested. Nevertheless, most examples provided will still run locally on a personal laptop for instructional purposes.

Les modèles compliqués, les vastes études de simulation et les ensembles de données volumineux entraînent des temps d'exécution de plus en plus longs sous R. Cet atelier expliquera comment accélérer le code et les scripts R en utilisant le parallélisme sur les systèmes de calcul à haute performance et en améliorant l'efficacité de l'exécution en tirant parti de Rcpp et d'autres paquets rapides basés sur C++. Aucune expérience préalable en C++ n'est requise ! Une compréhension des concepts sous-jacents au calcul parallèle et l'interopérabilité avec le code compilé, ainsi que la pratique seront les piliers de l'atelier.

Nous aborderons l'écriture de code parallélisé pour les machines locales et distantes, ainsi que des conseils et astuces pour tirer le meilleur parti de votre temps de calcul. Parmi les sujets clés : la parallélisation à l'aide de bibliothèques d'algèbre linéaire multithread, la parallélisation au sein même de R, l'utilisation efficace de la mémoire sur les machines à grand nombre de noyaux, et un peu d'intuition matérielle. Nous verrons comment réutiliser un code et des scripts existants en les modifiant légèrement pour permettre le parallélisme. De plus, nous apprendrons à profiler un code long à exécuter pour identifier les boucles pouvant être accélérées à l'aide de code compilé. Nous utiliserons des paquets tels que Rfast, Rcpp, et RcppArmadillo pour améliorer l'efficacité de calcul du code exécuté de façon répétée.

Pour faciliter le processus, veuillez installer RStudio et R, ainsi que Rtools (sous Windows) ou Xcode (sous Mac) ou gcc (sous Linux). Nous utiliserons également Rcpp, RcppArmadillo, Rfast, Rfast2, pbapply, profvis, et bench ; veuillez vous assurer de préinstaller ces paquets. Si vous n'avez pas encore accès à une grappe, la base de données Calcul Canada (CCDB) est en accès gratuit pour les chercheurs et étudiants, universitaires ou non. Veuillez consulter le site https://ccdb.computecanada.ca/security/login pour plus d'informations et pour demander un compte. L'approbation et l'activation d'un compte pouvant prendre quelques semaines, nous vous invitons à créer votre compte avant l'atelier si vous êtes intéressé. Néanmoins, la plupart des exemples fournis fonctionneront également sur un ordinateur portable personnel, à des fins pédagogiques.

Note :
Session 1 : (09h00-10h20)
Pause café : (10h20-10h40)

Note:
Session 1: (9:00-10:20)
Coffee break: (10:20-10:40)
Session 2: (10:40-12:00)
Lunch break: (12:00-13:30)
Session 3: (13:30-14:50)
Coffee break: (14:50-15:10).
Session 4: (15:10-16:30).

Session 2 : (10h40-12h00)
Pause déjeuner : (12h00-13h30)
Session 3 : (13h30-14h50)
Pause café : (14h50-15h10)
Session 4 : (15h10-16h30).

Statistics Education Workshop • Atelier d'éducation en statistique

Room/Salle: RB 1200

Date: Sunday May 28 / dimanche 28 mai

Time/Heure: 09:00-16:30

Omidali Aghababaei Jazi (University of Toronto Mississauga) **Katherine Daignault** (University of Toronto) **Chelsea Uggenti** (University of Waterloo)

Developing Students' Communication Skills in Statistics

Développer les compétences de communication en statistique des étudiants

It is critical to equip students with professional skills. In particular, they need to develop effective verbal and written communication skills and it is our responsibility to help them in the process. As statistics educators, we strive to include meaningful relevant course activities that provide our students with opportunities to develop reasoning skills and a deep understanding of concepts, we tend to face challenges teaching students how to appropriately verbalize and write about statistics. This workshop targets statistics educators who plan to support their students' learning about how to communicate the results of a scientific investigation of data, such as how to describe data succinctly, create effective visualizations, and accurately summarize statistical findings. In this workshop, we facilitate activities drawn from best practices in statistics education, utilizing the work by Deborah Nolan and Sara Stoudt in *Communicating with Data: The Art of Writing for Data Science* (2021) as well as sharing our own practical ideas for supporting students' development of their communications skills in statistics.

Omidali Aghababaei Jazi, PhD. University of Toronto Mississauga (UTM), Department of Mathematical and Computational Sciences

Bio: Omidali Aghababaei Jazi is an assistant professor, teaching stream, in the Department of Mathematical and Computational Sciences at the University of Toronto Mississauga (UTM). He has taught a variety of courses including Probability and Statistics, Stochastic Processes, and Experimental Design, supervised undergraduate students for their research projects, and facilitated SSC sessions and the case study competition over the last years. His research interests are statistics education, analysis of biased survival data, and analysis of longitudinal data with informative follow-up.

Contribution: Omid will help organize and plan for workshop activities with the help of other organizers and will facilitate activities from Part V of the book Communicating with Data: The Art of Writing for Data Science, concerning Science Writing and You.

Il est essentiel de doter les étudiants de compétences professionnelles. En particulier, ils doivent développer des compétences efficaces en matière de communication verbale et écrite et il est de notre responsabilité de les aider dans ce processus. En tant qu'enseignants en statistique, nous nous efforçons d'inclure des activités de cours pertinentes et significatives qui donnent à nos étudiants l'occasion de développer des capacités de raisonnement et une compréhension approfondie des concepts. Pourtant, nous rencontrons parfois des difficultés pour enseigner aux étudiants comment verbaliser et écrire de manière appropriée sur les statistiques. Cet atelier s'adresse aux enseignants en statistique qui prévoient de soutenir l'apprentissage de leurs étudiants sur la façon de communiquer les résultats d'une enquête scientifique sur les données, par exemple comment décrire les données de façon succincte, créer des visualisations efficaces et résumer avec précision les résultats statistiques. Dans cet atelier, nous animons des activités tirées des meilleures pratiques de l'enseignement de la statistique. Nous nous appuierons sur le travail de Deborah Nolan et Sara Stoudt dans *Communicating with Data : The Art of Writing for Data Science* (2021) et nous partagerons nos propres idées pratiques pour soutenir le développement des compétences de communication en statistique des étudiants.

Omidali Aghababaei Jazi, PhD. Université de Toronto Mississauga (UTM), Département des sciences mathématiques et informatiques

Bio : Omidali Aghababaei Jazi est professeur adjoint au Département des sciences mathématiques et informatiques de l'Université de Toronto Mississauga (UTM). Il a enseigné divers cours, notamment Probabilité et statistique, Processus stochastiques et Conception expérimentale, supervisé des étudiants de premier cycle dans le cadre de leurs projets de recherche et animé des sessions de la SSC et le concours d'études de cas ces dernières années. Il étudie l'enseignement de la statistique, l'analyse de données de survie biaisées et l'analyse de données longitudinales avec suivi informatif.

Contribution : Omid aidera à organiser et à planifier les activités de l'atelier avec l'aide d'autres organisateurs et animera les activités de la partie V du livre Communicating with Data : The Art of Writing for Data Science, concernant L'écriture scientifique et vous.

Katherine Daignault, PhD. University of Toronto, Department of Statistical Sciences
Bio: Katherine Daignault is an assistant professor, teaching stream, in the Department of Statistical Sciences at the University of Toronto. She has taught a variety of statistics courses, ranging from survey courses for non-statistics students to core 3rd/4th year methodology courses for statistics students. Her current teaching focus is on incorporating active learning into her courses, researching effective ways to create hybrid courses, and creating authentic assessments that encourage students to practice written and oral communication.
Contribution: Katherine will help organize and plan workshop activities with the help of other organizers, and will facilitate activities related to Part II and III of the book Communicating with Data: The Art of Writing for Data Science, concerning writing for various audiences, communication through figures, and organizing the story of written work.
Chelsea Uggenti, PhD. University of Waterloo, Department of Statistics and Actuarial Science
Bio: Chelsea Uggenti is an applied statistician with research interests in fire science, statistical education, active learning, and the development of graduate teaching assistants. She received her doctorate in statistics at the University of Western Ontario and is a Lecturer in the Department of Statistics and Actuarial Science at the University of Waterloo.
Contribution: Chelsea will help organize and plan for workshop activities with the help of other organizers, and will facilitate activities from Part IV of the book Communicating with Data: The Art of Writing for Data Science, concerning Editing and Revising. She will also cover oral communication in statistics within the field and the broader community.

Note:
Session 1: (9:00-10:20)
Coffee break: (10:20-10:40)
Session 2: (10:40-12:00)
Lunch break: (12:00-13:30)
Session 3: (13:30-14:50)
Coffee break: (14:50-15:10).
Session 4: (15:10-16:30).

Katherine Daignault, PhD. Université de Toronto, Département des sciences statistiques
Bio : Katherine Daignault est professeure adjointe, filière enseignement, au Département des sciences statistiques de l'Université de Toronto. Elle a enseigné une variété de cours de statistique, allant des cours d'enquête pour les étudiants non statisticiens aux cours de méthodologie de base de 3e/4e année pour les étudiants en statistique. Elle se concentre actuellement sur l'intégration de l'apprentissage actif dans ses cours, la recherche de moyens efficaces de créer des cours hybrides et la création d'évaluations authentiques qui encouragent les étudiants à pratiquer la communication écrite et orale.
Contribution : Katherine participera à l'organisation et à la planification des activités de l'atelier avec l'aide d'autres organisateurs, et animera les activités liées aux parties II et III du livre Communicating with Data : The Art of Writing for Data Science, concernant l'écriture pour différents publics, la communication par les chiffres et l'organisation de l'histoire du travail écrit.
Chelsea Uggenti, PhD. Université de Waterloo, Département de statistique et d'actuariat
Bio : Chelsea Uggenti est une statisticienne appliquée qui s'intéresse à la science du feu, à l'enseignement de la statistique, à l'apprentissage actif et au développement d'assistants d'enseignement diplômés. Elle a obtenu son doctorat en statistique à l'Université de Western Ontario et est chargée de cours au Département de statistique et d'actuariat de l'Université de Waterloo.
Contribution : Chelsea aidera à organiser et à planifier les activités de l'atelier avec l'aide d'autres organisateurs, et facilitera les activités de la partie IV du livre Communicating with Data : The Art of Writing for Data Science, concernant l'édition et la révision. Elle couvrira également la communication orale en statistique au sein du domaine et de la communauté au sens large.

Note :
Session 1 : (09h00-10h20)
Pause café : (10h20-10h40)
Session 2 : (10h40-12h00)
Pause déjeuner : (12h00-13h30)
Session 3 : (13h30-14h50)
Pause café : (14h50-15h10)
Session 4 : (15h10-16h30).

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

Room/Salle: RB 2311

Date: Sunday May 28 / dimanche 28 mai

Time/Heure: 09:00-16:30

Varada Kolhatkar (University of British Columbia) **Joel Ostblom** (University of British Columbia)

Introduction to Machine Learning in Python with scikit-learn

Introduction à l'apprentissage automatique en Python avec scikit-learn

In this workshop learners will be introduced to machine learning in the Python programming language using the most popular library, scikit-learn. By the end of the workshop, learners should be able to build a supervised machine learning pipeline in Python using scikit-learn. The workshop outline is shown below below:

- Workshop introduction (1̃5 mins)
- Terminology, baselines, and introduction to scikit-learn fit and predict (4̃5 mins)
- Fundamentals of machine learning (1 hour)
- Hands-on practice (1̃ hour)
- Analogy-based models, linear models (1̃ hour)
- Preprocessing, pipelines, column transformers (1 hour)
- Hands-on practice (1̃ hour)

Note:
Session 1: (9:00-10:20)
Coffee break: (10:20-10:40)
Session 2: (10:40-12:00)
Lunch break: (12:00-13:30)
Session 3: (13:30-14:50)
Coffee break: (14:50-15:10).
Session 4: (15:10-16:30).

Dans cet atelier, les apprenants seront initiés à l'apprentissage automatique dans le langage de programmation Python en utilisant la bibliothèque la plus populaire, scikit-learn. À la fin de l'atelier, ils devraient être en mesure de construire un pipeline d'apprentissage automatique supervisé en Python à l'aide de scikit-learn. Le plan de l'atelier est présenté ci-dessous :

- Introduction de l'atelier (1̃5 mins)
- Terminologie, lignes de base, et introduction à scikit-learn fit et predict (4̃5 mins)
- Principes fondamentaux de l'apprentissage automatique (1 heure)
- Travaux pratiques (1̃ heure)
- Modèles basés sur l'analogie, modèles linéaires (1̃ heure)
- Prétraitement, pipelines, transformateurs de colonnes (1 heure)
- Travaux pratiques (1̃ heure)

Note :
Session 1 : (09h00-10h20)
Pause café : (10h20-10h40)
Session 2 : (10h40-12h00)
Pause déjeuner : (12h00-13h30)
Session 3 : (13h30-14h50)
Pause café : (14h50-15h10)
Session 4 : (15h10-16h30).

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

Room/Salle: RB 3202

Date: Sunday May 28 / dimanche 28 mai

Time/Heure: 09:00-16:30

Jae-Kwang Kim (Iowa State University)

Statistical Methods for handling missing data

Méthodes statistiques pour le traitement des données manquantes

The workshop will focus on the theory and methods for missing data analysis and its application to latent variable models. Topics include maximum likelihood estimation under missing data, mean score theorem, EM algorithm, imputation, fractional imputation, multiple imputation, propensity scores, doubly robust estimation, and non-ignorable missing data.
Outline:
1. Basic Theory for missing data (80 minutes)
2. Imputation (80 minutes)
3. Propensity score approach (80 minutes)
4. Advanced topics (80 minutes)

Note:
Session 1: (9:00-10:20)
Coffee break: (10:20-10:40)
Session 2: (10:40-12:00)
Lunch break: (12:00-13:30)
Session 3: (13:30-14:50)
Coffee break: (14:50-15:10).
Session 4: (15:10-16:30).

L'atelier se concentrera sur la théorie et les méthodes d'analyse des données manquantes et leur application aux modèles de variables latentes. Les sujets abordés comprennent l'estimation du maximum de vraisemblance en cas de données manquantes, le théorème du score moyen, l'algorithme EM, l'imputation, l'imputation fractionnelle, l'imputation multiple, les scores de propension, l'estimation doublement robuste et les données manquantes non ignorables.
Aperçu :
1. Théorie de base pour les données manquantes (80 minutes)
2. Imputation (80 minutes)
3. Approche par score de propension (80 minutes)
4. Sujets avancés (80 minutes)

Note :
Session 1 : (09h00-10h20)
Pause café : (10h20-10h40)
Session 2 : (10h40-12h00)
Pause déjeuner : (12h00-13h30)
Session 3 : (13h30-14h50)
Pause café : (14h50-15h10)
Session 4 : (15h10-16h30).

Probability Workshop • Atelier du probabilité

Room/Salle: RB 3201

Date: Sunday May 28 / dimanche 28 mai

Time/Heure: 09:00-14:50

Anatoliy Swishchuk (University of Calgary)

Stochastic Modelling of Financial and Energy Markets Data

Modélisation stochastique des données des marchés financiers et énergétiques

Lecture 1 (2 hours): Introduction to Mathematical Modelling and Mathematical Finance

• Examples of Modelling: ODE, PDE, Interest Rate Modelling Process
• Modelling Involved Stochastic/Random Behaviour: Wiener and Gaussian Processes
• Stock Prices Modelling (Geometric Brownian Motion) and Forwards Simulation Roots and History of Math Finance
• Classics: Cox-Ross-Rubinstein (CRR) (discrete-time case) and Black-Scholes (continuous-time case) Models and Formulas, Call-Put Parity
• New Directions and Some Prospective in Math Finance

Lecture 2 (2 hours): Stochastic Modelling of Big Data in Financial and Energy Markets

• Big Data in Finance: High-Frequency and Algorithmic Trading (HFT) Data Limit Order Books (LOB)
• Stochastic Modelling of Big Data in Finance
• Forwards, Futures, Options on Forwards and on Futures Black-76 Formula (Positive Prices)
• Alternative to Black-76 Formulas (Negative Prices)
• Some Results for Alternative Models
• A Vision to Transition to 100

PS: 10 Multiple choice questions will be provided to test the students' training activities.

Note:
Session 1: (9:00-10:20)
Coffee break: (10:20-10:40)
Session 2: (10:40-12:00)
Lunch break: (12:00-13:30)
Session 3: (13:30-14:50).

Cours 1 (2 heures) : Introduction à la modélisation mathématique et à la finance mathématique

• Exemples de modélisation : EDO, EDP, processus de modélisation du taux d'intérêt
• Modélisation du comportement stochastique/aléatoire impliqué : Processus de Wiener et gaussiens
• Modélisation des prix des actions (mouvement brownien géométrique) et simulation à terme.
• Racines et histoire des classiques de la finance mathématique : Modèles et formules de Cox-Ross-Rubinstein (CRR) (cas en temps discret) et Black-Scholes (cas en temps continu), parité call-put.
• Orientations et quelques perspectives en finance mathématique

Conférence 2 (2 heures) : Modélisation stochastique de Big Data sur les marchés financiers et énergétiques

• Big Data en finance : Données sur la négociation algorithmique et à haute fréquence (HFT), ordres à cours limités
• Modélisation stochastique de Big Data en finance
• Contrats à terme (forwards, futures), options sur contrats à terme (Forwards et futures), formule Black-76 (prix positifs)
• Alternative aux formules Black-76 (prix négatifs)
• Quelques résultats pour des modèles alternatifs
• Une vision pour passer à 100

PS : 10 questions à choix multiples seront fournies pour tester les activités de formation des étudiants.

Note :
Session 1 : (09h00-10h20)
Pause café : (10h20-10h40)
Session 2 : (10h40-12h00)
Pause déjeuner : (12h00-13h30)
Session 3 : (13h30-14h50).

Scientific Program • Programme scientifique

Monday May 29 lundi 29 mai

08:30-09:50 **Invited / Sur invitation** (abstract/résumé 77) **RB 2200**

SSC Presidential Invited Address
Allocution de l'invité du président de la SSC

Chair/Président: Bruno N. Rémillard
Organizer/Responsable: Bruno N. Rémillard

08:30-09:50 **Pierre Del Moral** (INRIA, Bordeaux Research Center)
Some Theoretical Aspects of Particle Filters and Ensemble Kalman Filters / Quelques aspects théoriques des filtres à particules et des filtres de Kalman d'ensemble

---

10:20-11:50 **Invited / Sur invitation** (abstract/résumé 78) **AT 101**

Advances and applications in optimal dynamic treatment regimes
Progrès et applications dans les régimes de traitement dynamiques optimaux

Chair/Président: Mireille Schnitzer
Organizer/Responsable: Mireille Schnitzer, CONG JIANG

10:20-10:42 **Lu Wang** (University of Michigan)
Evidence-based Estimation of the Optimal Dynamic Personalized Health Care Decision Rules with Restrictions / Estimation fondée sur des données probantes des règles de décision dynamiques optimales en matière de soins de santé personnalisés avec restrictions

10:42-11:05 **Aaron Sarvet** (EPFL)
Optimal Regimes for Algorithm-Assisted Human Decision-Making / Régimes optimaux pour la prise de décision humaine assistée par algorithme

11:05-11:27 **Cong Jiang** (University of Montreal) **Mary Thompson** (University of Waterloo) **Michael Wallace** (University of Waterloo)
Estimating DTR with household interference for ordinal outcome / Estimation de régimes de traitement dynamiques (DTR) avec interférences ménagères pour des réponses ordinales

11:27-11:50 **Janie Coulombe** (Université de Montréal) **Erica E. M. Moodie** (McGill University) **Susan M. Shortreed** (Kaiser Permanente Washington) **Christel Renoux** (Lady Davis Institute for Clinical Research)
Using multiple imputations of longitudinal data and dynamic weighted survival modeling to choose an optimal antidepressant drug / Le recours à l'imputation multiple de données longitudinales et à la modélisation de survie dynamique afin d'identifier un médicament antidépresseur optimal

---

10:20-11:50 **Invited / Sur invitation** (abstract/résumé 81) **AT 301**

Maximizing learning capacity in a joyful environment in Statistical Education
Maximiser la capacité d'apprentissage dans un environnement joyeux dans l'éducation en statistique

Chair/Président: Suborna Shekhor Ahmed
Organizer/Responsable: Suborna Shekhor Ahmed
Sponsor/Commanditaires: Statistics Education Section/Statistics Education Section

10:20-10:50 **Bruce Dunham** (The University of British Columbia)
Implementing evidence-based pedagogy: When they don't know what's good for them / Implantation d'une pédagogie basée sur la preuve : lorsqu'ils ne savent pas ce qui est bon pour eux

10:50-11:20 **Douglas Whitaker** (Mount Saint Vincent University) **Alana Unfried** (California State University, Monterey Bay) **Leyla Batakci** (Elizabethtown College) **Marjorie E. Bond** (Pennsylvania State University) **April Kerby-Helm** (Winona State University) **Michael A. Posner** (Villanova University)
Measuring student enjoyment and effective teaching practices: what works? / Mesure de l'enthousiasme des étudiants et pratiques d'enseignement efficaces : qu'est-ce qui fonctionne ?

11:20-11:50 **Mike Marin** (The University of British Columbia)
Creating Educational Videos: Advice From A Decade of Experience / Création de vidéo éducatif : conseil basé sur dix ans d'expérience

---

10:20-11:50 **Invited / Sur invitation** (abstract/résumé 83) **TB 202**

Stochastic Spatial Models For Vegetation
Stochastic Spatial Models For Vegetation

Chair/Président: Priscilla E. Greenwood
Organizer/Responsable: Priscilla E. Greenwood
Sponsor/Commanditaires: Probability Section/Probability Section

10:20-10:50 **Priscilla E. Greenwood** (The University of British Columbia)
Desertification: Stochastic Models and Statistical Challenges / Désertification : modèles stochastiques et problèmes statistiques

10:50-11:20 **Luis F. Gordillo** (Utah State University) **Priscilla E. Greenwood** (University of British Columbia)
Intermittent Precipitation-Dependent Interactions, Encompassing Allee Effect, May Yield Vegetation Patterns in a Transitional Parameter Range / Des interactions intermittentes dépendant des précipitations, englobant l'effet Allee, peuvent donner lieu à des schémas de végétation dans une plage de paramètres transitoire

11:20-11:50 **Chunyi Gai** (The University of British Columbia) **Yuxin Chen** **Theodore Kolokolnikov** (Dalhousie University) **Justin Tzou** (Macquarie University)
Patterned Vegetation, Tipping Points and the Rate of Climate Change / Végétation schématique, point de bascule et vitesse du changement climatique

---

10:20-11:50 **Invited / Sur invitation** (abstract/résumé 85) **AT 102**

Data Science in Industry: A Showcase
Science des données dans l'industrie : vitrine

Chair/Président: Matthew Greenberg
Organizer/Responsable: Matthew Greenberg
Sponsor/Commanditaires: Data Science and Analytics Section/Data Science and Analytics Section

10:20-10:50 **Jennifer Zhu** (Amazon)
Data Science in Industry: A Showcase / Présentation de la science des données dans l'industrie

10:50-11:20 **Tingzhou Yu** (University of Victoria)
Data Science and Optimization: An Industry Perspective / Point de vue de l'industrie sur la science et l'optimisation des données

---

10:20-11:50 **Invited / Sur invitation** (abstract/résumé 87) **TB 208**

Stochastic Optimization in Insurance and Finance
Optimisation stochastique en assurance et finance

Chair/Président: Jean-François Renaud
Organizer/Responsable: Jean-François Renaud
Sponsor/Commanditaires: Actuarial Science Section/Actuarial Science Section

10:20-10:50 **Bin Li** (University of Waterloo)
Equilibrium Pricing of Variable Annuities within a Principal-Agent Framework / Tarification d'équilibre des rentes variables dans un cadre mandant-mandataire Ⓔ E

10:50-11:20 **Alexandre Roch** (UQAM) **Simone Scotti** (Université Paris Diderot)
Optimal Ratcheting Dividends Policy with Resets / Politique optimale de dividendes de type cliquet avec réinitialisation Ⓔ E

11:20-11:50 **Abel Cadenillas** (University of Alberta) **Wenyue Liu** (University of Alberta)
Optimal Insurance Contracts for a Shot-Noise Cox Claim Process / Contrats d'assurance optimaux pour un processus de réclamations Cox à bruit quantique Ⓔ E

10:20-11:50 **Contributed / Communications libres** (abstract/résumé 89) **TB 236**

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

Chair/Président: Hugh Chipman

10:20-10:35 **Larry Dong** (University of Toronto: Dalla Lana School of Public Health) **David W. Johnson** (Department of Pediatrics, Faculty of Medicine, University of Calgary, Alberta Children's Hospital, Calgary, Canada) **Terry P. Klassen** (Manitoba Institute of Child Health, Winnipeg, Canada) **Anna Heath** (Child Evaluative Services, the Hospital for Sick Children, Toronto, Canada; Division of Biostatistics, Dalla Lana School of Public Health, University of Toronto, Toronto, Canada; Department of Statistical Science, University College London, London, UK) **Amy Plint** (CHEO Research Institute, Ottawa, Canada; Department of Emergency Medicine, University of Ottawa, Ottawa, Canada)
The Probability of Reducing Hospitalisation Rates for Bronchiolitis: A Bayesian Analysis / La probabilité de réduire les taux d'hospitalisation pour la bronchiolite : une analyse bayésienne Ⓔ E

10:35-10:50 **Fatemeh Mahmoudi** (University of Calgary) **Xuewen Lu** (University of Calgary)
Bi-Level Selection with the Sparse Group BAR for Multivariate Interval-Censored Data / Sélection à deux niveaux avec la régression ridge adaptative brisée (BAR) par groupe parcimonieux pour des données multivariées à intervalles censurés Ⓔ E

10:50-11:05 **Zixuan Zhao** (Queen's University) **Yanglei Song** (Queen's University) **Wenyu Jiang** (Queen's University) **Dongsheng Tu** (Queen's University)
Consistent covariances estimation for stratum imbalances under marginal design for covariate adaptive randomization / Estimation convergente des covariances pour les déséquilibres de strate sous plan marginal pour la randomisation adaptative selon les covariables Ⓔ E

11:05-11:35 **Raphael Robert McDonald** (Dalhousie University) **David Keith** (Fisheries and Oceans Canada) **Jessica Sameoto** (Fisheries and Oceans Canada) **Joanna Elizabeth Mills Flemming** (Dalhousie University)
Integrating habitat features into spatio-temporal biomass dynamics models for a better understanding of stock productivity: A case study of Sea Scallop in the Bay of Fundy / L'intégration de différents habitats à l'intérieur d'un modèle spatio-temporel de dynamique de la biomasse pour une meilleure compréhension de la productivité des stocks : Étude de cas sur le pétoncle de mer dans la Baie de Fundy ⒺⒻ E

10:20-11:50 **Contributed / Communications libres** (abstract/résumé 92) **TB 240**

Design and Analysis Approaches for Complex Survey Data
Méthodes de conception et d'analyse des données d'enquêtes complexes

Chair/Président: Zhiyang Zhou

10:20-10:35 **Wei Liang** (University of Waterloo) **Changbao Wu** (University of Waterloo)
Propensity Score Weighting with Post-Treatment Survey Data / Pondération du score de propension avec des données d'enquête post-traitement Ⓔ E

10:35-10:50 **Jia Ning Zhang** (University of Ottawa) **David Haziza** (University of Ottawa) **Sixia Chen** (University of Oklahoma)
Robust imputation in the presence of influential units in surveys / Imputation robuste en présence d'unités influentes dans les enquêtes Ⓔ Ⓕ Ⓔ

10:50-11:05 **Caren Hasler** (Université de Neuchâtel) **Esther Eustache** (Université de Neuchâtel)
Quasi-Model-Assisted Estimators under Nonresponse in Sample Surveys / Estimateurs quasi-assistés par modèle pour non-réponse dans les enquêtes Ⓔ Ⓔ

11:05-11:20 **Beili Huang** (University of Manitoba)
Statistical Models for Multilevel Data with "Don't know" category: Implications for program evaluation / Modèles statistiques pour des données multiniveaux avec la catégorie «ne sait pas» : conséquences sur l'évaluation de programme Ⓔ Ⓔ

10:20-11:50 **Contributed / Communications libres** (abstract/résumé 95) **TB 238**

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

Chair/Président: Simon J. Bonner

10:20-10:35 **Jervis Gallanosa** (University of Manitoba) **Yuliya V. Martsynyuk** (University of Manitoba) **Shivani Bhardwaj** (University of Manitoba)
Confidence Intervals for the Mean Based on the Functional CLT for a Weighted Student Process / Intervalles de confiance pour une moyenne basée sur le théorème de la limite centrale (CLT) fonctionnel dans un processus de Student pondéré Ⓔ Ⓔ

10:35-10:50 **Armin Hatefi** (Memorial University of Newfoundland)
Bayesian Mixture Modeling with Ranked Set Samples / Modélisation de mélanges bayésiens avec plans d'échantillonnage d'ensemble classé Ⓔ Ⓔ

10:50-11:05 **Karim Barigou** (Université Laval) **Pierre-Olivier Goffard** (Université de Strasbourg) **Stéphane Loisel** (Université Lyon 1) **Yahia Salhi** (Université Lyon 1)
Bayesian model averaging for mortality forecasting using leave-future-out validation / Modèle moyen bayésien pour prédire la mortalité par validation leave-future-out Ⓔ Ⓔ

11:05-11:20 **Mufan Li** (University of Toronto)
The Neural Covariance SDE: Shaped Infinite Depth-and-Width Networks at Initialization / La covariance neuronale SDE : réseaux à largeur-profondeur infinie modelés à l'initialisation Ⓔ Ⓔ

11:20-11:35 **Xu (Sunny) Wang** (Wilfrid Laurier University) **Yang Liu** (Wilfrid Laurier University) **Herteg Kohar** (Wilfrid Laurier University) **Doug Brown** (CIBC)
Enhancing Named Entity Recognition with Data Augmentation via Abstractive Summarization / Rehausser la reconnaissance d'entités nommées avec l'augmentation de données par l'entremise de synthèse par abstraction Ⓔ Ⓔ

11:35-11:50 **Ankana Dey** (Université de Sherbrooke) **F. Guillaume Blanchet** (Université de Sherbrooke)
Modelling the Human Microbiome using Metacommunity Theory / Modélisation du microbiome humain via la théorie des métacommunautés Ⓔ Ⓔ

10:20-11:50 **Contributed / Communications libres** (abstract/résumé 99) **RB 1200**

New Approaches for Dependence Modeling
Nouvelles approches de la modélisation de la dépendance

Chair/Président: Sudhir Paul

10:20-10:35 **Samantha-Jo Caetano** (University of Toronto) **Xiao Wu** (University of Toronto) **Wai Yu Amanda Ng** (University of Toronto) **Quynh Phuong Vu** (University of Toronto)
Development and implementation of a statistics consulting group / Développement et mise en place d'un groupe de consultation statistique Ⓔ Ⓔ

10:35-10:50 **Adel Ahmadi Nadi** (University of Waterloo) **Stefan Steiner** (University of Waterloo) **Nathaniel T. Stevens** (University of Waterloo)
Assessing the Agreement Between Multi-Operator Measurement Systems Using the Probability of Agreement / Évaluation de la concordance entre systèmes de mesure multi-opérateurs à l'aide de la probabilité de concordance Ⓔ E

10:50-11:05 **Zelalem Firisa Negeri** (University of Waterloo)
Identifying and Accommodating Outlying Studies in Diagnostic Test Meta-Analyses: a Mixture Modelling Approach / Identification et prise en compte des études aberrantes dans les méta-analyses de tests diagnostiques : une approche de modélisation des mélanges Ⓔ E

11:05-11:20 **Jairo Diaz-Rodriguez** (York University)
Spatial Estimation of Virus Infection Propensity from GPS Spatio-temporal Locations / Estimation spatiale de la propension d'infection virale à partir d'emplacements spatiotemporels GPS Ⓔ E

11:20-11:35 **Yuan Fang** (Binghamton University) **Jiachen Chen** (Boston University) **Joanne Murabito** (Boston University) **Kathryn Lunetta** (Boston University)
Modeling heterogeneity in cognitive trajectories in the Framingham Heart Study / Modélisation de l'hétérogénéité des trajectoires cognitives dans la Framingham Heart Study (FHS) Ⓔ E

11:35-11:50 **Kunasekaran Nirmalkanna** (Fisheries and Marine Institute of Memorial University of Newfoundland) **Noel Cadigan** (Fisheries and Marine Institute of Memorial University of Newfoundland) **S. J. W. W. M. M. P Weerasekera** (Fisheries and Marine Institute of Memorial University of Newfoundland)
Spatiotemporal relationship between weight and length of Atlantic cod on the southern Grand Bank of Newfoundland / Relation spatiotemporelle entre le poids et la longueur de la morue de l'Atlantique dans le sud des Grands Bancs de Terre-Neuve Ⓔ E

10:20-11:50 **Invited / Sur invitation** (abstract/résumé 103) **TB 210**

NSERC Discovery Grant Application Assistance
NSERC Discovery Grant Application Assistance

Chair/Président: David Haziza
Organizer/Responsable: W. John Braun
Sponsor/Commanditaires: Research Committee/Comité de la recherche

10:20-11:50 **Adele Ngi-Song** (NSERC) **Rachel Desrochers** (Natural Sciences and Engineering Research Council of Canada) **David Haziza** (University of Ottawa)
NSERC Discovery Grant Application Assistance / Aide à la soumission d'une demande de subvention à la découverte du CRSNG Ⓔ E

10:20-11:50 **Contributed / Communications libres** (abstract/résumé 104) **RB 1201**

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

Chair/Président: Wanhua Su

10:20-10:35 **Johanna de Haan-Ward** (University of Western Ontario) **Douglas G. Woolford** (University of Western Ontario) **Simon J. Bonner** (University of Western Ontario)
Modelling Predicted Probabilities of Rare Events Using Stratified Sampling, with Application to Human-Caused Wildland Fire / Modélisation des probabilités prédites d'événements rares à l'aide d'un échantillonnage stratifié, avec application aux incendies de forêt d'origine humaine Ⓔ E

10:35-10:50 **Meixi Chen** (University of Waterloo) **Martin Lysy** (University of Waterloo) **Reza Ramezan** (University of Waterloo)
Decoding Neural Population Dynamics Through Continuous-Time Latent Factor Models / Décodage de la dynamique d'une population neuronale à l'aide de modèles de facteurs latents en temps continu Ⓔ E

10:50-11:05 **Jianchu Chen** (University of Waterloo) **Richard J. Cook** (University of Waterloo)
A Copula Model for Recurrent and Terminal Events / Modèle de copule pour événement récurrent et terminal

11:05-11:20 **Fabiha Binte Farooq** (Simon Fraser University) **Christina Nieuwoudt** (Simon Fraser University)
Statistical prioritization of sequencing variants in disease-enriched families / Priorisation statistique des variables de séquençage dans des familles avec surreprésentation d'une maladie

11:20-11:35 **Jingyue Huang** (University of Waterloo) **Leilei Zeng** (University of Waterloo) **Changbao Wu** (University of Waterloo)
Sample Empirical Likelihood Approaches for Causal Inference / Approches de vraisemblance empirique fondée sur l'échantillon pour une inférence causale

11:35-11:50 **Ashani N. Wickramasinghe** (University of Manitoba) **Saman Muthukumarana** (University of Manitoba) **Surajith N. Wanasundara** (University of Manitoba) **Matt Schaubroeck** (ioAirFlow)
Identifying Locations with Abnormal Behavior of Temperature in School Buildings Using Anomaly Detection / Repérage de régions où la température se comporte anormalement dans les établissements scolaires au moyen de détection d'anomalie

---

13:30-15:00 **Invited / Sur invitation** (abstract/résumé 108) **AT 102**

Statistical design of Bayesian adaptive and platform trials
Conception statistique d'essais adaptatifs et de plates-formes bayésiennes

Chair/Président: Luke Hagar
Organizer/Responsable: Shirin Golchi

13:30-14:00 **Anna Heath** (The Hospital for Sick Children)
Efficient Design of Platform Trials with Complex Outcomes / Conception efficace d'études plateformes avec résultats complexes

14:00-14:30 **Steffen Ventz** (Univrsity of Minnesota)
Bayesian Multi-Arm De-Intensification Designs / Plans bayésiens de désintensification à plusieurs bras

14:30-15:00 **Shirin Golchi** (McGill University) **James Willard** (McGill University)
Estimating the Sampling Distribution of a Bayesian Probability Statement with Application to Clinical Trials / Estimation de la distribution échantillonnale d'un énoncé de probabilité bayésien avec application à des essais cliniques

---

13:30-15:00 **Invited / Sur invitation** (abstract/résumé 110) **AT 101**

In celebration of 50 years of research in Stochastics at Carleton University: Session in Honour of Miklós Csörgő
Célébrations de 50 ans de recherche en stochastique à l'Université Carleton : Session en l'honneur de Miklós Csörgö

Chair/Président: Barbara Szyszkowicz
Organizer/Responsable: Barbara Szyszkowicz
Sponsor/Commanditaires: Probability Section/Probability Section

13:30-14:00 **Barbara Szyszkowicz** (Carleton University)
Revisiting ICAMPS'97, 8-13 July 1997, Ottawa. / Revisiter ICAMPS' 97, du 8 au 13 juillet 1997, Ottawa.

14:00-14:30 **Yifan Li** (University of Western Ontario) **Reg Kulperger** (University of Western Ontario) **Hao Yu** (University of Western Ontario)
Do we still have a nice CLT under general model uncertainty? / Avons-nous encore un bon théorème central limite en cas d'incertitude générale du modèle ?

14:30-15:00 **Masoud Nasari** (Bank of Canada and Carleton University)
A new inferential framework for the mean / Nouveau cadre inférentiel pour la moyenne

---

13:30-15:00 **Invited / Sur invitation** (abstract/résumé 112) **TB 210**

Mortality Modelling and Longevity Risk
Modélisation de la mortalité et risque de longévité

Chair/Président: Jun Cai
Organizer/Responsable: Hong Li
Sponsor/Commanditaires: Actuarial Science Section/Actuarial Science Section

13:30-14:00 **Jianxi Su** (PURDUE UNIVERSITY)
Some statistical properties of the multivariate truncated normal distributions with actuarial applications in view / Quelques propriétés statistiques des distributions normales tronquées multivariées en vue d'applications actuarielles

14:00-14:30 **Kenneth Q. Zhou** (Arizona State University) **Hong Li** (University of Guelph) **Ze Chen** (Renmin University of China) **Yu Mao** (University of International Business and Economics)
Modeling and Managing Pandemic Mortality Risk via a Stochastic Compartmental Model / Modélisation et gestion du risque de mortalité lié à une pandémie à l'aide d'un modèle compartimental stochastique

14:30-15:00 **Hong Li** (University of Guleph) **Jianxi Su** (Purdue University)
Mitigating Wildfire Losses via Insurance-Linked Securities: Modeling and Risk Management Perspectives / Mitiger les pertes relatives aux feux incontrôlés grâce aux contrats d'assurance titrisés : perspectives de modélisation et de gestion de risques

13:30-15:00 **Invited / Sur invitation** (abstract/résumé 114) **TB 238**

Early Career Mentoring through the SSC Accreditation Program
Mentorat de début de carrière dans le programme d'accréditation de la SSC

Chair/Président: Darcy C. Pickard
Organizer/Responsable: Fabrice Larribe, Beatrice D. Baribeau
Sponsor/Commanditaires: Accreditation Committee/Accreditation Committee

13:30-14:00 **Judy-Anne W. Chapman** (Queen's University (retired))
Premise for Professional Mentoring of A.Stat. by P.Stat / Présentation de mentorat professionnel pour statisticiens associés (A. Stat) par des statisticiens professionnels (P.Stat.)

14:00-14:30 **Kevonn Morgan** (RBC - Manager, Global AML Insights & Analytics)
Accreditation Mentoring from the perspective of an A.Stat / Accréditation au mentorat selon un A.Stat

14:30-15:00 **May Raad** (HDR Corporation)
Accreditation Mentoring from the perspective of a P.Stat. / Mentorat au programme d'accréditation dans la perspective d'un statut de statisticien professionnel (P.Stat.)

13:30-15:00 **Invited / Sur invitation** (abstract/résumé 116) **AT 301**

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/ICSA Canada Chapter

13:30-14:00 **Oscar Hernan Madrid Padilla** (UCLA) **Hangjian Li** (University of California, Los Angeles) **Qing Zhou** (University of California, Los Angeles)
Learning Gaussian DAGs from Network Data / Apprentissage de DAG gaussiens à partir de données de réseau

14:00-14:30 **Peijun Sang** (University of Waterloo)
Nonlinear function-on-function Regression by RKHS / Régression de fonction sur fonction non linéaire à partir du RKHS Ⓔ E

14:30-15:00 **Archer Yang** (McGill University) **Yue Zhao** (University of York) **Yi Lian** (University of Pennsylvania) **Jun Fan** (McGill University) **Yuwen Gu** (University of Connecticut)
Flexible Regularized Estimating Equations: Some New Perspectives / Équations d'estimation régularisées flexibles : Quelques nouvelles perspectives Ⓔ E

13:30-15:00 **Invited / Sur invitation** (abstract/résumé 118) **TB 240**

Some recent research contributions at Statistics Canada
Quelques contributions récentes en recherche à Statistique Canada

Chair/Président: Jean-François Beaumont
Organizer/Responsable: Jean-François Beaumont
Sponsor/Commanditaires: Survey Methods/Survey Methods

13:30-14:00 **Keven Bosa** (Statistics Canada) **Jean-François Beaumont** (Statistics Canada)
How to allocate the sample to maximize benefits from small area estimation techniques? / Comment répartir l'échantillon pour maximiser les avantages des techniques d'estimation sur petits domaines ? Ⓔ E

14:00-14:30 **Abel C. Dasylva** (Statistics Canada) **Jean-François Beaumont** (Statistics Canada) **Keven Bosa** (Statistics Canada) **Guillaume Maranda** (Statistics Canada)
Measuring the accuracy of a prediction for a finite population total / Mesure de l'exactitude d'une prédiction du total d'une population finie Ⓔ E

14:30-15:00 **Marie-Hélène Toupin** (Statistics Canada)
Improving the coverage of confidence intervals using degrees of freedom / Améliorer la couverture des intervalles de confiance à l'aide des degrés de liberté Ⓕ E

13:30-15:00 **Invited / Sur invitation** (abstract/résumé 120) **TB 208**

Advances in spatial epidemiology
Progrès en épidémiologie spatiale

Chair/Président: Rob Deardon
Organizer/Responsable: Rob Deardon
Sponsor/Commanditaires: Biostatistics/Biostatistics

13:30-13:52 **Patrick Brown** (University of Toronto)
Daily air pollution and short-term health effects / Pollution atmosphérique quotidienne et effets à court terme sur la santé Ⓔ E

13:52-14:15 **Madeline Ward** (University of Calgary) **Rob Deardon** (University of Calgary) **Lorna Deeth** (University of Guelph) **Caitlin Ward** (University of Minnesota)
Incorporating Behavioural Change into Infectious Disease Transmission Models / Incorporation de changements comportementaux dans les modèles de transmission de maladies infectieuses Ⓔ E

14:15-14:38 **Dirk Douwes-Schultz** (McGill) **Alexandra Schmidt** (McGill University) **Yannan Shen** (McGill University) **David Buckeridge** (McGill University)
A Three-state Coupled Markov Switching Model for COVID-19 Outbreaks across Quebec based on Hospital Admissions / Un modèle de Markov à changement entre trois états pour étudier les épidémies de COVID-19 au Québec mesurées par les hospitalisations Ⓔ E

14:38-15:00 **Longhai Li** (University of Saskatchewan)
Cross-validatory Residual Diagnostics for Bayesian Spatial Models / Diagnostics résiduels à validation croisée pour les modèles spatiaux bayésiens Ⓔ E

13:30-15:00 **Invited / Sur invitation** (abstract/résumé 123) **RB 2200**

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

Chair/Président: Wesley S. Burr
Organizer/Responsable: Wesley S. Burr
Sponsor/Commanditaires: Statistics Education Section/Statistics Education Section

13:30-15:00 **Alison L. Gibbs** (University of Toronto)
Teaching Statistics through Changing Times: Lessons from my mother and others / Enseigner la statistique d'une époque à une autre : enseignements de ma mère et d'autres Ⓔ Ⓔ

13:30-15:00 **Contributed / Communications libres** (abstract/résumé 124) **TB 206**

Advanced Models in Finance and Insurance
Modèles avancés en finance et en assurance

Chair/Président: Hanna K Jankowski

13:30-13:45 **IsmaelAfolabi Assani** (Université de Montréal) **Maciej Augustyniak** (Université de Montréal)
Optimal Quadratic Hedging with Basis Risk under Bivariate Q-Affine GARCH Models / Couverture quadratique sous risque de base dans des modèles GARCH Q-affines Ⓔ Ⓔ

13:45-14:00 **Juan-Sebastian Yanez** (Université du Québec à Montréal) **Mathieu Pigeon** (Université du Québec à Montréal)
Dependence in Claim Processing Time: A Frailty Analysis Perspective / Dépendance dans le temps de traitement des réclamations : une perspective basée sur l'analyse de la fragilité Ⓔ Ⓔ

14:00-14:15 **Hashan Savinda Peiris Kalugama Gardige** (Simon Fraser University)
Integration of Traditional and Telematics data for Efficient Insurance Claims Prediction / Intégration de données télématiques et traditionnelles pour une prédiction efficace des réclamations d'assurance Ⓔ Ⓔ

14:15-14:30 **Philipp Ratz** (Université du Québec à Montréal)
Solving censored regression problems using a multitask approach / Données censurées sous l'angle d'une regression multitâche Ⓔ Ⓔ

14:30-14:45 **Sulalitha Bowala Mudiyanselage** (University of Manitoba)
Daily and Hourly Fuzzy Forecasts of Electricity Demand Using Neural Network Dynamic Regression Models / Prévisions horaires et quotidiennes floues de la demande en électricité à l'aide de modèles de régression dynamique par réseaux neuronaux Ⓔ Ⓔ

13:30-15:00 **Contributed / Communications libres** (abstract/résumé 127) **TB 202**

Missing Data and Patient-Focused Statistical Methods
Données manquantes et méthodes statistiques axées sur le patient

Chair/Président: Janie Coulombe

13:30-13:45 **Abdoulaye Dioni** (Université Laval)
Sensitivity analysis in the presence of a missing not at random categorical explanatory variable / Analyse de sensibilité en présence d'une variable explicative, catégorielle manquant non aléatoirement Ⓕ Ⓔ

13:45-14:00 **Md. Erfanul Hoque** (Thompson Rivers Univeristy) **Mahmoud Torabi** (University of Manitoba) **Elif Acar** (University of Manitoba)
A Heterogeneous Random Effects Covariance Matrix in Longitudinal Data with Missing Responses and Mismeasured Covariates / Matrice de covariance des effets aléatoires hétérogènes pour données longitudinales avec des réponses manquantes et des covariables mal mesurées Ⓔ Ⓔ

14:00-14:15 **Chendong Li** (University of Manitoba) **Depeng Jiang** (University of Manitoba) **Robert Tate** (University of Manitoba) **Philip St. John** (University of Manitoba)
Modeling developmental trajectories with nonrandomly missing data: Investigating trajectories of frailty using data from the Manitoba Follow-up Study / Modélisation de trajectoires de développement avec des données manquantes de manière non aléatoire : enquête sur les trajectoires de fragilité à l'aide de données de l'étude Manitoba Follow-up Study

14:15-14:30 **Marc Angelo Parsons** (McGill University) **Andrea Benedetti** (McGill University) **Russell Steele** (McGill University)
Flexible Methods for Trajectory Estimation in Individual Participant Data Meta-Analyses / Méthodes flexibles pour l'estimation de la trajectoire dans les méta-analyses de données individuelles sur les patients

14:30-14:45 **William Ruth** (Simon Fraser University) **Richard Lockhart** (Simon Fraser University)
A Monte Carlo EM Analysis of COVID-19 Outbreaks in Long-Term Healthcare Facilities / Analyse espérance-maximisation (EM) Monte-Carlo des éclosions de la COVID-19 dans les centres de soins de longue durée

14:45-15:00 **HensleyHubert Mariathas** (Memorial University of Newfoundland) **Shabnam Asghari** (Memorial University of Newfoundland) **Oliver Hurley** (Memorial University of Newfoundland)
Sample size calculation for stepped wedge cluster randomized trial when the number of clusters available is limited to small / Calcul de la taille d'échantillon pour les essais randomisés en grappes avec permutation séquentielle lorsque le nombre de grappes est limité ou faible

---

13:30-15:00 **Contributed / Communications libres** (abstract/résumé 131) **TB 236**

Recent Developments in models for event-history, case-cohort and mixed-type trait data
Développements récents dans les modèles pour données d'histoire d'événements, de cohortes de cas et de traits de type mixte

Chair/Président: Celia M.T. Greenwood

13:30-13:45 **Marcus Hlady** (University of Manitoba) **Yuliya V. Martsynyuk** (University of Manitoba)
Importance of Reliability Ratio for Properties of Asymptotic Confidence Intervals in Linear Structural Errors-in-Variables Models / Importance du rapport de fiabilité pour les propriétés des intervalles de confiance asymptotiques dans les modèles structurels linéaires d'erreurs dans les variables

13:45-14:00 **Jacqueline A. May** (University of Guelph) **Zeny Feng** (University of Guelph) **Sarah J. Adamowicz** (University of Guelph)
A real data-driven simulation strategy to select an imputation method for mixed-type trait data / Stratégie de simulation pilotée par des données réelles pour la sélection d'une méthode d'imputation pour des données de traits de type mixte

14:00-14:15 **Hameed A. Jimoh** (University of Saskatchewan) **Shahedul Khan** (University of Saskatchewan)
The proportional hazards assumption for joint models of longitudinal and time-to-event data / Hypothèse de risques proportionnels pour les modèles conjoints de données longitudinales et de temps avant l'événement

14:15-14:30 **Christian Chan** (University of Calgary) **Xuewen Lu** (University of Calgary) **Xiaotian Dai** (University of Calgary) **Thierry Chekouo** (University of Minnesota) **Quan Long** (University of Calgary)
Broken Adaptive Ridge Method for Variable Selection in Logistic Partly Linear Models with application to CATHGEN data / Méthode de crête adaptative brisée pour la sélection de variables dans les modèles logistiques partiellement linéaires avec application aux données CATHGEN

14:30-14:45 **Yan Liu** (University of Montreal) **Mireille Schnitzer** (Université de Montréal) **Ronald Herrera** (Université de Montréal; Bayer AG, Berlin, Germany) **Iván Díaz** (NYU Langone Health) **Jennifer O'Loughlin** (Université de Montréal) **Marie-Pierre Sylvestre** (Université de Montréal)
The application of target trials with longitudinal targeted maximum likelihood estimation to assess the effect of alcohol consumption in adolescence on depressive symptoms in adulthood / Application d'essais ciblés avec une estimation par maximum de vraisemblance ciblée longitudinale pour évaluer l'effet de la consommation d'alcool à l'adolescence sur les symptômes de dépression à l'âge adulte

13:30-15:00 **Contributed / Communications libres** (abstract/résumé 135) **RB 1201**

Advances in Functional and Longitudinal Data Analysis
Progrès en analyse des données fonctionnelles et longitudinales

Chair/Président: Nicholas Mitsakakis

13:30-13:45 **Myoungjin Ko** (Department of Statistics and Actuarial Science, University of Waterloo)
Adaptive Particle Markov Chain Monte Carlo for Jump Diffusion Models / MCMC à particules adaptatif pour modèles de diffusion à sauts

13:45-14:00 **Menelaos Konstantinidis** (University of Toronto: Dalla Lana School of Public Health) **Myanca Rodriques** (Department of Health Research Methods, Evidence, and Impact, McMaster University) **Sofia Tsokani** (Department of Primary Education, School of Education, University of Ioannina) **Ian R. White** (Medical Research Council Clinical Trials Unit at UCL, Institute of Clinical Trials and Methodology, University College London) **Julian P. T. Higgins** (Population Health Sciences, NIHR Applied Research Collaboration West (ARC West), University Hospitals Bristol and Weston NHS Foundation Trust) **Guido Schwarzer** (Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center, University of Freidburg) **Dimitris Mavridis** (Department of Primary Education, School of Education, University of Ioannina) **Andrea Tricco** (Division of Epidemiology & Institute of Health Policy, Management and Evaluation, University of Toronto. Li Ka Shing Knowledge Institute, St. Michael's Hospital, Unity Health Toronto) **Areti-Angeliki Veroniki** (Institute of Health Policy, Management and Evaluation, University of Toronto. Li Ka Shing Knowledge Institute, St. Michael's Hospital, Unity Health Toronto)
Assessing network meta-analysis inconsistency in the design-by-treatment interaction model: A simulation study / Évaluation de l'incohérence dans les méta-analyses en réseau dans le modèle d'interaction conçu par traitement (DBT) : une étude par simulation

14:00-14:15 **Sidi Wu** (Simon Fraser University) **Cédric Beaulac** (Université du Québec à Montréal) **Jiguo Cao** (Simon Fraser University)
Autoencoder for Discrete Functional Data Representation Learning and Smoothing / Auto-encodeur pour l'apprentissage et le lissage de représentations de données fonctionnelles discrètes

14:15-14:30 **Luis Ledesma** (University of Toronto: Dalla Lana School of Public Health) **Eleanor M. Pullenayegum** (Hospital for Sick Children)
Intercept estimation of semi-parametric joint models in the context of longitudinal data subject to irregular observations / Estimation de l'ordonnée à l'orgine de modèles conjoints semi-paramétriques dans un contexte de données longitudinales faisant l'objet d'observations irrégulières

14:30-14:45 **Chi Zhang** (University of Waterloo) **Peijun Sang** (University of Waterloo) **Yingli Qin** (University of Waterloo)
Two-sample inferences for sparse functional data / Inférences à deux échantillons pour des données fonctionnelles éparses

14:45-15:00 **Chi-Kuang Yeh** (University of Waterloo) **Gregory Rice** (University of Waterloo) **Joel A. Dubin** (University of Waterloo)
Functional Spherical Autocorrelation: Robust Autocorrelation Estimation of a Functional Time Series / Autocorrélation sphérique fonctionnelle : estimation d'autocorrélation robuste de séries temporelles fonctionnelles

13:30-15:00 **Contributed / Communications libres** (abstract/résumé 139) **RB 1200**

Data science methods and applications I
Méthodes et applications de la science des données I

Chair/Président: Jinko Graham

13:30-13:45 **Rose Garrett** (University of Toronto) **Eleanor M. Pullenayegum** (Hospital for Sick Children)
The Role of Sensitivity Analysis in Handling Untestable Assumptions in Analyses of EHR Data / Rôle de l'analyse de sensibilité dans le traitement des hypothèses non vérifiables dans les analyses de données de DSE

13:45-14:00 **Preethi Ravikumar** (McGill University)
Detecting Influential Points in High-Dimensional MRI Data Using Random Projections and Kernel Density Estimation / Détection de points influents dans des données IRM en haute dimension à l'aide de projections aléatoires et de l'estimation de la densité du noyau

14:00-14:15 **Jérémie Boudreault** (Institut National de la Recherche Scientifique) **Celine Campagna** (Institut national de santé publique du Québec (INSPQ)) **Fateh Chebana** (Institut national de la recherche scientifique (INRS))
Does machine and deep learning outperform statistical models in predicting the health effects of extreme heat? / L'apprentissage automatique et profond est-il meilleur que les modèles statistiques pour prédire les effets de la chaleur extrême sur la santé ?

14:15-14:30 **Ziqian Zhuang** (University of Toronto: Dalla Lana School of Public Health) **Wei Xu** (Princess Margaret Cancer Centre)
A Supervised Distance-based Feature Selection Algorithm for Multiple Mixed Outcomes / Un algorithme de sélection de caractéristiques basé sur la distance supervisée pour des résultats multiples mixtes

14:30-14:45 **Jasper Zhongyuan Zhang** (University of Toronto) **Wei Xu** (University of Toronto, Princess Margaret Cancer Centre) **Pingzhao Hu** (Western University)
Integrating Multiomics Data using Deep Tensor Factorization for Survival Outcomes Prediction / Intégration de données multiomiques en utilisant la factorisation de tenseur profond pour les résultats de survie dans le cancer

14:45-15:00 **Quanhan Xi** (The University of British Columbia) **Bloem-Reddy Benjamin** (University of British Columbia)
Indeterminacy in Generative Models: Characterization and Strong Identifiability / Indétermination dans les modèles génératifs : Caractérisation et identifiabilité forte

---

13:30-15:00 **Poster / Poster** (abstract/résumé **??**) **RB 2220-2224**

Case Study I: Optimizing Patient Flow in Emergency Department
Étude de cas I : Optimisation du flux de patients dans le service des urgences

Chair/Président: Chel Hee Lee
Organizer/Responsable: Chel Hee Lee

13:30-15:00 **Zehui Wang** (Queen's University) **Yangsai Lyu** (Queen's University) **Yuwei Ke** (Queen's University) **Xinrui Wang** (Queen's University)
Queen's University / Université Queen's

13:30-15:00 **Ravish Kamath** (York University) **My Luu** (York University) **Zayeeda Shahreen Labiba** (York University) **Zhiyin Chen** (York University)
York University / Université York

13:30-15:00 **Nayanthi Karunanayake** (University of Manitoba) **Samuel Morrissette** (University of Manitoba) **Ashani N. Wickramasinghe** (University of Manitoba) **Balage Don Harshani Hiranthika De Silva** (University of Manitoba)
University of Manitoba 1 / Université du Manitoba 1

13:30-15:00 **Ayman Arik Kazi** (McMaster University) **Ela Bandari** (University of Western Ontario) **Corey Fletcher** (Queen's University) **Morgan Rosenberg** (McMaster University)
McMaster University / University of Western Ontario / Queen's University / Université McMaster / Université de Western Ontario / Université Queen's

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) **Qiyue Zhang** (Queen's University)
University of Toronto / Queen's University / Université de Toronto / Université Queen's

13:30-15:00 **Jiayin He** (University of Alberta) **Ying Mei** (University of Alberta) **Jingyi Xu** (University of Alberta)
University of Alberta / Université de l'Alberta

13:30-15:00 **Jiayang He Xiyou Lin** (Dalhousie University)
Dalhousie University / Université Dalhousie

13:30-15:00 **Alexandra Mossman** (University of Waterloo) **Bryn Crandles** (University of Waterloo) **Gradon Nicholls** (University of Waterloo) **Megan French** (University of Waterloo)
University of Waterloo 1 / Université de Waterloo 1

13:30-15:00 **Junwei Shen** (McGill University) **Haoyu Wu** (McGill University) **Xianglin Zhao** (McGill University)
McGill University / Université McGill

13:30-15:00 **Christian Chan** (University of Calgary) **Chinmoy Roy Rahul** (University of Calgary) **Madeline Ward** (University of Calgary)
University of Calgary / Université de Calgary

13:30-15:00 **Patrick Fournier** (Université du Québec à Montréal) **Niki Z. Petrakos** (McGill University) **Mélanie Raymond** (Université du Québec à Montréal -UQAM) **Julien St-Pierre** (McGill University)
Université du Québec à Montréal / McGill University / Université du Québec à Montréal / Université McGill

13:30-15:00 **Nirodha M. Epasinghege Dona** (Simon Fraser University) **Yiting Chen** (Simon Fraser University) **Tak Shing (Dustin) Poon** (Simon Fraser University) **Hashan Savinda Peiris Kalugama Gardige** (Simon Fraser University)
Simon Fraser University 1 / Université Simon Fraser 1

13:30-15:00 **Yuan Xia** (The University of British Columbia) **Tae Yoon Lee** (University of British Columbia) **Laura Lim Huey Mien** (National University of Singapore)
University of British Columbia / National University of Singapore / Université de la Colombie-Britannique / Université nationale de Singapour

---

13:30-15:00 **Poster / Poster** (abstract/résumé **??**) **RB 2220-2224**

Case Study 2: Understanding how Canada's economy might be impacted by climate change
Étude de cas 2 : Comprendre comment l'économie canadienne pourrait être affectée par les changements climatiques

Chair/Président: Chel Hee Lee
Organizer/Responsable: Chel Hee Lee

13:30-15:00 **Yuxin Shi** (University of Toronto: Dalla Lana School of Public Health) **Deen Xu** (University of Toronto) **Yu-chun Chien** (University of Toronto) **Zhenhuan Xu** (University of Toronto)
University of Toronto 2 / Université de Toronto 2

13:30-15:00 **Greg Forkutza** (McMaster University) **Manan Mukherjee** (McMaster University) **Cameron Roopnarine** (McMaster University)
McMaster University 2 / Université McMaster 2

13:30-15:00 **Shreena Nisha Kalaria** (BC Cancer Research Centre)
University of Victoria / Université de Victoria

13:30-15:00 **Sonny Min** (Simon Fraser University) **Donghui Son** (Simon Fraser University)
Simon Fraser University 2 / Université Simon Fraser 2

13:30-15:00 **Shamsia Sobhan** (University of Manitoba) **Amin Abed** (University of Manitoba) **Md. Hasan** (University of Manitoba) **Justin Dyck** (University of Manitoba)
University of Manitoba 2 / Université du Manitoba 2

13:30-15:00 **Sunqiaohe Zheng** (York University) **Young Jo Chung** (University of Toronto) **Jinqi Guo** (York University)
York University / University of Toronto / Université York / Université de Toronto

13:30-15:00 **Erin Zhang** (Simon Fraser University) **Rina Wang** (Simon Fraser University) **Sidi Wu** (Simon Fraser University) **Tianxing Yan** (Simon Fraser University)
Simon Fraser University 3 / Université Simon Fraser 3

13:30-15:00 **Shiyu He** (University of Waterloo) **Yuying Huang** (University of Waterloo) **Trang Bui** (University of Waterloo) **Wenling Zhang** (University of Waterloo)
University of Waterloo 2 / Université de Waterloo 2

13:30-15:00 **Minh Chau Nguyen** (Department of Statistics and Actuarial Science, University of Waterloo) **David Awosoga** (University of Waterloo) **Alireza Ghaffartehrani** (University of Waterloo)
University of Waterloo 3 / Université de Waterloo 3

13:30-15:00 **Shijie Min** (University of Toronto) **Mu Yang** (University of Toronto) **Aoqi Xie** (University of Toronto) **Yushu Zou** (University of Toronto)
University of Toronto 1 / Université de Toronto 1

13:30-15:00 **Abhiroop Chowdhury** (McMaster University) **Jason Pekos** (McMaster University) **Shiheng Huang** (McMaster University)
McMaster University 1 / Université McMaster 1

13:30-15:00 **Abhinav Kansal** (University of British Columbia) **Divyansh Sharma** (University of British Columbia) **Aditya Chinchure** (University of British Columbia) **Kashish Hemal Joshipura** (University of British Columbia)
University of British Columbia 1 / Université de la Colombie-Britannique 1

13:30-15:00 **Saurav Neupane** (University of Prince Edward Island) **Aadesh Nunkoo** (University of Prince Edward Island)
University of Prince Edward Island / Université de l'Île-du-Prince-Édouard

---

15:30-17:00 **Invited / Sur invitation** (abstract/résumé 143) **AT 301**

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

Chair/Président: Kevin McGregor
Organizer/Responsable: Kevin McGregor
Sponsor/Commanditaires: The Committee on New Investigators/The Committee on New Investigators

15:30-16:00 **Jingyi Cao** (York University) **Dongchen Li** (Brock University) **Virginia R. Young** (University of Michigan) **Bin Zou** (University of Connecticut)
Reinsurance Games with Two Reinsurers: Tree Versus Chain / Jeux de réassurance avec deux réassureurs : arborescence versus chaîne

16:00-16:30 **Pratheepa Jeganathan** (McMaster University) **Shiheng Huang** (McMaster University) **Jamie McNicol** (McMaster University)
Probabilistic Topic Models and Tessellation to Uncover Spatial Compartmentalization in Spatial Omics / Modèles probabilistes de sujets et tessellation pour découvrir la compartimentation spatiale dans les données omiques spatiales

16:30-17:00 **Mohamed Belalia** (University of Windsor) **Guanjie Lyu** (University of Windsor)
Nonparametric hypotheses testing in copula models: Some recent avenues. / Test d'hypothèses non paramétrique dans des modèles de copule : quelques avenues récentes

---

15:30-17:00 **Invited / Sur invitation** (abstract/résumé 145) **TB 210**

Nonlinear and Nonparametric Regression Models in Environmental and Social Health
Modèles de régression non linéaires et non paramétriques en santé environnementale et sociale

Chair/Président: Alex Stringer
Organizer/Responsable: Alex Stringer

15:30-16:00 **Glen McGee** (University of Waterloo)
Testing for Non-Additive Interaction in Flexible Bayesian Models for Multi-Pollutant Mixtures / Test pour l'interaction non additive dans des modèles bayésiens flexibles pour des mélanges de multi-polluants Ⓔ E

16:00-16:30 **Michael Y.C. Chong** (University of Toronto) **Monica Alexander** (University of Toronto)
Estimating the Timing of Stillbirths in Countries Worldwide Using a Bayesian Hierarchical Penalized Splines Regression Model / Estimation des périodes de mortinatalité dans les pays du monde entier à l'aide d'un modèle de régression hiérarchique bayésien à splines pénalisées Ⓔ E

16:30-17:00 **Liqun Diao** (University of Waterloo) **Grace Y. Yi** (University of Western Ontario) **Haoxin Zhuang** (BeiGene Inc.)
Polya Tree Based Nearest Neighborhood Regression / Régression par méthode des plus proches voisins, basée sur l'arborescence de Pólya Ⓔ E

15:30-17:00 **Invited / Sur invitation** (abstract/résumé 147) **AT 102**

Modern Statistical Learning with Complex Large-scale Data
Apprentissage statistique moderne avec données complexes à grande échelle

Chair/Président: Chen Xu
Organizer/Responsable: Chen Xu

15:30-16:00 **Shili Lin** (The Ohio State University)
Sparse Estimation in Finite Mixture of Varying Coefficient Regression Models / Estimation éparse dans des mélanges finis de modèles de régression à coefficients variables Ⓔ E

16:00-16:30 **Shaista Jaffer** (University of Ottawa)
Event-driven Trading Strategies with Explainable Machine Learning / Stratégies de négociation événementielle à l'aide de l'apprentissage automatique explicable Ⓔ E

16:30-17:00 **Kaili Jing** (Xi'an JiaoTong University)
Low Gradient-based Subsampling for Corrupted Massive Data / Sous-échantillonnage basé sur un faible gradient pour mégadonnées corrompues Ⓔ E

15:30-17:00 **Invited / Sur invitation** (abstract/résumé 149) **AT 101**

In celebration of 50 years of research in Stochastics at Carleton University: Session in Honour of Donald A. Dawson
Célébrations de 50 ans de recherche en stochastique à l'Université Carleton : Session en l'honneur de Donald A. Dawson

Chair/Président: Bouchra Nasri
Organizer/Responsable: Bouchra Nasri
Sponsor/Commanditaires: Probability Section/Probability Section

15:30-16:00 **Bruno N. Rémillard** (HEC Montréal) **Jean Vaillancourt** (HEC Montreal)
Central limit theorem for martingales / Un théorème central limite pour martingales Ⓔ E

16:00-16:30 **Shui Feng** (McMaster University)
Relative Entropy and Hierarchical Dirichlet Process / Entropie relative et processus de Dirichlet hiérarchique Ⓔ E

16:30-17:00 **Jean Vaillancourt** (HEC Montréal) **Bruno N. Rémillard** (HEC Montréal)
Another central limit theorem plus a few applications / Un autre théorème limite central et quelques applications Ⓕ E

15:30-17:00 **Invited / Sur invitation** (abstract/résumé 151) **TB 208**

New Tools in Teaching Statistics
Nouveaux outils pour l'enseignement de la statistique

Chair/Président: Wanhua Su
Organizer/Responsable: Wanhua Su
Sponsor/Commanditaires: Statistics Education Section/Statistics Education Section

15:30-16:00 **Becky Wei Lin** (Simon Fraser University)
The Design and Use of Shiny Apps for Teaching Introductory Statistics Courses / Conception et utilisation des applications Shiny dans l'enseignement de cours d'introduction aux statistiques

16:00-16:30 **Jonathan Herman** (University of Toronto at Mississauga)
MathMatize: A New Tool for Mathematical and Statistical Education / MathMatize : un nouvel outil pour l'enseignement des mathématiques et des statistiques

15:30-17:00 **Invited / Sur invitation** (abstract/résumé 152) **RB 1200**

New Developments at the Interface of Finance and Pension
Nouveaux développements à l'interface des finances et des pensions

Chair/Président: Jean-François Bégin
Organizer/Responsable: Jean-François Bégin
Sponsor/Commanditaires: Actuarial Science Section/Actuarial Science Section

15:30-16:00 **Thomas Salisbury** (York University)
Regulatory Constraints and the Riccati Tontine / Les contraintes réglementaire et la tontine de Riccati

16:00-16:30 **Barbara Sanders** (Simon Fraser University) **Jean-François Bégin** (Simon Fraser University) **Wenyuan Zhou** (Simon Fraser University)
To Merge or not to Merge? On the Impacts of Pension Plan Consolidation / Fusionner ou ne pas fusionner ? Des impacts de la consolidation des régimes de retraite

16:30-17:00 **David Saunders** (University of Waterloo)
Risk Sharing in Pension Plans: Optimality and Fairness / Partage du risque dans les régimes de pension : optimalité et équité

15:30-17:00 **Contributed / Communications libres** (abstract/résumé 154) **TB 236**

Analytic Approaches for Novel Data Sources
Approches analytiques pour nouvelles sources de données

Chair/Président: Hugh Chipman

15:30-15:45 **Cédric Beaulac** (Université du Québec à Montréal)
Developing neural network architectures for functional data analysis / Développer de nouvelles architectures de réseaux de neurones pour l'analyse de données fonctionnelles

15:45-16:00 **Mahsa Panahi** (Department of Statistics and Actuarial Science, University of Waterloo) **Stefan Steiner** (University of Waterloo) **Jeroen de Mast** (University of Amsterdam)
Identifying Dominant Causes using Leveraged Study Designs / Identification des causes dominantes à l'aide de plans d'étude à effet de levier

16:00-16:15 **Matthew Baxter** (University of Guelph)
Comparing Ensemble Methods for a Two-Level Individual Level Model Parameter Estimation / Comparaison de méthodes d'ensemble pour l'estimation des paramètres d'un modèle à deux niveaux individuels

16:15-16:30 **Jinxin Yuan** (University of Manitoba)
Bayesian Analysis of Dolly Varden Mark-recapture Data in Canadian Arctic / Analyse bayésienne de données de marquage-recapture du Dolly Varden dans l'archipel arctique canadien

16:30-16:45 **Trang Bui** (University of Waterloo) **Stefan Steiner** (University of Waterloo) **Nathaniel T. Stevens** (University of Waterloo)
General Additive Network Effect Models / Modèles généraux d'effet additif de réseau

16:45-17:00 **Dena L. Schanzer** (Public Health Agency of Canada (retired)) **Douglas G. Woolford** (University of Western Ontario)
Perspectives on Data Translation from Fields Outside of the Statistical Science / Perspectives sur la traduction de données provenant de domaines externes aux sciences statistiques

---

15:30-17:00 **Contributed / Communications libres** (abstract/résumé 158) **TB 238**

On Causal Inference Methods and Applied Statistics
Des méthodes d'inférence causale et de la statistique appliquée

Chair/Président: Celia M.T. Greenwood

15:30-15:45 **Nirodha M. Epasinghege Dona** (Simon Fraser University)
A Causal Investigation of Pace of Play in Soccer / Étude causale du rythme de jeu au soccer

15:45-16:00 **Jinko Graham** (Simon Fraser University) **Payman Nickchi** (Simon Fraser University) **Charith Karunarathna** (Simon Fraser University)
An Exploration of Linkage Fine-Mapping on Sequences from Case-Control Studies / Une exploration de la cartographie fine de liaison sur les séquences tirées d'études cas-témoins

16:00-16:15 **Yuliang Shi** (University of Waterloo) **Yeying Zhu** (University of Waterloo) **Joel A. Dubin** (University of Waterloo)
Variable Selection for Causal Modeling in Missing Exposure Problems / Sélection de variables pour la modélisation causale dans des problèmes d'exposition manquante

16:15-16:30 **Mengjie Bian** (McMaster University) **Angelo J. Canty** (McMaster University)
Investigating the Winner's Curse in Mendelian Randomization / Enquête sur la malédiction du vainqueur dans la randomisation mendélienne

---

15:30-17:00 **Contributed / Communications libres** (abstract/résumé 161) **TB 202**

New Tools for Bayesian and Hierarchical Analysis
Nouveaux outils pour l'analyse bayésienne et hiérarchique

Chair/Président: Mai Ghannam

15:30-15:45 **Balage Don Harshani Hiranthika De Silva** (University of Manitoba) **Saman Muthukumarana** (University of Manitoba) **Tony Szturm** (University of Manitoba)
Bayesian Analysis for Comparing Computerized Manual Dexterity Performance Metrics Among Children with Cerebral Palsy / Comparaison de mesures de dextérité manuelle informatisées chez les enfants atteints d'infirmité motrice cérébrale par analyse bayésienne

15:45-16:00 **Arjun Banik** (University of Victoria) **Laura Cowen** (University of Victoria) **Saman Muthukumarana** (University of Manitoba) **Colin Gallagher** (Fisheries and Ocean, Canada)
Analyzing demographic parameters of anadromous Dolly Varden (Salvelinus malma malma) using Bayesian multi-state capture-recapture modelling / Analyse de données démographiques du Dolly Varden anadrome (Salvelinus malma malma) à l'aide d'une modélisation multi-états de capture-recapture bayésienne

16:00-16:15 **Masud Rana** (University of Saskatchewan) **Shahedul Khan** (University of Saskatchewan) **Punam Pahwa** (University of Saskatchewan)
Covariate measurement error in spatial analysis of count data / Erreur de mesure de covariables dans l'analyse spatiale de données de dénombrement

16:15-16:30 **Mingwei Xu** (University of Waterloo)
Bayesian Integration Method for Parameter Estimation in Ordinary Differential Equation / Méthode d'intégration bayésienne pour l'estimation de paramètre dans une équation différentielle ordinaire

16:30-16:45 **Ajmery Jaman** (McGill University)
Penalized G-estimation for Effect Modifier Selection in the Structural Nested Mean Model for Repeated Outcomes / G-estimation pénalisée pour la sélection de modificateurs d'effet dans le modèle structurel de moyenne imbriquée pour des résultats répétés

16:45-17:00 **Samuel Morrissette** (University of Manitoba) **Saman Muthukumarana** (University of Manitoba) **Maxime Turgeon** (University of Manitoba)
Parsimonious Bayesian Model-Based Clustering with Dissimilarities / Regroupement avec dissimilarité basé sur un modèle bayésien parcimonieux

---

15:30-17:00 **Contributed / Communications libres** (abstract/résumé 165) **TB 240**

Data science methods and applications II
Méthodes et applications de la science des données II

Chair/Président: Pingzhao Hu

15:30-15:45 **Sebastian F. Calcetero** (University of Toronto) **Andrei L. Badescu** (University of Toronto) **X. Sheldon Lin** (University of Toronto)
A Credibility Index Approach for Effective a Posteriori Ratemaking with Large Insurance Portfolios / Une approche d'indice de crédibilité pour une tarification a posteriori efficace avec de grandes portefeuilles d'assurance

15:45-16:00 **Shimon Nauenberg** (University of Toronto) **Nicholas Mitsakakis** (CHEO Research Institute and University of Toronto)
Weak Supervision for Detecting Mental Health Issues from Social Media Data / Supervision faible pour la détection des problèmes de santé mentale à partir des données des médias sociaux

16:00-16:15 **Zhaoran Hou** (University of Waterloo)
Estimating Boltzmann Averages for Protein Structural Quantities Using Sequential Monte Carlo / Estimer les moyennes de Boltzmann pour les quantités structurelles protéiques au moyen de la séquentielle de Monte Carlo

16:15-16:30 **Jie Jian** (University of Waterloo)
Dynamic Community Detection of the International Apple Trading Network / Détection dynamique de communautés dans le réseau international de commerce de pommes

16:30-16:45 **McNealis Vanessa** (McGill University) **Erica E. M. Moodie** (McGill University) **Nema Dean** (University of Glasgow)
Joint Mixed-Effects Models for Causal Inference in Network-Based Observational Studies / Modèles à effets mixtes conjoints pour l'estimation d'effets causaux en présence d'interférence réseau

16:45-17:00 **Karina Kwan** (McGill University) **Celia M. T. Greenwood** (McGill University) **David Soave** (Wilfrid Laurier University)
A Weighted Casebase Framework for Analyzing Case-Cohort Study Data / Cadre de travail basé sur les cas pondérés pour l'analyse de données d'études cas-cohortes

Tuesday May 30 — mardi 30 mai

08:30-09:50 — **Invited / Sur invitation** (abstract/résumé 169) — **RB 2200**

SSC 2022 Gold Medal Address
Allocution du récipiendaire de la Médaille d'or de la SSC 2022

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

08:30-09:50 **David A. Stephens** (McGill University)
What Makes a 'Bayesian' Analysis Bayesian ? / Qu'est-ce qui fait qu'une analyse bayésienne est « bayésienne » ?

10:20-11:50 — **Invited / Sur invitation** (abstract/résumé 170) — **AT 102**

Practical challenges and novel methods for estimating causal effect from observational data
Défis pratiques et nouvelles méthodes d'estimation de l'effet causal à partir de données d'observation

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

10:20-10:50 **Lan Wen** (University of Waterloo)
Robust Estimation of the Effect of Intervening Variable in Settings with Unmeasured Confounding / Estimation robuste de l'effet de la variable d'intervention dans les situations comportant des confondantes non mesurées

10:50-11:20 **Mireille Schnitzer** (Université de Montréal) **Ashkan Ertefaie** (University of Rochester) **Denis Talbot** (Université Laval) **Guanbo Wang** (Harvard University) **David Berger** (Université de Montréal) **Jennifer O'Loughlin** (Université de Montréal) **Marie-Pierre Sylvestre** (Université de Montréal)
Longitudinal outcome-adaptive and marginal fused LASSO for confounder selection and model pooling with time-varying treatments / LASSO fusionné adaptatif longitudinal pour la sélection de facteurs de confusion et regroupement de modèles à temps de traitement variables

11:20-11:50 **Denis Talbot** (Université Laval)
A generalized double-robust Bayesian model averaging approach to causal effect estimation / Une approche généralisée et doublement robuste de moyenne de modèles bayésienne pour l'estimation d'un effet causal

---

10:20-11:50 — **Invited / Sur invitation** (abstract/résumé 172) — **TB 210**

The Collaborative Statistician: Learning the Tricks of the Trade
Le statisticien collaboratif : Apprendre les ficelles du métier

Chair/Président: Bouchra Nasri
Organizer/Responsable: Tolu Sajobi
Sponsor/Commanditaires: Biostatistics/Biostatistics

10:20-10:35 **Lehana Thabane** (McMaster University)
Statistical Leadership in the Time of Crisis: Lessons from COVID Trials / Leadership statistique en temps de crise : Leçons des essais cliniques COVID

10:35-10:50 **Lisa M. Lix** (University of Manitoba)
Lessons on Collaborations in Health Services & Policy Research: Integrating Knowledge Translation with Statistical Research / Leçons sur les collaborations dans le domaine de la recherche sur les services et les politiques de santé : Intégrer l'application des connaissances à la recherche statistique

10:50-11:05 **Joanna Elizabeth Mills Flemming** (Dalhousie University)
Lessons Learned while Developing Statistical Methods to Better Understand the Ocean and its Inhabitants / Leçons apprises lors du développement de méthodes statistiques cherchant à comprendre davantage l'océan et ses habitants

11:05-11:20 **Tolu Sajobi** (University of Calgary)
Time is Brain: Lessons on Statistical Collaborations in a Fast-Paced Multidisciplinary World of Acute Stroke / Le temps, c'est du cerveau : Leçons sur les collaborations statistiques dans le monde multidisciplinaire et rapide de l'accident vasculaire cérébral aigu Ⓔ E

11:20-11:50 **Bouchra Nasri** (Université de Montréal)
Panel Discussion and Q&A / Débat et questions-réponses Ⓔ E

10:20-11:50 **Invited / Sur invitation** (abstract/résumé 175) **AT 101**

Statistics in Schools
La statistique dans les écoles

Chair/Président: Samantha-Jo Caetano
Organizer/Responsable: Samantha-Jo Caetano
Sponsor/Commanditaires: Statistics Education Section/Statistics Education Section

10:20-10:50 **Chrystal E. Smith** (York University)
Using R-programming in the K-12 Classroom for Mathematical modelling and statistical literacy development. / Utiliser le logiciel R dans la salle de classe K-12 pour la modélisation mathématique et l'apprentissage statistique Ⓔ E

10:50-11:20 **Diana Katherine Skrzydlo** (University of Waterloo)
High School Outreach – Machine Learning, Finance, and Data Bias / Rayonnement au niveau secondaire – Apprentissage automatique, finance et biais dans les données Ⓔ E

11:20-11:50 **Nathalie Sjarova** (Statistics Canada) **Georgia Fakiolas** (Statistics Canada / Statistique Canada) **Dmitri Bakker** (Statistics Canada / Statistique Canada) **Miranda Cox** (Statistics Canada / Statistique Canada) **Warren McDougald** (Statistics Canada / Statistique Canada)
Statistics Canada Census Education Toolkits / Trousses éducatives du Recensement de Statistique Canada ⒺⒻ E F

10:20-11:50 **Invited / Sur invitation** (abstract/résumé 177) **AT 301**

Recent Advances in mixture-model based classification
Avancées récentes de la classification basée sur des modèles de mélange

Chair/Président: Ryan P Browne
Organizer/Responsable: Ryan P Browne
Sponsor/Commanditaires: Business and Industrial Statistics Section (BISS)/Business and Industrial Statistics Section (BISS)

10:20-10:50 **Jeffrey L. Andrews** (University of British Columbia, Okanagan)
Variable Importance in Clustering with Finite Mixture Models / Importance des variables dans le regroupement avec des modèles de mélange finis Ⓔ E

10:50-11:20 **Sanjeena Dang** (Carleton University) **Anjali Silva** (University of Toronto) **Xiaoke Qin** (Carleton University) **Steven Rothstein** (University of Guelph) **Paul David McNicholas** (McMaster University)
Clustering matrix-variate discrete data / Regroupement de données discrètes à variables matricielles Ⓔ E

11:20-11:50 **Alexander Sharp** (University of Waterloo) **Ryan P. Browne** (University of Waterloo)
Maximum Contribution to the Likelihood: Precise Estimation with the Stochastic Expectation-Maximization Algorithm / Contribution des maxima à la vraisemblance : Estimation précise avec l'algorithme espérance-maximisation stochastique Ⓔ E

10:20-11:50 **Invited / Sur invitation** (abstract/résumé 179) **TB 208**

Data Science Ethics and Fairness
Éthique et équité en science des données

Chair/Président: Joel Ostblom
Organizer/Responsable: Joel Ostblom
Sponsor/Commanditaires: Data Science and Analytics Section/Data Science and Analytics Section

10:20-11:05 **Sina Fazelpour** (Northeastern University, Canada)
Disciplining deliberation: On interpreting machine learning trade-offs in sociotechnical systems / Discipliner la délibération : de l'interprétation des compromis de l'apprentissage automatique dans les systèmes sociotechniques

11:05-11:50 **Heather Krause** (We All Count)
Data Equity: How not to use data like a racist / Équité des données : comment ne pas utiliser les données comme un raciste

10:20-11:50 **Invited / Sur invitation** (abstract/résumé 181) **RB 2200**

Spotlight on CANSSI postdocs
Vitrine des postdocs de l'INCASS

Chair/Président: Andrea Benedetti
Organizer/Responsable: Andrea Benedetti
Sponsor/Commanditaires: CANSSI/CANSSI

10:20-10:38 **Gracia Dong** (University of Toronto - St. Michael's Hospital)
Analyzing Indian Mortality with Temperature Attribution using Case-Crossover Models / Analyse de la mortalité indienne avec attribution de température au moyen de modèles de cas croisés

10:38-10:56 **François-Michel Boire** (HEC Montréal) **Hatem Ben-Ameur** (HEC Montréal) **Mark Reesor** (Wilfrid Laurier University) **Lars Stentoft** (Western University)
Capital Structure Modelling with Dynamic Programming Under Alternative Processes / Modélisation de la structure du capital par programmation dynamique en présence des processus alternatifs

10:56-11:14 **Ismaila Ba** (York University)
Linear response variational Bayes (LRVB) approximation for a zero-inflated probabilistic PCA (ZIP-PCA) model / Approche bayésienne variationnelle de réponse linéaire (LRVB) pour un modèle d'ACP probabiliste avec excès de zéros (ZIPPCA)

11:14-11:32 **Dylan Z. Spicker** (McGill University) **Erica E. M. Moodie** (McGill University)
Preserving Patient Privacy in Dynamic Treatment Regimes / Préservation de la vie privée des patients dans les régimes de traitement dynamiques

11:32-11:50 **Antonio Herrera Martin** (University of Toronto) **Gwendolyn Eadie** (University of Toronto) **Radu Craiu** (University of Toronto) **David Stenning** (Simon Fraser University) **Derek Bingham** (Simon Fraser University) **Bryan Gaensler** (University of Toronto)
Classification of FRBs / Classification des sursauts radio rapides (FRB)

10:20-11:50 **Invited / Sur invitation** (abstract/résumé 184) **TB 236**

Recent Advances in Small Area Estimation
Progrès récents en estimation des petites surfaces

Chair/Président: Zeinab Mashreghi
Organizer/Responsable: Mahmoud Torabi
Sponsor/Commanditaires: Survey Methods/Survey Methods

10:20-10:50 **Jonnagadda N. Rao** (Carleton University) **Victor Esteavo** (Statistics Canada) **Jean-François Beaumont** (Statistics Canada) **Keven Bosa** (Statistics Canada)
Comparison of area level and unit level estimators when unit level data are available / Comparaison des estimateurs au niveau du domaine et au niveau d'unité quand des données au niveau de l'unité sont disponibles Ⓔ E

10:50-11:20 **Samuel Sombo** (Statistics Canada) **Jean-François Beaumont** (Statistics Canada) **Keven Bosa** (Statistics Canada) **Cynthia Bocci** (Statistics Canada)
Using Random Forest for Small Area Estimation / L'utilisation de forêts aléatoires pour l'estimation sur petits domaines Ⓔ E F

11:20-11:50 **Gauri Datta** (U.S. Census Bureau) **Juhyung Lee** (University of Florida) **Jiacheng Li** (University of Georgia)
Pseudo-Bayesian Small Area Estimation / Estimation pseudo-bayésienne pour petits domaines Ⓔ E

10:20-11:50 **Contributed / Communications libres** (abstract/résumé 186) **RB 1200**

Probability, Stochastic processes and Inference Methods
Probabilité, processus stochastiques et méthodes d'inférence

Chair/Président: Yifan Li

10:20-10:35 **Deli Li** (Lakehead University) **Yu Miao** (Henan Normal University, China) **Yongcheng Qi** (University of Minnesota Duluth, USA)
A new type of results on probabilities of moderate deviations for i.i.d. random variables / Nouveau type de résultats sur les probabilités de déviations modérées pour les variables aléatoires iid Ⓔ E

10:35-10:50 **Ziang Zhang** (University of Toronto) **Alex Stringer** (University of Waterloo) **Patrick Brown** (University of Toronto) **James Stafford** (University of Toronto)
Bayesian Smoothing and Inference of Derivatives through Integrated Wiener Processes / Lissage bayésien et inférence de dérivées par l'entremise de processus intégrés de Wiener Ⓔ E

10:50-11:05 **Kilani Ghoudi** (United Arab Emirates University) **Mohamed Chaouch** (Qatar University) **Naamane Laib** (CY University)
Joint parametric specification checking of conditional mean and volatility in time series models with martingale difference innovations / Vérification de spécification paramétrique conjointe de la moyenne conditionnelle et volatilité dans des modèles de séries temporelles avec innovations de la différence de martingale Ⓔ E

11:05-11:20 **Haodi Liang** (University of British Columbia) **Haodi Liang** (University of British Columbia) **Jiahua Chen** (University of British Columbia)
Global Consistency of Empirical Likelihood / Cohérence globale et vraisemblance empirique Ⓔ E

11:20-11:35 **Pankaj Uttam Bhagwat** (University of Sherbrooke) **William Strawderman** (Rutgers University, USA)
On the construction of Bayes minimax multiple shrinkage estimators / Construction d'estimateurs minimax de rétrécissement multiples bayésiens Ⓔ E

11:35-11:50 **Yunhong Lyu** (University of Windsor) **Sévérien Nkurunziza** (University of Windsor)
On generalized square-root diffusion processes and Inference problems / Sur les processus de diffusion de type racines carrées et problèmes d'inférence Ⓔ E

10:20-11:50 **Contributed / Communications libres** (abstract/résumé 190) **TB 238**

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

Chair/Président: Yan Yuan

10:20-10:35 **Mehdi Dagdoug** (University of Ottawa) **Camelia Goga** (Université de Franche-Comté) **David Haziza** (University of Ottawa)
High-dimensional model-assisted estimation for survey data / Estimation assistée par modèle dans un contexte de haute dimension pour des données d'enquête Ⓔ E

10:35-10:50 **Boyi Hu** (Simon Fraser University)
Simultaneous Functional Quantile Regression / Régression fonctionnelle simultanée des quantiles

10:50-11:05 **Andrea Payne** (Carleton University) **Anjali Silva** (University of Toronto) **Steven Rothstein** (University of Guelph) **Paul David McNicholas** (McMaster University) **Sanjeena Dang** (Carleton University)
Utilizing a Family of Mixtures of Multivariate Poisson Log-Normal Distributions for Clustering High Dimensional Multivariate Count Data / Utilisation d'une famille de mélanges de distributions log-normales de Poisson multivariées pour le regroupement de données de comptage multivariées en haute dimension

11:05-11:20 **Yuan Zhong** (York University) **Wei Xu** (University Health Network, University of Toronto) **Xin Gao** (York University)
High-dimensional Data Integration with Multiple Heterogeneous Tasks / Intégration de données en haute dimension avec de multiples tâches hétérogènes

11:20-11:35 **Victoire Michal** (McGill University) **Jon Wakefield** (University of Washington) **Alexandra Schmidt** (McGill University)
Comparison of Small Area Prediction methods with high dimensional auxiliary information / Comparaison de méthodes de prédiction pour des petits domaines en présence d'information auxiliaire de haute dimension

---

10:20-11:50 **Contributed / Communications libres** (abstract/résumé 193) **TB 240**

Novel Methods in generalized linear models
Nouvelles méthodes pour les modèles linéaires généralisés

Chair/Président: Sayantee Jana

10:20-10:35 **Chel Hee Lee** (Alberta Health Services & University of Calgary) **Jason Waechter** (University of Calgary) **Christopher Doig** (University of Calgary) **Mikelis Bickis** (University of Saskatchewan)
Imprecise Inference for Clinical Reasoning under Near or Complete Ignorance / Inférence imprécise pour le raisonnement clinique selon une ignorance complète ou presque complète

10:35-10:50 **Peizhi Li** (Queen's University) **Yingwei Peng** (Queen's University) **Jianing Zheng** (Dongbei University of Finance and Economics)
An Empirical Comparison between Gradient Boosting Methods and Cox's Proportional Hazards Model for Right-censored Survival Data / Une comparaison empirique entre les méthodes de boosting de gradient et le modèle de risques proportionnels de Cox pour les données de survie censurées à droite

10:50-11:05 **Kelly Ramsay** (York University) **Shojaeddin Chenouri** (University of Waterloo)
FKWC Tests for Differences in the Covariance Structure of Functional Data / Des tests FKWC pour reconnaître les différences dans la structure de covariance des données fonctionnelles

11:05-11:20 **Alex Stringer** (University of Waterloo) **Blair Bilodeau** (University of Toronto) **Yanbo Tang** (Imperial College London)
Stochastic Convergence Rates and Applications of Adaptive Quadrature in Bayesian Inference / Taux de convergence stochastique et applications de quadrature adaptative dans l'inférence bayésienne

11:20-11:35 **James H. McVittie** (University of Regina)
Nonparametric Survival Function Estimation using Current Lifetimes: An alternative to the Grenander Estimator / Estimation de fonction de survie non paramétrique à l'aide de durées de vie actuelles : une solution de remplacement à l'estimateur Grenander

11:35-11:50 **Gurbakhshash Singh** (Central Connecticut State University) **Gordon Fick** (University of Calgary (Emeritus))
Considerations for the use of a logit or a log link in generalized linear models / Réflexions sur l'utilisation d'une fonction logit ou log dans les modèles linéaires généralisés

---

10:20-11:50 **Contributed / Communications libres** (abstract/résumé 197) **TB 202**

Models for Clustered and Recurrent Data
Modèles pour les données regroupées et récurrentes

Chair/Président: Md. Erfanul Hoque

10:20-10:35 **Yixiu Liu** (University of Manitoba) **Depeng Jiang** (University of Manitoba)
Joint Modeling of Longitudinal and Survival Data: An Approach via Poisson Models for the Survival Part under Bayesian Framework / Modélisation conjointe de données de survie et longitudinales : une approche par modèles de Poisson pour la survie dans un cadre bayésien Ⓔ E

10:35-10:50 **Aida Eslami** (Université Laval) **Hervé Abdi** (School of Behavioral and Brain Sciences, The University of Texas at Dallas, USA)
Cluster analysis of qualitative and mixed variables / Regroupement de variables qualitatives et mixtes Ⓔ E

10:50-11:05 **Xiaoke Qin** (Carleton University)
Clustering Human Gut Microbiome Data Using Finite Mixtures of Generalized Dirichlet-Multinomial Distributions / Regroupement de données sur le microbiome intestinal humain à l'aide de mélanges finis de distributions multinomiales de Dirichlet généralisées Ⓔ E

11:05-11:20 **FrançoisA Marshall** (Queen's University)
Efficient Reconstruction of Low-frequency Oscillations in Long Cyclostationary Processes: Novel Coherence Diagnostics / Reconstruction efficace des oscillations à basse fréquence dans des processus quasi-périodique : nouveaux diagnostics de cohérence Ⓔ E

13:30-15:00 **Invited / Sur invitation** (abstract/résumé 200) **TB 208**

Recent Advances in Compositional Data Analysis and Applications
Avancées récentes dans l'analyse et les applications des données compositionnelles

Chair/Président: Connie Stewart
Organizer/Responsable: Connie Stewart, Michail Tsagris

13:30-14:00 **Michail Tsagris** (University of Crete) **Abdulaziz Alenazi** (Northern Border University) **Connie Stewart** (University of New Brunswick)
A Jensen-Shannon divergence based k-NN algorithm for missing value imputation in compositional data / Algorithme k-NN basé sur la divergence de Jensen-Shannon pour l'imputation des valeurs manquantes dans les données de composition Ⓔ E

14:00-14:30 **Janice Lea Scealy** (Australian National University)
Score matching for microbiome compositional data / Appariement de score pour les données de composition du microbiome Ⓔ E

14:30-15:00 **Hongzhe Lee** (University of Pennsylvania)
A Tensor Decomposition Model for Longitudinal Microbiome Studies / Modèle de décomposition tensorielle pour études longitudinales du microbiome Ⓔ E

13:30-15:00 **Invited / Sur invitation** (abstract/résumé 202) **AT 301**

Statistical Modelling and Computational Intelligence for High-Dimensional Biological Data
Modélisation statistique et intelligence computationnelle pour les données biologiques de haute dimension

Chair/Président: Aerambamoorthy A Thavaneswaran
Organizer/Responsable: You Liang
Sponsor/Commanditaires: Biostatistics/Biostatistics

13:30-14:00 **Xuekui Zhang** (University of Victoria)
Statistical methods for cell type annotation using single-cell genomic data / Méthodes statistiques pour l'annotation de type de cellule au moyen de données génomiques de cellule unique Ⓔ E

14:00-14:30 **Li Xing** (University of Saskatchewan)
grouppred: A novel machine learning approach for gene module identification and prediction via a co-expression network of single-cell sequencing data / grouppred : une nouvelle approche d'apprentissage automatique pour l'identification et la prédiction de module génétique par l'entremise d'un réseau de co-expression de données de séquençage à cellule unique

14:30-15:00 **You Liang** (Toronto Metropolitan University)
Hyperspectral microscopy-based semi-automatic segmentation of eye tissues / Segmentation semi-automatique basée sur une microscopie hyperspectrale pour les tissus oculaires

---

13:30-15:00 **Invited / Sur invitation** (abstract/résumé 204) **AT 101**

Recent Developments by Early Career Researchers in Bayesian Modelling for Correlated Data
Développements récents par les chercheurs en début de carrière en modélisation bayésienne pour les données corrélées

Chair/Président: Osvaldo Espin-Garcia
Organizer/Responsable: Osvaldo Espin-Garcia

13:30-14:00 **Kuan Liu** (University of Toronto)
Bayesian causal analysis with clustered data / Analyse causale bayésienne à l'aide de données en grappes

14:00-14:30 **Zihang Lu** (Queen's University)
A Bayesian latent class model for clustering multiple longitudinal features / Modèle bayésien de classes latentes pour le regroupement de plusieurs variable longitudinales

14:30-15:00 **Aya A. Mitani** (University of Toronto) **Chen Chen** (University of Toronto) **Dipankar Bandyopadhyay** (Virginia Commonwealth University)
Fitting bivariate skew-t linear mixed models for longitudinal periodontal data: A Stan recourse / Ajustement de modèles mixtes linéaires bivariés à distribution skew-t pour des données parodontales longitudinales : une procédure Stan

---

13:30-15:00 **Invited / Sur invitation** (abstract/résumé 206) **TB 238**

Statistical Inference in Stochastic Models
Inférence statistique dans les modèles stochastiques

Chair/Président: Natalia Stepanova
Organizer/Responsable: Natalia Stepanova
Sponsor/Commanditaires: Probability Section/Probability Section

13:30-14:00 **Mohamedou Ould Haye** (Carleton University) **Anne Philippe** (Nantes University, France) **Caroline Robet** (Nantes University, France)
Inference for continuous-time long memory randomly sampled processes / Inference for continuous-time long memory randomly sampled processes

14:00-14:30 **Andrei Volodin** (University of Regina) **Rita Giuliano** (University of Pisa, Italy) **Manuel Ordóñez Cabrera** (University of Sevilla, Spain)
On the sub-Gaussianity of the r-correlograms / De la sous-gaussianité des r-corrélogrammes

14:30-15:00 **Marie Turcicova** (Institute of Atmospheric Physics CAS) **Natalia Stepanova** (Carleton University)
Sparse signal recovery in regression models with Gaussian white noise / Récupération de signaux "sparses" dans les modèles de régression avec bruit blanc gaussien

---

13:30-15:00 **Invited / Sur invitation** (abstract/résumé 208) **TB 210**

Statistics in transportation
Statistique dans les transports

Chair/Président: Pratheepa Jeganathan
Organizer/Responsable: Aurelie Labbe
Sponsor/Commanditaires: Business and Industrial Statistics Section (BISS)/Business and Industrial Statistics Section (BISS)

13:30-14:00 **Lijun Sun** (McGill University)
Bayesian Complementary Kernelized Learning for Multidimensional Spatiotemporal Data / Apprentissage noyauté bayésien complémentaire pour les données spatiotemporelles multidimensionnelles

14:00-14:30 **Alexandra Schmidt** (McGill University) **Hassan Rezaee** (HEC, Montreal) **Joshua Stipancic** (Intact Insurance) **Aurélie Labbe** (HEC, Montreal)
A process convolution model for crash count data on a network / Modèle de convolution de processus pour les données de dénombrement des accidents sur un réseau

14:30-15:00 **Aurélie Labbe** (HEC, Montreal)Devon **Mengying Lei** (McGill University) **Lijun Sun** (McGill University)
BKTR - Bayesian Kernelized Tensor Regression: application to bike-sharing demand modeling / BKTR - Bayesian Kernelized Tensor Regression : application à la modélisation de la demande de vélos en libre service

13:30-15:00 **Invited / Sur invitation** (abstract/résumé 210) **RB 1200**

Model Uncertainty in Insurance and Finance
Incertitude des modèles en assurance et finance

Chair/Président: Fangda Liu
Organizer/Responsable: Fangda Liu
Sponsor/Commanditaires: Actuarial Science Section/Actuarial Science Section

13:30-14:00 **Silvana Manuela Pesenti** (University of Toronto) **Sebastian Jaimungal** (University of Toronto) **Yuri Saporito** (Getulio Vargas Foundation (FGV)) **Rodrigo Targino** (Getulio Vargas Foundation (FGV))
Risk Budgeting Allocation for Dynamic Risk Measures / Allocation de la budgétisation du risque pour des mesures de risque dynamiques

14:00-14:30 **Wenjun Jiang** (University of Calgary) **Tim Boonen** (University of Amsterdam)
Distributionally Robust Insurance with the Wasserstein Distance / Assurance à distribution robuste avec la distance de Wasserstein

14:30-15:00 **Haiyan Liu** (Michigan State University)
Worst-Case Risk with Unspecified Risk Preferences / Risque le plus défavorable avec préférences de risque non spécifiées

13:30-15:00 **Invited / Sur invitation** (abstract/résumé 212) **AT 102**

Statistical Inference by Combining Information from Multiple Sources
Inférence statistique par combinaison d'informations provenant de sources multiples

Chair/Président: Wen Wilson Lu
Organizer/Responsable: Changbao Wu
Sponsor/Commanditaires: Survey Methods/Survey Methods

13:30-14:00 **Peisong Han** (University of Michigan)
Causal inference with improved efficiency by integrating external summary information in the possible presence of heterogeneous study populations / Inférence causale avec efficacité accrue en intégrant une information sommaire externe en présence possible de populations d'étude hétérogènes

14:00-14:30 **Jae-Kwang Kim** (Iowa State University) **Zhonglei Wang** (Xiamen University)
Repreducing kernel method for data integration / Méthode à noyau reproduisant pour l'intégration des données Ⓔ E

14:30-15:00 **Changbao Wu** (University of Waterloo)
Issues with Analysis of Non-Probability Survey Samples / Problèmes liés à l'analyse d'échantillons d'enquête non probabiliste Ⓔ E

13:30-15:00 **Invited / Sur invitation** (abstract/résumé 214) **RB 2200**

SSC Impact Award
Prix Impact de la SSC

Chair/Président: Tolu Sajobi
Organizer/Responsable: Tolu Sajobi

13:30-15:00 **Lehana Thabane** (McMaster University)
My journey as a biostatistician: "EBM=Statistical thinking in health" / Mon parcours de biostatisticien : « Médecine fondée sur des données probantes = pensée statistique dans le domaine de la santé » Ⓔ E

13:30-15:00 **Poster / Poster** (abstract/résumé 215) **RB 2220-2224**

Poster Presentations
Présentations d'affiches

13:30-15:00 **Simon Snyman** (University of British Columbia) **John Braun** (University of British Columbia Okanagan) **Lengyi Han** (University of British Columbia Okanagan)
On Data Sharpening in Nonparametric Autoregressive Models / Sur l'affinement des données dans les modèles autorégressifs non paramétriques Ⓔ E

13:30-15:00 **Ying Miao** (University of Manitoba) **Po Yang** (University of Manitoba)
Optimal single-stratum and multi-stratum designs for the dual purpose of model discrimination and parameter estimation / Conceptions optimales avec multi-strates et une seule strate visant deux buts : la discrimination du modèle et l'estimation des paramètres Ⓔ E

13:30-15:00 **Brittany Alana Perry** (Brock University) **William Marshall** (Brock University) **Mei Ling Huang** (Brock University)
Boosting Methods for Classification with Small Sample Size / Méthodes de boosting pour la classification avec de petites tailles d'échantillon Ⓔ E

13:30-15:00 **Wanye Gao** (University of British Columbia)
From simulation to observation – comparing spatial heterogeneity of radiation at the cellular level / De la simulation à l'observation – comparaison de l'hétérogénéité spatiale de la radiation au niveau cellulaire Ⓔ E

13:30-15:00 **Danika M. Lipman** (University of Calgary)
A Bayesian Regression model with Gaussian Process for Semi-Continuous Response / Modèle de régression bayésien avec processus gaussien pour une réponse semi-continue Ⓔ E

13:30-15:00 **Haoxuan Ge** (University of Toronto at Mississauga) **Jue Wang** (University of Toronto)
Spatial Non-Stationarity Effects of Unhealthy Food Environments and Green Spaces for Type-2 Diabetes in Toronto / Effets de non-stationnarité spatiale des environnements alimentaires malsains et des espaces verts pour le diabète de type 2 à Toronto Ⓔ E

13:30-15:00 **Yan Yuan** (University of Alberta) **Yutong Han** (University of Alberta)
How a Misclassified Binary Outcome Y in Training Data Affects Model Prediction Performance: a Simulation Study / Une étude en simulations relative aux effets d'un résultat binaire Y classé incorrectement dans des données d'apprentissage sur la performance de prédiction de modèle Ⓔ E

13:30-15:00 **Dana Ivancevic** (University of Ottawa) **The Minh Luong** (Bureau of Food Surveillance and Science Integration, Food Directorate, Health Products and Food Branch, Health Canada) **Lidia Loukine** (Bureau of Food Surveillance and Science Integration, Food Directorate, Health Products and Food Branch, Health Canada) **Dominique Ibañez** (Bureau of Food Surveillance and Science Integration, Food Directorate, Health Products and Food Branch, Health Canada) **Ali Karimnezhad** (Bureau of Food Surveillance and Science Integration, Food Directorate, Health Products and Food Branch, Health Canada; University of Ottawa) **Kelly Burkett** (University of Ottawa)
Is There an Association Between Dietary Sodium Intake and Life Expectancy in the Canadian Adult Population? / Existe-t-il une association entre l'apport en sodium alimentaire et l'espérance de vie dans la population adulte canadienne ?

13:30-15:00 **Xiao Wu** (University of Toronto) **Wai Yu Amanda Ng** (University of Toronto)
Statistical Methods of Variant Calling in Next-Generation Sequence Data Analysis / Méthodes statistiques d'appel de variants dans l'analyse des données de séquençage de nouvelle génération

13:30-15:00 **Yiting Chen** (Simon Fraser University) **Xiaoping Shi** (University of British Columbia)
Multivariate Two-sample Test Statistics Based on Data Depth / Variables à tester pour deux échantillons multivariés basés sur la profondeur de données.

13:30-15:00 **Quynh Phuong Vu** (University of Toronto)
The Conundrum of COVID-19 Self-Testing Devices / L'énigme des autotests COVID-19

13:30-15:00 **Karen A. Kopciuk** (Cancer Care Alberta, AHS)
Data and Knowledge: Considerations for Working with Data from Indigenous People / Données et savoir : les précautions à adopter en travaillant avec des données se rapportant aux peuples autochtones

13:30-15:00 **Sévérien Nkurunziza** (University of Windsor) **Yunhong Lyu** (University of Windsor)
Inference in Generalized Exponential Ornstein-Uhlenbeck Processes with Change-point / Inférence dans les processus d'Ornstein-Uhlenbeck exponentiels généralisés

13:30-15:00 **Linke Li** (University of Toronto: Dalla Lana School of Public Health)
Efficiently Evaluating the Economic Benefits of Group Sequential Design using Bayesian Decision Theory / Évaluation efficace des avantages économiques d'une méthodologie séquentielle par groupes au moyen de la théorie de décision bayésienne

13:30-15:00 **Christopher Salim** (Brock University) **S. Ejaz Ahmed** (Brock University) **Tianyu Guan** (Brock University)
A Study of Soccer Space Gain in Pass Sequences Using Logistic Regression / Étude du gain d'espace au soccer dans les séquences de passes à l'aide de la régression logistique

13:30-15:00 **Wai Yu Amanda Ng** (University of Toronto) **Sofia Panasiuk** (University of Toronto)
A Clustering Approach to Bottom-up Theories of Subjective Well-being: Is Life Satisfaction Differentially Predicted by Levels of Domain Satisfaction / Une approche de regroupement des théories ascendantes du bien-être subjectif : La satisfaction à l'égard de la vie est-elle différemment prédite selon les niveaux des domaines de satisfaction ?

13:30-15:00 **Alexandra Mossman** (University of Waterloo)
Dynamic Treatment Regimes for Clustered Data with Between- and Within-Group Interference / Régimes de traitement dynamiques pour données en grappes avec interférence entre les groupes et à l'intérieur des groupes

13:30-15:00 **EmanSreeni Abbas** (University of Manitoba) **Brad C. Johnson** (University of Manitoba)
A jackknife type procedure for modelling networks / Procédure de type jackknife pour la modélisation de réseaux

13:30-15:00 **Guanyu Chen** (The University of British Columbia) **Ching-Lin Shih** (National Sun Yat-Sen University) **Yan Liu** (Carleton University)
Exploring How Response Time predict Student Engagement in English Language Proficiency using Computerized Adaptive Test Data / Étude du temps de réponse comme prédicteur de l'engagement des étudiants à maîtriser la langue anglaise à l'aide de données de test adaptatif informatisé

13:30-15:00 **Raymond Romaniuk** (Brock University) **William Marshall** (Brock University) **Mei Ling Huang** (Brock University)
Boosting Methods for Imbalanced Data Classification / Méthodes de boosting pour la classification de données déséquilibrées

13:30-15:00 **Kyuson Lim** (McMaster University) **Utkarsh J. Dang** (Carleton University) **Sanjeena Dang** (Carleton University)
Mixtures of contaminated multivariate Poisson-lognormal models / Mélanges de modèles multivariés Poisson-lognormale contaminés

13:30-15:00 **Samopriya Basu** (Simon Fraser University (Department of Statistics and Actuarial Science)) **Faezeh Yazdi** (Simon Fraser University)
Bayesian inversion and uncertainty quantification for muon tomography / Inversion bayésienne et quantification des incertitudes pour la tomographie muonique

---

13:30-15:00 **Contributed / Communications libres** (abstract/résumé 227) **TB 240**

Statistical Models for Clinical and Healthcare Data
Modèles statistiques pour les données cliniques et de santé

Chair/Président: Lisa M. Lix

13:30-13:45 **Yanqing Yi** (Memorial University of Newfoundland)
Optimality and Efficiency of Sequential Designs of Clinical Trials / Optimalité et efficacité des plans séquentiels d'essais cliniques

13:45-14:00 **Kehinde I. Olobatuyi** (University of Victoria, Canada) **Laura Cowen** (University of Victoria) **Caroline Colijn** (Simon Fraser University) **Patrick Brown** (University of Toronto)
Hierarchical Zero-Inflated Occupancy Model for Estimating Population Abundance of Bacteria Serotypes with Low-Detection Probabilities / Modèle hiérarchique d'occupation avec excès de zéros pour l'estimation de l'abondance de la population de sérotypes bactériens avec de faibles probabilités de détection

14:00-14:15 **Wanhua Su** (MacEwan University)
Pre-Post Test Using Partial Pairs with Missing Identities / Pre-post test utilisant des paires partielles avec identités manquantes

14:15-14:30 **Wenshu Dai** (Binghamton University) **Yuan Fang** (Binghamton University) **Sanjeena Dang** (Carleton University)
Mixtures of Regression Models for Microbiome Data / Mélanges de modèles de régression pour des données sur le microbiome

14:30-14:45 **Mohammad Ehsanul Karim** (The University of British Columbia) **Momenul Haque Mondol** (University of British Columbia)
Guideline for Choosing the Number of Splits for Double Cross-fit Targeted Maximum Likelihood Estimators / Lignes directrices pour le choix du nombre de fractionnements pour des estimateurs du maximum de vraisemblance ciblé par ajustement croisé double

14:45-15:00 **Nam-Anh Tran** (The Hospital for Sick Children)
A Comparison of Alternative Ranking Methods in Two-Stage Clinical Trials with Multiple Interventions: An Application to the Anxiolysis for Laceration Repair in Children Trial / Une comparaison des méthodes de classement alternatives dans des essais cliniques à deux étapes avec multiples interventions : une application à l'anxiolyse pour la réparation de lacérations en essai pédiatrique

---

13:30-15:00 **Contributed / Communications libres** (abstract/résumé 230) **RB 1201**

On Goodness-of-fit and Inference Methods
Sur la qualité de l'ajustement et des méthodes d'inférence

Chair/Président: Mohamed Belalia

13:30-13:45 **Luke Hagar** (University of Waterloo) **Nathaniel T. Stevens** (University of Waterloo)
Fast Sample Size Determination for Two-Group Equivalence Tests with Unequal Variances / Détermination rapide de taille d'échantillon pour les tests d'équivalence en deux groupes avec variances inégales

13:45-14:00 **Lenin Arango-Castillo** (Bank of Mexico) **Francisco J. Martínez-Ramírez** (Banco de México) **María José Orraca Corona** (Banco de México)
Univariate Measures of Persistence: A Comparative Analysis / Analyse comparative des mesures univariées de la persistance

14:00-14:15 **Wangshu Tu** (Carleton University) **Sanjeena Dang** (Carleton University) **Utkarsh J. Dang** (Carleton University)
Changepoint detection using Gaussian mixture models / Détection de points de changement à l'aide de modèles de mélange gaussien

14:15-14:30 **Shiyu He** (University of Waterloo) **Samuel W.K. Wong** (University of Waterloo)
Integrating Satellite Images and in situ Datasets for Inference and Prediction of Harmful Algal Blooms with Application to Western Lake Erie / Intégration d'images satellitaires et de données in situ pour l'inférence et la prévision d'efflorescences algales nuisibles, avec application à l'ouest du lac Érié

14:30-14:45 **Tatiana Krikella** (University of Waterloo) **Joel A. Dubin** (University of Waterloo)
Personalized Predictive Model that Jointly Optimizes Discrimination and Calibration / Un modèle prédictif personnalisé qui optimise conjointement la discrimination et la calibration

14:45-15:00 **Mai Ghannam** (University of Windsor) **Sévérien Nkurunziza** (University of Windsor)
Inference in a tensor regression model with a possible change-point / Inférence dans un modèle de régression tensoriel avec un éventuel point de rupture

---

13:30-15:00 **Contributed / Communications libres** (abstract/résumé 234) **TB 236**

Methods for Genetic Association Studies
Méthodes pour les études d'association génétique

Chair/Président: Anna Heath

13:30-13:45 **Yiran Wang** (University of Waterloo) **Martin Lysy** (University of Waterloo) **Audrey Béliveau** (University of Waterloo)
Genetic Mark-Recapture Methods for Estimating Population Size of Stock Populations / Méthodes de marquage et de recapture génétique pour estimer la taille de population des populations de stocks

13:45-14:00 **Fousseni Sama** (Université Laval)
Comparative study of Cis-Regulatory Hubs observed from bulk organized cell data (Hi-C) to those observed in single-cell data (scHi-C) / Étude comparative des pôles cis-régulateurs observés à partir des données des cellules organisées en vrac (Hi-C) à ceux observées dans les données unicellulaires (scHi-C)

14:00-14:15 **Devan G. Becker** (Wilfrid Laurier University)
Estimating Proportions of Genetic Variants and the Variants Themselves from Wastewater Using Spatiotemporal Information / Estimation des proportions de variants génétiques et des variants mêmes dans les eaux usées à l'aide de données spatio-temporelles

14:15-14:30 **Lucy Gao** (University of British Columbia) **Anna Neufeld** (University of Washington) **Joshua Popp** (Johns Hopkins University) **Alexis Battle** (Johns Hopkins University) **Daniela Witten** (University of Washington)
Inference after latent variable estimation for single-cell RNA sequencing data / Inférence après estimation de la variable latente pour des données de séquençage de l'ARN unicellulaire

14:30-14:45 **Pingzhao Hu** (Western University) **Leann Lac** (University of Manitoba) **Eric Lin** (University of Toronto) **Boyuan Liu** (University of Toronto) **Daryl L.X. Fung** (University of Manitoba) **Carson K. Leung** (University of Manitoba)
scGMM-VGAE: A Gaussian Mixture Model-based Variational Graph Autoencoder Algorithm for Clustering Single-cell RNA-seq Data / scGMM-VGAE : un algorithme à autoencodeur de graphe variationnel basé sur un modèle de mélanges gaussiens pour regrouper les données de séquençage de l'ARN à cellule unique Ⓔ E

15:30-17:00 **Invited / Sur invitation** (abstract/résumé 237) **AT 102**

Advanced statistical methods for complex data with measurement error
Méthodes statistiques avancées pour les données complexes avec erreur de mesure

Chair/Président: Li-Pang Chen
Organizer/Responsable: Li-Pang Chen

15:30-16:00 **Li-Pang Chen** (National Chengchi University) **Grace Y. Yi** (University of Western Ontario)
Variable selection and estimation for the average treatment effect with error-prone confounders / Sélection de variables et estimation pour l'effet de traitement moyen avec confondants sujets à erreur Ⓔ E

16:00-16:30 **Jou-Chin Wu** (National Chengchi University) **Li-Pang Chen** (National Chengchi University)
Semiparametric Estimation for Error-prone Partial Linear Single-index Models / Estimation semiparamétrique pour les modèles linéaires partiels à indice unique et sujets à erreur Ⓔ E

16:30-17:00 **Qihuang Zhang** (McGill University) **Grace Y. Yi** (University of Western Ontario)
Generalized Network Structured Models with Mixed Responses subject to Measurement Error and Misclassification / Modèles structurés de réseau généralisés avec des réponses mixtes sujettes à l'erreur de mesure et à la mauvaise classification Ⓔ E

15:30-17:00 **Invited / Sur invitation** (abstract/résumé 239) **TB 240**

Emerging researchers in statistical ecology
Nouveaux chercheurs en écologie statistique

Chair/Président: Johanna de Haan-Ward
Organizer/Responsable: Johanna de Haan-Ward

15:30-16:00 **Rowenna Gryba** (The University of British Columbia) **Andrew VonDuyke** (North Slope Borough Department of Wildlife Management) **Henry Huntington** (Huntington Consulting) **Billy Adams** (North Slope Borough Department of Wildlife Management) **Brower Frantz** (Utqiaġvik, Alaska) **Justin Gatten** (Utqiaġvik, Alaska) **Qaiyaan Harcharek** (Utqiaġvik, Alaska) **Robert Sarren** (Utqiaġvik, Alaska) **Greg Henry** (University of British Columbia) **Marie Auger-Méthé** (University of British Columbia)
Disrupting the Traditions of Science: Indigenous Knowledge and Habitat Modelling / Perturbation des traditions scientifiques : connaissances autochtones et modélisation de l'habitat Ⓔ E

16:00-16:30 **Ethan Lawler** **Chris Field** (Dalhousie University) **Kim Whoriskey** (Dalhousie University) **Hugues Benoit** (Fisheries and Oceans Canada) **Joanna Elizabeth Mills Flemming** (Dalhousie University)
Estimating Space Use from Animal Telemetry Data using Non-Parametric Langevin Diffusion / Estimation de l'utilisation de l'espace à partir de données de télémétrie animale à l'aide de la diffusion de Langevin non paramétrique Ⓔ E

16:30-17:00 **Inesh Prabuddha Munaweera Arachchilage** (University of Manitoba) **Saman Muthukumarana** (University of Manitoba) **Darren Gillis** (University of Manitoba)
Bayesian modelling and simulation methods for acoustic telemetry studies / Modélisation bayésienne et méthodes de simulation pour des études de télémétrie acoustique Ⓔ E

15:30-17:00 **Invited / Sur invitation** (abstract/résumé 241) **AT 101**

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
Sponsor/Commanditaires: Biostatistics/Biostatistics

15:30-16:00 **Osvaldo Espin-Garcia** (Western University)
Leveraging Genetic Correlation Methods for Polygenic Risk Score Construction in Multivariate Outcomes via Penalized Linear Regression Models / Tirer avantage des méthodes de corrélation génétique pour la construction de score de risque polygénique dans des résultats multivariés au moyen de modèles de régression linéaire pénalisés Ⓔ E

16:00-16:30 **Kevin McGregor** (York University)
Dependence modelling in count-based genomic data in the presence of zero-inflation / Modélisation de la dépendance des données génomiques basée sur le comptage en présence d'un excès de zéros Ⓔ E

16:30-17:00 **Elena Tuzhilina** (University of Toronto) **Trevor Hastie** (Stanford University) **Mark Segal** (University of California, San Francisco)
Statistical Curve Models for Inferring 3D Chromatin Architecture / Modèles de courbes statistiques pour déduire l'architecture de la chromatine en 3D Ⓔ E

15:30-17:00 **Invited / Sur invitation** (abstract/résumé 243) **AT 301**

Causal inference for complex data
Inférence causale pour les données complexes

Chair/Président: Archer Yang
Organizer/Responsable: Archer Yang

15:30-15:53 **Yunan Wu** (University of Texas at Dallas) **Lan Wang** (University of Miami) **Haoda Fu** (Eli Lilly and Company)
Model-Assisted Uniformly Honest Inference for Optimal Treatment Regimes in High Dimension / Inférence uniformément honnête assistée par modèle pour les régimes de traitements optimaux en grande dimension Ⓔ E

15:53-16:15 **Eric Morenz** (University of Washington) **Marco Carone** (University of Washington) **Charles Wolock** (University of Washington)
Debiased machine learning for survival functionals / Apprentissage machine débiaisé pour des fonctionnelles de survie Ⓔ E

16:15-16:37 **Ziang Niu** (University of Pennsylvania)
Inference for ATE under Unmeasured Confounding with Arbitrary Number of Invalid IVs: Beyond Majority and Plurality Voting / Inférence pour l'effet moyen du traitement en cas de confusion non mesurée avec un nombre arbitraire de VI invalides : au-delà du vote à la majorité et à la pluralité Ⓔ E

16:37-17:00 **Guanbo Wang** (CausaLab)
Examining subgroup-specific treatment effects in multi-source data: source-specific inference and transportability to an external population / Examen des effets de traitement spécifiques à un sous-groupe dans des données multi-sources : inférence spécifique à la source et transférabilité à une population externe Ⓔ E

15:30-17:00 **Invited / Sur invitation** (abstract/résumé 246) **TB 208**

Recent Developments in High-Dimensional Inference
Développements récents en matière d'inférence en haute dimension

Chair/Président: Farouk Nathoo
Organizer/Responsable: Farouk Nathoo
Sponsor/Commanditaires: Business and Industrial Statistics Section (BISS)/Business and Industrial Statistics Section (BISS)

15:30-16:00 **S. Ejaz Ahmed** (Brock University)
Improving the Prediction Performance in High-Dimensional Data Analysis / Améliorer l'efficacité de prédiction dans l'analyse des données de grande dimension

16:00-16:30 **Linglong Kong** (University of Alberta)
Regularized tensor quantile regression with applications to neuroimaging data analysis / Régression quantile tensorielle régularisée et applications à l'analyse de données de neuro-imagerie

16:30-17:00 **Abbas Khalili** (McGill University) **Gilberto Chavez-Martinez** (McGill Univeristy)
Sparse Estimation in Markov Regime-Switching Models / Estimation éparse dans les modèles de Markov à changement de régime

15:30-17:00 **Contributed / Communications libres** (abstract/résumé 248) **TB 202**

Novel methods in Actuarial Sciences
Nouvelles méthodes en sciences actuarielles

Chair/Président: Thierry Chekouo

15:30-15:45 **Hyukjun (Jay) Gweon** (Western University) **Shu Li** (Western University)
A hybrid data mining framework for variable annuity portfolio valuation / Cadre hybride d'exploration de données pour l'évaluation de portefeuilles de rentes variables

15:45-16:00 **Alexandru Badescu** (University of Calgary) **Maciej Augustyniak** (Université de Montréal) **Jean-François Bégin** (Simon Fraser University) **Sarath Kumar Jayaraman** (University of Calgary)
On the Weak Limits of Affine Option Pricing Models / Des limites faibles des modèles affines d'évaluation des options

16:00-16:15 **Maciej Augustyniak** (Université de Montréal)
Basis Risk in Variable Annuities / Le risque de base dans les rentes variables

16:15-16:30 **Frédéric Godin** (Concordia University)
Risk allocation through Shapley decompositions, with applications to variable annuities / Allocation du risque à l'aide de décompositions de Shapley et applications aux fonds distincts

16:30-16:45 **Marie-Claude Vachon** (Université du Québec à Montréal) **Anne Mackay** (Université de Sherbrooke)
Optimal stopping with discontinuous, unbounded, and time-dependent reward and applications to variable annuities / Problème de temps d'arrêt optimal avec fonction de gain discontinue, non bornée et dépendante du temps et applications aux fonds distincts

16:45-17:00 **Emma Kroell** (University of Toronto) **Silvana Manuela Pesenti** (University of Toronto) **Sebastian Jaimungal** (University of Toronto)
Optimal robust reinsurance with multiple insurers / Réassurance robuste optimale pour de multiples assureurs

15:30-17:00 **Contributed / Communications libres** (abstract/résumé 251) **TB 238**

Innovations in Statistical and Data Science Education
Innovations dans l'enseignement de la statistique et de la science des données

Chair/Président: Wanhua Su

15:30-15:45 **Bethany J. G. White** (University of Toronto)
Are Undergraduate Life Science Students Prepared for Quantitative Research? / Les étudiants de premier cycle en sciences de la vie sont-ils préparés à la recherche quantitative ?

15:45-16:00 **Sohee Kang** (University of Toronto Scarborough) **Justin Slater** (University of Toronto)
Authoring Peer-Reviewed Questions in Introductory Statistics / Créer des questions évaluées par les pairs en introduction aux statistiques

16:00-16:15 **Kevin Mongeon** (University of Ottawa) **Janique Dubois** (University of Ottawa)
Reconciling The Grade Gap Between Indigenous And Non-Indigenous Students / Réconcilier l'écart de niveau entre les élèves autochtones et non autochtones

16:15-16:30 **Suborna Shekhor Ahmed** (University of British Columbia) **Chuyi Zheng** (University of British Columbia)
Assessing Interactive Open Access Self-Assessment Tool to Enhance Statistical Learning Capacity / Évaluation d'outil d'auto-évaluation interactif et libre accès pour augmenter les capacités d'apprentissage en statistique

---

17:00-18:00 **Invited / Sur invitation** (abstract/résumé 254) **RB 2200**

NSERC – Updates and 2023 Competition Results
Mises à jour du CRSNG et résultats du concours 2023

Chair/Président: W. John Braun
Organizer/Responsable: W. John Braun

17:00-18:00 **Adele Ngi-Song** (NSERC) **Rachel Desrochers** (Natural Sciences and Engineering Research Council of Canada) **David Haziza** (University of Ottawa)
NSERC – Updates and 2023 Competition Results / Mises à jour du CRSNG et résultats du concours 2023

Wednesday May 31 — mercredi 31 mai

08:30-09:50 **Invited / Sur invitation** (abstract/résumé 255) **RB 2200**

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

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

08:30-09:50 **Zhou Zhou** (University of Toronto)
Time-frequency analysis, a non-stationary time series perspective / L'analyse temps-fréquence, une perspective de séries temporelles non stationnaires

---

10:20-11:50 **Invited / Sur invitation** (abstract/résumé 256) **AT 102**

Securing Team-Based and Industrial Grants
Obtenir des subventions d'équipe et des subventions industrielles

Chair/Président: Linglong Kong
Organizer/Responsable: Mateen Shaikh
Sponsor/Commanditaires: Committee on New Investigators/Comité des nouveaux chercheurs

10:20-11:50 **Jim V. Zidek** (The University of British Columbia) **Celia M. T. Greenwood** (McGill University)
Securing Team-Based and Industrial Grants / Obtenir des subventions d'équipe et des subventions industrielles

---

10:20-11:50 **Invited / Sur invitation** (abstract/résumé 257) **AT 101**

Advances in Bayesian Modelling and Computation
Progrès en modélisation et calcul bayésiens

Chair/Président: V. Radu Craiu
Organizer/Responsable: V. Radu Craiu

10:20-10:50 **Mylène Bédard** (Université de Montréal) **Gabriel Boisvert-Beaudry** (Université de Montréal)
On a Generalized, Balanced Version of MALA / À propos d'une version généralisée et équilibrée de MALA

10:50-11:20 **Morris Greenberg** (University of Toronto) **Radu Craiu** (University of Toronto) **Kieran R Campbell** (University of Toronto)
Restricted Search Space MCMC with Adaptive Weighting and Sparsity Parameterization for Graph Inference / Méthode de Monte-Carlo par chaînes de Markov à espace de recherche restreint avec pondération adaptative et paramétrage de la parcimonie pour l'inférence de graphes

11:20-11:50 **Daniel M. Roy** (University of Toronto)
Admissibility is Bayes Optimality with Infintesimals / L'admissibilité est l'optimalité de Bayes avec des infinitésimaux

---

10:20-11:50 **Invited / Sur invitation** (abstract/résumé 259) **AT 301**

Measuring Statistical Evidence
Mesure de la preuve statistique

Chair/Président: Éric P. Marchand
Organizer/Responsable: Éric P. Marchand

10:20-10:50 **Michael J. Evans** (University of Toronto)
The Role of Frequentism in Bayesian Inference / Le rôle du fréquentisme dans l'inférence bayésienne

10:50-11:20 **Minge Xie** (Rutgers University)
Repro Samples Method for Uncertainty Quantification in Irregular Inference Problems and More / Méthode des échantillons repro pour la quantification de l'incertitude dans les problèmes d'inférence irréguliers et autres

11:20-11:50 **Éric P. Marchand** (Université de Sherbrooke)
On the behaviour of Bayesian credible intervals for some restricted parameter space problems / Sur le comportement d'intervalles de crédibilité bayésiens en présence de contraintes paramétriques

---

10:20-11:50 **Invited / Sur invitation** (abstract/résumé 261) **TB 238**

Mentorship for Statisticians
Mentorat pour statisticiens et statisticiennes

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

10:20-11:50 **Thérèse A. Stukel** (ICES/ University of Toronto) **Erica E. M. Moodie** (McGill University) **Béatrice Baribeau** (Statistics Canada) **Jennifer McNichol** (University of Victoria) **Kathryn Mills** (Canada Revenue Agency)
Mentorship for Statisticians / Mentorat pour statisticiens

10:20-11:50 **Invited / Sur invitation** (abstract/résumé 262) **RB 2200**

CJS Award Address
Allocution du récipiendaire du Prix de la RCS

Chair/Président: Andrei Volodin
Organizer/Responsable: Andrei Volodin

10:20-11:50 **Evan Sidrow** (The University of British Columbia)
Modelling multi-scale, state-switching functional data with hidden Markov models / Modélisation de données fonctionnelles multi-échelles et à changement d'état à l'aide de modèles de Markov cachés

---

10:20-11:50 **Contributed / Communications libres** (abstract/résumé 263) **TB 208**

Advances in Distribution Theory and Asymptotic Methods
Progrès en théorie de la distribution et méthodes asymptotiques

Chair/Président: Mohamed Belalia

10:20-10:35 **Hélène Guérin** (Université du Québec à Montréal)
Asymptotic distribution of the Elephant Random Walk / Distribution asymptotique de la marche aléatoire de l'éléphant

10:35-10:50 **Camila P. E. de Souza** (University of Western Ontario) **Pedro H. T. O. Souza** (Universidade de Campinas) **Ronaldo Dias** (Universidade de Campinas)
Bayesian adaptive selection of basis functions for functional data representation / Sélection adaptative bayésienne de fonctions de base pour la représentation de données fonctionnelles

10:50-11:05 **Klaus Peter Herrmann** (Université de Sherbrooke) **Marius Hofert** (The University of Hong Kong) **Johanna G. Nešlehová** (McGill University)
On a Class of Distortions that Transform GEV Distributions into GEV Distributions / D'une classe de distorsions qui transforment les distributions GEV en distributions GEV

11:05-11:20 **Serge B. Provost** (The University of Western Ontario)
Certain Results Pertaining to Distributional Moments / Certains résultats relatifs aux moments de distribution

11:20-11:35 **Zilin Wang** (Wilfrid Laurier University) **Mary Thompson** (University of Waterloo)
Modelling Missing-Not-At-Random in Complex Surveys / Modélisation de données omises de façon non aléatoire dans des enquêtes complexes

11:35-11:50 **Bailey Trenton Janeczko** (University of Manitoba) **Yuliya V. Martsynyuk** (University of Manitoba)
Functional CLT Based Confidence Intervals for the Slope in Linear Structural Errors-in-Variables Models / Intervalles de confiance basés sur le théorème central limite fonctionnel (FCLT) pour la pente dans les modèles structurels linéaires d'erreurs de variables

---

10:20-11:50 **Contributed / Communications libres** (abstract/résumé 267) **TB 210**

Novel Methods in complex data
Nouvelles méthodes pour les données complexes

Chair/Président: Nan Zheng

10:20-10:35 **Simon J. Bonner** (Western University) **Braedan Walker** (Western University) **Ron Bonner**
Incorporating Isotope Patterns in the Analysis of Mass-Spectrometry Data / Intégration des motifs isotopiques dans l'analyse des données de spectrométrie de masse

10:35-10:50 **Hao He** (University of Ottawa) **David Haziza** (University of Ottawa) **Song Cai** (Carleton University)
Efficient and Robust Empirical Likelihood Approach for Estimating Equations with Missing Data / Approche efficace et robuste de la vraisemblance empirique pour l'estimation d'équations avec données manquantes

10:50-11:05 **Shaun McDonald** (Simon Fraser University) **Dave Campbell** (Carleton University)
FRODO: A Novel Approach to Micro-Macro Multilevel Regression / FRODO : Une nouvelle approche de la régression micro-macro multiniveaux

11:05-11:20 **Jonathan Jalbert** (Polytechnique Montreal) **Gabriel Gobeil** (Polytechnique Montréal) **Philippe Roy** (Institut de recherche d'Hydro-Québec)
Post-processing the bulk and the tail of simulated precipitations / Post-traitement du coeur et de la queue des précipitations simulées

11:20-11:35 **Jeffrey D. Picka** (University of New Brunswick)
Is scientific realism compatible with statistical inference? / Le réalisme scientifique est-il compatible avec l'inférence statistique ?

11:35-11:50 **Paritosh Kumar Roy** (McGill University) **Alexandra Schmidt** (McGill University) **Thaís C. O. Fonseca** (Federal University of Rio de Janeiro, Brazil)
Dynamical Non-Gaussian Modeling of Multivariate Spatial Processes / Modélisation dynamique non gaussienne de processus spatiaux multivariés

---

10:20-11:50 **Contributed / Communications libres** (abstract/résumé 270) **TB 240**

Advances in Discriminant, Classification and Learning Methods
Progrès dans les méthodes de discrimination, de classification et d'apprentissage

Chair/Président: Suborna Shekhor Ahmed

10:20-10:35 **Liyuan Zheng** (University of Waterloo) **Yilei Wu** (Pinterest) **Yingli Qin** (University of Waterloo) **Mu Zhu** (University of Waterloo)
Quadratic Discriminant Analysis for High-Dimensional Data / Analyse discriminante quadratique pour les données de grande dimension

10:35-10:50 **Nicholas Mitsakakis** (Children's Hospital of Eastern Ontario Research Institute) **Samer El Kababji** (Children's Hospital of Eastern Ontario Research Institute) **Khaled El Emam** (Children's Hospital of Eastern Ontario Research Institute)
Determining Adequate Sample Size for Studies Using Machine Learning / Détermination de la taille adéquate de l'échantillon pour les études utilisant l'apprentissage automatique

10:50-11:05 **Steve Ferreira Guerra** (McGill University) **Steve Ferreira Guerra** (McGill University) **Robert Platt** (McGill University) **Michal Abrahamowicz** (McGill University)
SIMEX Adaptation to Correct for Misclassification in Binary Time-Varying Prescription-Based Exposures / SIMEX adapté pour corriger les erreurs de classification dans des expositions binaires basées sur des ordonnances et qui varient dans le temps

11:05-11:20 **Yanglei Song** (Queen's University) **Na Li** (Queen's University) **Chunfang Devon Lin** (Queen's University) **Dongsheng Tu** (Queen's University)
Bootstrap Adjusted Predictive Classification under Generalized Linear Models / Classification prédictive ajustée par bootstrap selon des modèles linéaires généralisés

11:20-11:35 **Qingrun Zhang** (University of Calgary) **Jingni He** (University of Calgary) **Qing Li** (University of Calgary) **Qingrun Zhang** (University of Calgary)
rvTWAS: identifying rare variants underlying complex traits using a data-bridge model / rvTWAS : identification de variants rares sous-jacents à des traits complexes au moyen d'un modèle de pont de données

---

13:30-15:00 **Invited / Sur invitation** (abstract/résumé 273) **AT 301**

The modeling approaches for the proportional/count data with extra variations
Approches de modélisation pour les données proportionnelles/de comptage avec variations supplémentaires

Chair/Président: Dianliang Deng
Organizer/Responsable: Dianliang Deng

13:30-14:00 **Cindy Feng** (Dalhousie University)
Zero-Inflated Models for Adjusting Varying Exposures: A Cautionary Note on the Pitfalls of Using Offset / Modèles à surreprésentation de zéros pour l'ajustement d'expositions variant avec le temps ; mise en garde contre les pièges d'une utilisation d'un "offset"

14:00-14:30 **Guohua Yan** (University of New Brunswick) **Xiaolei Zhang** (Yunnan Normal University) **Renjun Ma** (University of New Brunswick) **Jiaxiu Li** (University of New Brunswick)
Analysis of longitudinal binomial data with positive association between the number of successes and the number of failures: an application to stock instability study / Analyse de données binomiales longitudinales avec association positive entre le nombre de réussites et le nombre d'échecs : une application à l'étude de l'instabilité des stocks

14:30-15:00 **Xiaoqing Zhang** (University of Regina) **Dianliang Deng** (University of Regina)
The Generalized Lindley Binomial Model for Analyzing the Proportions with End-point Inflation / Le modèle binomial généralisé de Lindley pour l'analyse de proportions avec inflation terminale

---

13:30-15:00 **Invited / Sur invitation** (abstract/résumé 275) **AT 102**

Modelling heterogeneity in complex high-dimensional data
Modélisation de l'hétérogénéité dans les données complexes à haute dimension

Chair/Président: Abbas Khalili
Organizer/Responsable: Abbas Khalili

13:30-14:00 **Tudor A. Manole** (Carnegie Mellon University) **Patrick Bryant** (Carnegie Mellon University) **John Alison** (Carnegie Mellon University) **Mikael Kuusela** (Carnegie Mellon University) **Larry Wasserman** (Carnegie Mellon University)
Homogeneity Testing in the Search for Pairs of Higgs Bosons at the Large Hadron Collider / Test d'homogénéité à la recherche de paires des bosons de Higgs dans le Grand collisionneur de hadrons

14:00-14:30 **Pengqi Liu** (McGill University)
Regularization in Finite Mixture of Sparse GLMs with Ultra-High Dimensionality / Régularisation dans les mélanges finis de GLM épars avec une dimensionnalité ultra-élevée

---

13:30-15:00 **Invited / Sur invitation** (abstract/résumé 277) **AT 101**

Isobel Loutit Invited Address
Allocution Isobel-Loutit

Chair/Président: Reza Ramezan
Organizer/Responsable: Reza Ramezan
Sponsor/Commanditaires: Business and Industrial Statistics Section (BISS)/Business and Industrial Statistics Section (BISS)

13:30-15:00 **Derek Bingham** (Simon Fraser University)
Fast Emulation and Modular Calibration for Simulators with Functional Response / Émulation rapide et étalonnage modulaire pour simulateurs à réponse fonctionnelle

---

13:30-15:00 **Invited / Sur invitation** (abstract/résumé 278) **RB 2200**

Pierre Robillard Invited Address
Allocution du récipiendaire du Prix Pierre-Robillard

Chair/Président: Yingwei (Paul) Peng
Organizer/Responsable: Yingwei (Paul) Peng

13:30-15:00 **Saifuddin Syed** (The University of British Columbia)
Scalable Bayesian inference with non-reversible parallel tempering / Inférence bayésienne évolutive avec atténuation parallèle non réversible

---

13:30-15:00 **Contributed / Communications libres** (abstract/résumé 279) **TB 208**

Methods for Non-Normal and Misclassified Data
Méthodes pour les données non normales et mal classées

Chair/Président: Zelalem Firisa Negeri

13:30-13:45 **Rick E. Danielson** (Fisheries and Oceans Canada)
Modelling Data and Phenomena in Complementary Ways / Modéliser des données et des phénomènes de façons complémentaires

13:45-14:00 **Jean-François Bégin** (Simon Fraser University)
Ensemble Economic Scenario Generators: Unity Makes Strength / Générateurs de scénarios économiques ensemblistes : L'union fait la force

14:00-14:15 **Xintong Li** (University of Waterloo) **Ben Mingbin Feng** (University of Waterloo)
Machine learning proxies for estimating tail risk measures of dynamic hedging errors via nested simulation / Proxys dans l'apprentissage automatique pour l'estimation des mesures de risque de queue des erreurs de couverture dynamique par simulation imbriquée

14:15-14:30 **Ziming Chen** (University of Toronto) **Ewan Goligher** (Toronto General Hospital, University Health Network, Canada) **Eddy Fan** (Toronto General Hospital, University Health Network, Canada) **Anders Granholm** (Copenhagen University Hospitals, Denmark) **Michael Harhay** (Epidemiology and Medicine, University of Pennsylvania, USA) **Danny Mcauley** (Queen's University Belfast, UK) **Martin Urner** (Toronto General Hospital, University Health Network, University of Toronto, Canada) **Christopher Yarnell** (Toronto General Hospital, Canada) **Anna Heath** (The Hospital for Sick Children, University of Toronto, Canada, University College London, London,UK)
Comparing Analysis Methods for Combining Mortality and Ventilator Time in Critical Care Trials: A Simulation Study / Comparaison de méthodes d'analyse pour combiner la mortalité et le temps de respirateur dans les essais en soins intensifs : une étude de simulations

14:30-14:45 **Lin Liu** (University of California, San Diego) **Ruohui Chen** (University of California, San Diego) **Loki Natarajan** (University of California, San Diego)
A Linear Mixed Model Approach for Measurement Error Adjustment / Approche de modèle linéaire mixte pour l'ajustement des erreurs de mesure

14:45-15:00 **Philip J. Schmidt** (University of Waterloo) **Monica Emelko** (University of Waterloo)
Probabilistic Modelling: A Not-so-new Way to Fix Problems with Least Squares Regression of qPCR Standard Curves / Modélisation probabiliste : une pas si nouvelle façon de régler des problèmes avec la régression des moindres carrés de courbes standard à qPCR Ⓔ E

13:30-15:00 **Contributed / Communications libres** (abstract/résumé 283) **TB 210**

Improved Methods for Linear and Non-Linear Models
Méthodes améliorées pour les modèles linéaires et non linéaires

Chair/Président: Denis Talbot

13:30-13:45 **Alwell Oyet** (Memorial University of Newfoundland) **Brajendra Sutradhar** (Memorial University of Newfoundland) **R. Prabhakar Rao** (Sri Sathya Sai Institute of Higher Learning)
Estimation of Multinomial Dynamic Logit Models / Estimation de modèles logit dynamiques multinomiaux Ⓔ E

13:45-14:00 **Michael John Ilagan** (McGill University)
Using latent class analysis to counter trolling of chatbot safety judgments / Utilisation d'une analyse de classes latentes pour neutraliser le trollage des jugements des robots conversationnels en matière de sécurité Ⓔ E

14:00-14:15 **Simon Lévesque** (Université de Sherbrooke) **Félix Camirand Lemyre** (Université de Sherbrooke) **Jean-François Ethier** (Université de Sherbrooke)
Distributed Inference for Generalized Linear Models / Inférence distribuée pour les modèles linéaires généralisés Ⓔ F

14:15-14:30 **Guanjie Lyu** (University of Windsor) **Mohamed Belalia** (University of Windsor)
Smooth estimation of conditional quantile function using Bernstein polynomials / Estimation lisse de fonction quantile conditionnelle au moyen des polynômes de Bernstein Ⓔ E

14:30-14:45 **Anita Shahrokhian** (Queen's University) **Chunfang Devon Lin** (Queen's University)
Adaptive Designs for Contour Estimation from Computer Experiments with Quantitative and Qualitative Inputs / Devis adaptatifs pour une estimation de contour tirée d'expériences informatiques avec entrées qualitatives et quantitatives Ⓔ E

14:45-15:00 **Fateh Chebana** (Institut national de la recherche scientifique)
Multivariate L-moments and copula, Statistical inference with hydrological applications / L-moments et copules multivariés, inférence statistique et applications hydrologiques Ⓔ E

15:30-17:00 **Invited / Sur invitation** (abstract/résumé 287) **AT 102**

Advances in Clustering
Progrès en regroupement

Chair/Président: Jeffrey L. Andrews
Organizer/Responsable: Paul David McNicholas

15:30-15:52 **Mackenzie Neal** (McMaster University) **Paul David McNicholas** (McMaster University)
Variable Selection for Skewed Clustering and Classification / Sélection de variables pour regroupement asymétrique et classification Ⓔ E

15:52-16:15 **Michael Gallaugher** (Baylor University) **Eman Alamer** (McMaster University) **Paul David McNicholas** (McMaster University)
Mixture Modeling of Skewed Mixed-Type Data / Modélisation de mélanges de données de type mixte asymétriques Ⓔ E

16:15-16:38 **Brian Franczak** (MacEwan University)
Recent developments in using mixtures of multivariate asymmetric distributions for classification / Avancées récentes dans l'utilisation de mélanges de distributions asymétriques multivariées pour la classification Ⓔ E

16:38-17:00 **Paul David McNicholas** (McMaster University)
Some Current Challenges in Clustering / Certains défis actuels relatifs au regroupement

---

15:30-17:00 **Invited / Sur invitation** (abstract/résumé 289) **AT 301**

How should we conduct sensitivity analysis when making inferences that have policy implications? Learning from our colleagues in health economics.
Comment mener une analyse de sensibilité lorsque l'on effectue des inférences qui ont des implications politiques ? Leçons de nos collègues en économie de la santé

Chair/Président: Kuan Liu
Organizer/Responsable: Eleanor M. Pullenayegum

15:30-15:30 **Eleanor M. Pullenayegum** (Hospital for Sick Children) **Petros Pechlivanoglou** (The Hospital for Sick Children)
Roundtable: How should we conduct sensitivity analysis when making inferences that have policy implications? Learning from our colleagues in health economics. / Table ronde : comment réaliser une analyse de sensibilité lorsque l'élaboration d'inférences a des incidences politiques ? Apprendre de nos collègues en économie de la santé.

15:30-17:00 **Invited / Sur invitation** (abstract/résumé 290) **AT 101**

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

Chair/Président: Wen Wilson Lu
Organizer/Responsable: Wen Wilson Lu
Sponsor/Commanditaires: Survey Methods/Survey Methods

15:30-17:00 **J. Michael Brick** (Westat Inc) **Jill M. DeMatteis** (Westat)
Efficient Sample Designs for Multimode Surveys / Des plans efficaces pour des enquêtes multimodes

---

15:30-17:00 **Contributed / Communications libres** (abstract/résumé 291) **TB 238**

Modeling Time-to-Event and survival analysis
Modélisation de la durée de vie et analyse de survie

Chair/Président: Yifan Li

15:30-15:45 **Yun-hee Choi** (Western University) **Seungwoo Lee** (Western University) **Laurent Briollais** (Lunenfeld-Tanenbaum Research Institute; Dalla Lana School of Public Health, University of Toronto)
A Correlated Competing-risks Model with Time-varying Covariates: An Application to Hereditary Breast and Ovarian Cancer Families / Modèle de risques concurrents corrélés avec covariables variables dans le temps : application aux familles touchées par les cancers héréditaires du sein et de l'ovaire

15:45-16:00 **Ayesha Madhushani Rathnayake Nayaka Bandaralage** (Memorial University of Newfoundland) **Ashokan Mulayath Variyath** (Memorial University Newfoundland) **Zhaozhi Fan** (Memorial University Newfoundland)
Weighted Accelerated Failure Time Model / Modèle de temps de défaillance accéléré pondéré

16:00-16:15 **Rob Deardon** (University of Calgary)
Population-dependent time-varying transmission rates in epidemic models / Taux de transmission variant dans le temps dépendant de la population dans les modèles d'épidémie

16:15-16:30 **Chinmoy Roy Rahul** (University of Calgary) **Rob Deardon** (University of Calgary)
Semi-parametric spatial infectious disease modeling / Modélisation spatiale semi-paramétrique pour les maladies infectieuses

16:30-16:45 **Sayantee Jana** (Indian Institute of Technology Hyderabad) **Sze Fu** (Dalla Lana School of Public Health, Centre for Global Health Research, St. Michael's Hospital, University of Toronto) **Helen Gellband** (Dalla Lana School of Public Health, Centre for Global Health Research, St. Michael's Hospital, University of Toronto) **Prabhat Jha** (Dalla Lana School of Public Health, Centre for Global Health Research, St. Michael's Hospital, University of Toronto) **Patrick Brown** (Dalla Lana School of Public Health, Centre for Global Health Research, St. Michael's Hospital, University of Toronto)
Spatio-temporal modelling of malaria mortality in India from the Million Death Study / Modélisation spatio-temporelle de la mortalité causée par le paludisme en Inde à partir de l'étude Million Death Study Ⓔ E

16:45-17:00 **Wenling Zhang** (University of Waterloo) **Cecilia A. Cotton** (University of Waterloo)
Causal Inference for Recurrent Events via Aggregated Marginal Odds Ratio / Inférence causale pour les événements récurrents par l'entremise du rapport de cotes marginal agrégé Ⓔ E

15:30-17:00 **Contributed / Communications libres** (abstract/résumé 295) **TB 210**

Recent Bayesian Inference and Change-Point Methods
Méthodes récentes en inférence bayésienne et points de rupture

Chair/Président: Guanjie Lyu

15:30-15:45 **Félix Camirand Lemyre** (Université de Sherbrooke) **Jean-François Quessy** (Université du Québec à Trois-Rivières)
Change-Point Tests and Estimators for Gradually Changing Dependence Structures Based on Kendall's Tau / Procédures inférentielles pour la détection de changements graduels au sein de structures de dépendance, basées sur le tau de Kendall Ⓔ E

15:45-16:00 **Ana Carolina da Cruz** (Western University)
Bayesian Clustering Approach Based on Constant-Wise Change-Point Models / Approche bayésienne de regroupement basée sur des modèles de points de changement constants Ⓔ E

16:00-16:15 **Junwei Shen** (McGill University) **Erica E. M. Moodie** (McGill University) **Shirin Golchi** (McGill University)
Privacy-Preservation Through Two-Stage Bayesian Meta-Analysis of Individualized Treatment Rules / Préservation de la confidentialité grâce à une méta-analyse bayésienne en deux étapes des règles de traitement individualisé Ⓔ E

16:15-16:30 **Chengqian Xian** (Western University) **Camila P. E. de Souza** (University of Western Ontario) **Wenqing He** (University of Western Ontario) **Felipe F. Rodrigues** (King's University College at Western University) **Renfang Tian** (King's University College at Western University)
Variational Bayes Inference of Survival Data using Log-logistic Accelerated Failure Time Model / Inférence bayésienne variationnelle sur les données de survie à l'aide d'un modèle log-logistique de temps de défaillance accéléré Ⓔ E

16:30-16:45 **Leif Erik Lovblom** (University of Toronto) **Laurent Briollais** (University of Toronto) **Bruce Perkins** (University of Toronto) **George Tomlinson** (University of Toronto)
Bayesian Estimation of a Joint Model for Multiple Longitudinal Outcomes and a Multistate Process under a Mixed Observation Scheme with Interval-Censoring / Estimation bayésienne d'un modèle conjoint pour des résultats longitudinaux multiples et un processus multi-états dans le cadre d'un schéma d'observation mixte avec censure par intervalles Ⓔ E

16:45-17:00 **Thierry Chekouo** (University of Minnesota) **Xiaotian Dai** (University of Calgary) **Lu Xuewen** (University of Calgary)
A Bayesian variable selection approach incorporating prior feature ordering and population structures / Une approache bayésienne de sélection de variables qui incorpore a priori l'ordre des variables et la structure de la population Ⓔ E

SSC Presidential Invited Address
Allocution de l'invité du président de la SSC

Chair/Président: Bruno N. Rémillard

Organizer/Responsable: Bruno N. Rémillard

Room/Salle: RB 2200

Date: Monday May 29 / lundi 29 mai

Time/Heure: 08:30-09:50

Abstract/Résumé

[08:30-09:50]

Pierre Del Moral (INRIA, Bordeaux Research Center)

Some Theoretical Aspects of Particle Filters and Ensemble Kalman Filters

Quelques aspects théoriques des filtres à particules et des filtres de Kalman d'ensemble

In the last three decades, Particle Filters (PF) and Ensemble Kalman Filters (EnKF) have become one of the main numerical techniques in data assimilation, Bayesian statistical inference and nonlinear filtering. Both particle algorithms can be viewed as mean field type particle interpretation of the filtering equation and the Kalman recursion. In contrast with conventional particle filters, the EnKF is defined by a system of particles evolving as the signal in some state space with an interaction function that depends on the sample covariance matrices of the system. Despite widespread usage, little is known about the mathematical foundations of EnKF. Most of the literature on EnKF amounts to design different classes of useable observer-type particle methods. To design any type of consistent and meaningful filter, it is crucial to understand their mathematical foundations and their learning/tracking capabilities. This talk discusses some theoretical aspects of these numerical techniques. We present some recent advances on the stability properties of these filters. We also initiate a comparison between these particle samplers and discuss some open research questions.

Au cours des trois dernières décennies, les filtres de particules (PF) et les filtres d'ensemble de Kalman (EnKF) sont devenus l'une des principales techniques numériques d'assimilation de données, d'inférence statistique bayésienne et de filtrage non linéaire. Les deux algorithmes de particules peuvent être vus comme une interprétation particulaire de type champ moyen de l'équation de filtrage et de la récursivité de Kalman. Contrairement aux filtres à particules conventionnels, l'EnKF est défini par un système de particules évoluant comme le signal dans un espace d'état avec une fonction d'interaction qui dépend des matrices de covariance d'échantillon du système. Malgré une utilisation répandue, on sait peu de choses sur les fondements mathématiques d'EnKF. La plupart de la littérature sur l'EnKF revient à concevoir différentes classes de méthodes particulaires utilisables de type observateur. Pour concevoir tout type de filtre cohérent et significatif, il est crucial de comprendre leurs fondements mathématiques et leurs capacités d'apprentissage/de suivi. Cet exposé aborde certains aspects théoriques de ces techniques numériques. Nous présentons quelques avancées récentes sur les propriétés de stabilité de ces filtres. Nous initions également une comparaison entre ces échantillonneurs de particules et discutons de certaines questions de recherche ouvertes.

Advances and applications in optimal dynamic treatment regimes
Progrès et applications dans les régimes de traitement dynamiques optimaux

Chair/Président: Mireille Schnitzer

Organizer/Responsable: Mireille Schnitzer, CONG JIANG

Room/Salle: AT 101

Date: Monday May 29 / lundi 29 mai

Time/Heure: 10:20-11:50

Abstract/Résumé

[10:20-10:42]

Lu Wang (University of Michigan)

Evidence-based Estimation of the Optimal Dynamic Personalized Health Care Decision Rules with Restrictions

Estimation fondée sur des données probantes des règles de décision dynamiques optimales en matière de soins de santé personnalisés avec restrictions

We present recent advances and statistical developments for evaluating Dynamic Treatment Regimes (DTR), which allow the treatment to be dynamically tailored according to evolving subject-level data. Identification of an optimal DTR is a key component for precision medicine and personalized health care. We develop a tree-based doubly robust reinforcement learning (T-RL) method, and a new Stochastic-Tree Search method, ST-RL, for evaluating optimal DTRs, which contributes to the existing literature in its non-greedy policy search and demonstrates outstanding performances. We further develop a Restricted Tree-based Reinforcement Learning (RT-RL) method that searches for an interpretable DTR to maximize the expected outcome, given a (set of) user-specified restriction(s), which specifies treatment options that ought not to be considered as part of the estimated tree-based DTR. The method is illustrated using an observational dataset to estimate a two-stage stepped-up DTR for guiding the level of care placement for adolescents with substance use disorder.

Nous présentons les avancées récentes et les développements statistiques pour l'évaluation des régimes de traitement dynamiques (DTR), qui permettent d'adapter dynamiquement le traitement en fonction de l'évolution des données au niveau du sujet. L'identification d'un DTR optimal est un élément clé de la médecine de précision et des soins de santé personnalisés. Nous développons une méthode d'apprentissage par renforcement doublement robuste basée sur les arbres (T-RL), ainsi qu'une nouvelle méthode de recherche par arbre stochastique, ST-RL, pour évaluer les DTR optimaux, qui contribue à la littérature existante par sa recherche de politique non avide et démontre des performances exceptionnelles. Nous développons également une méthode d'apprentissage par renforcement basée sur un arbre restreint (RT-RL) qui recherche un DTR interprétable pour maximiser le résultat attendu, compte tenu d'une (série de) restriction(s) spécifiée(s) par l'utilisateur, qui précise(nt) les options de traitement qui ne devraient pas être prises en compte dans le cadre du DTR basé sur l'arbre estimé. Nous illustrons la méthode à l'aide d'un ensemble de données d'observation afin d'estimer un DTR progressif en deux étapes pour guider le placement au niveau des soins des adolescents souffrant de troubles liés à l'utilisation de substances psychoactives.

[10:42-11:05]

Aaron Sarvet (EPFL)

Optimal Regimes for Algorithm-Assisted Human Decision-Making

Régimes optimaux pour la prise de décision humaine assistée par algorithme

Foundational work on causal inference and dynamic treatment regimes presents a promising pathway towards precision medicine. In a precision-medicine system, decision rules might be algorithmically individual-

Les travaux fondamentaux sur l'inférence causale et les régimes de traitement dynamiques constituent une voie prometteuse vers la médecine de précision. Dans un système de médecine de précision, les règles de décision pourraient être algorithmiquement

ized based on an optimal rule previously learned from non-experimental or experimental data. However, there is some resistance to the notion that implementation of an optimal regime, successfully learned from the data, will result in better expected outcomes on average, compared to existing human-decision rules: care providers may be inclined to override the treatment recommendations provided by the identified optimal regimes, based on their privileged patient observations. In this talk, I will review existing methodology for learning optimal regimes and clarify the validity of the care provider's skepticism. Then, I will present methodology for leveraging human intuition by identifying a super-optimal regime using data generated by either nonexperimental or experimental studies, and clarify when a fusion of such data is beneficial. The superoptimal regime will indicate to a care provider – in an algorithm-assisted decision setting – precisely when expected outcomes would be maximized if the care provider would override the optimal regime recommendation and, importantly, when the optimal regime recommendation should be followed regardless of the care-provider's assessment.

individualisées sur la base d'une règle optimale préalablement apprise à partir de données non expérimentales ou expérimentales. Cependant, l'idée que la mise en œuvre d'un régime optimal, appris avec succès à partir des données, se traduira par de meilleurs résultats escomptés en moyenne, par rapport aux règles de décision humaines existantes, se heurte à une certaine résistance : les prestataires de soins peuvent être enclins à passer outre les recommandations de traitement fournies par les régimes optimaux identifiés, sur la base de leurs observations privilégiées des patients. Dans cet exposé, je passerai en revue la méthodologie existante pour l'apprentissage des régimes optimaux et je clarifierai la validité du scepticisme du prestataire de soins. Je présenterai ensuite une méthodologie permettant de tirer parti de l'intuition humaine en identifiant un régime super-optimal à l'aide de données générées par des études non expérimentales ou expérimentales, et je préciserai quand la fusion de ces données est bénéfique. Le régime super-optimal indiquera à un prestataire de soins - dans un contexte de décision assistée par algorithme - précisément quand les résultats attendus seraient maximisés si le prestataire de soins passait outre la recommandation du régime optimal et, surtout, quand la recommandation du régime optimal devrait être suivie quelle que soit l'évaluation du prestataire de soins.

[11:05-11:27]

Cong Jiang (University of Montreal) **Mary Thompson** (University of Waterloo) **Michael Wallace** (University of Waterloo)

Estimating DTR with household interference for ordinal outcome

Estimation de régimes de traitement dynamiques (DTR) avec interférences ménagères pour des réponses ordinales

The focus of precision medicine is on decision support, often in the form of dynamic treatment regimes (DTRs), which are sequences of decision rules. At each decision point, the decision rules determine the next treatment based on the patient's characteristics, previous treatment and response, and current status and tests. DTR estimation with ordinal outcomes is rarely studied, and rarer still in the context of interference - where one patient's treatment may affect another's outcome. In this talk, I will introduce our proposed robust method — the dynamic weighted proportional odds model (dWPOM) for ordinal outcomes. Moreover, in the presence of household interference, considering the possible correlation between treatments in the same household, I will present the covariate balancing weights, which rely on the joint propensity score, and methods for estimating that score. Lastly, I will illustrate dWPOM in the analysis of the Population Assessment of Tobacco and Health data.

La médecine de précision est axée sur l'aide à la décision, souvent sous la forme de régimes de traitement dynamiques (DTR) qui sont des séquences de règles de décision. À chaque point de décision, ces règles déterminent le prochain traitement en fonction des caractéristiques du patient, du traitement et de la réponse antérieures de même que de l'état actuel du patient et de tests. L'estimation d'un DTR avec des réponses ordinales est rarement étudiée et plus rarement encore dans un contexte d'interférence – lorsque le traitement d'un patient peut influer sur les réponses d'un autre. Je présente la méthode robuste que nous proposons – le modèle de cotes proportionnelles pondérées dynamiques (dWPOM) pour des réponses ordinales. De plus, en présence d'une interférence ménagère, et prenant en compte la corrélation possible entre des traitements dans un même ménage, je présente des poids d'équilibrage des covariables qui dépendent du score de propension conjoint et des méthodes pour l'estimation de ce score. Pour terminer, j'illustre le dWPOM en analysant les données de l'étude Population Assessment of Tobacco and Health.

[11:27-11:50]

Janie Coulombe (Université de Montréal) **Erica E. M. Moodie** (McGill University) **Susan M. Shortreed** (Kaiser Permanente Washington) **Christel Renoux** (Lady Davis Institute for Clinical Research)

Using multiple imputations of longitudinal data and dynamic weighted survival modeling to choose an optimal antidepressant drug

Le recours à l'imputation multiple de données longitudinales et à la modélisation de survie dynamique afin d'identifier un médicament antidépresseur optimal

We created a cohort of new users of antidepressant drugs between 2008 and 2018 using observational data from Kaiser Permanente Washington. The data contain information on patients' demographics, anthropometric measurements, medication use, and patient health questionnaires (PHQ). The PHQ is a depression severity score ranging from 0 to 27, with 27 the most severe. Our goal is to find tailoring variables for the choice of an antidepressant drug based on patient characteristics to minimize patients' PHQ score in the first year of follow-up. Since PHQ measurements are not always available at cohort entry and they are measured sporadically during follow-up, we propose a sequential approach to multiple imputations to recover the baseline and monthly PHQ values. Dynamic weighted survival modeling is further used to develop an adaptive treatment strategy for the choice of antidepressant drug. The results are discussed and compared with those from a previous study using United Kingdom's data.

Nous avons créé une cohorte de nouveaux utilisateurs d'antidépresseurs entre 2008 et 2018 à partir de données de Kaiser Permanente Washington. Les données contiennent l'information sur les variables démographiques, anthropométriques, la médication et le questionnaire de santé du patient (QSP). Le QSP est un score de sévérité de la dépression qui varie de 0 à 27 avec 27 le plus sévère. Notre but est de trouver des variables adaptives pour le choix d'un antidépresseur en se basant sur les caractéristiques des patients pour minimiser le QSP des patients dans la première année. Comme le QSP est mesuré de façon sporadique, nous proposons une approche séquentielle d'imputation multiple pour remplacer les valeurs manquantes du QSP à chaque mois. La modélisation de survie dynamique pondérée est utilisée pour développer la règle de traitement adaptive servant à choisir un médicament antidépresseur. Les résultats sont discutés et comparés à ceux d'une étude utilisant les données du Royaume-Uni.

Maximizing learning capacity in a joyful environment in Statistical Education
Maximiser la capacité d'apprentissage dans un environnement joyeux dans l'éducation en statistique

Chair/Président: Suborna Shekhor Ahmed

Organizer/Responsable: Suborna Shekhor Ahmed

Room/Salle: AT 301

Date: Monday May 29 / lundi 29 mai

Time/Heure: 10:20-11:50

Abstract/Résumé

[10:20-10:50]

Bruce Dunham (The University of British Columbia)

Implementing evidence-based pedagogy: When they don't know what's good for them

Implantation d'une pédagogie basée sur la preuve : lorsqu'ils ne savent pas ce qui est bon pour eux

Research and personal experience indicate one of the challenges in implementing many of the approaches known to be effective for learning (such as repeated testing, group work, distributed practice, and in-class activities) is that they are unpopular with some students. The main issue is that such approaches are effortful for students and may appear, superficially at least, less efficient than being told information. Described are tactics to overcome student objections, including the sharing of evidence about the methods adopted, information about student performances, and arguments to address concerns.

Les recherches et l'expérience personnelle indiquent que l'un des défis d'intégration de nombreuses approches reconnues pour être efficaces pour apprendre (comme le test à répétition, le travail en équipe, la pratique distribuée et les activités en classe) est qu'elles sont impopulaires auprès de certains étudiants. Le problème principal est que de telles approches demandent beaucoup d'efforts des étudiants et peuvent sembler (superficiellement, du moins) moins efficaces que de simplement recevoir de l'information. Nous décrivons des tactiques pour surmonter l'objection étudiante, y compris le partage de preuve concernant les méthodes adoptées, des renseignements sur les performances étudiantes et des arguments pour aborder les inquiétudes.

[10:50-11:20]

Douglas Whitaker (Mount Saint Vincent University) **Alana Unfried** (California State University, Monterey Bay) **Leyla Batakci** (Elizabethtown College) **Marjorie E. Bond** (Pennsylvania State University) **April Kerby-Helm** (Winona State University) **Michael A. Posner** (Villanova University)

Measuring student enjoyment and effective teaching practices: what works?

Mesure de l'enthousiasme des étudiants et pratiques d'enseignement efficaces : qu'est-ce qui fonctionne ?

Knowledge is not the only important outcome from statistics courses: students' attitudes and beliefs about statistics have been widely recognized as important outcomes since the 1990s. Within the statistics education research literature, there has been a proliferation of work using surveys to measure student attitudes (e.g., their interest/enjoyment). However, with most widely used research tools for measuring attitudes about statistics, finding evidence of change in attitudes from precourse to post-course has been a recurring challenge. A new family of instruments - the Surveys of Motivational Attitudes toward Statistics (SOMAS) - has been devel-

La connaissance n'est pas le seul aboutissement important des cours de statistique ; les attitudes et convictions des étudiants au sujet de la statistique ont été largement reconnues comme des résultats importants depuis les années 1990. Dans la documentation de recherche sur l'enseignement de la statistique, on trouve une abondance de travaux utilisant les enquêtes pour mesurer les attitudes des étudiants (par ex. : leur intérêt/enthousiasme). Cependant, avec la plupart des outils de recherche largement utilisés pour mesurer les attitudes à l'égard de la statistique, la découverte d'une preuve d'un changement des attitudes entre les précours et postcours a été un problème récurrent. Une nouvelle famille d'instruments – l'Enquête sur les attitudes motivationnelles en-

oped to measure student attitudes, instructor attitudes and practices, and characteristics of the learning environment to provide researchers with a robust view of statistics attitudes within a course. With an eye toward 'what works', this presentation provides an overview of these tools along with results from a large pilot implementation.

vers les statistiques (SOMAS) (Surveys of Motivational Attitudes toward Statistics) – a été développée pour mesurer les attitudes des étudiants, les attitudes et pratiques des enseignants et les caractéristiques de l'environnement d'apprentissage afin de fournir aux chercheurs une vue statistique robuste des attitudes pendant un cours. Dans une perspective de « ce qui fonctionne », cette présentation fournit une vue d'ensemble de ces outils de même que des résultats d'une vaste implémentation pilote.

[11:20-11:50]

Mike Marin (The University of British Columbia)

Creating Educational Videos: Advice From A Decade of Experience

Création de vidéo éducatif : conseil basé sur dix ans d'expérience

With online and blended learning becoming more and more common, many of us are creating (or wanting to create) video lectures to supplement our classes. While the idea of creating a 'video' may be daunting at first, there are actually many easy entry-points into this. As academics, we have many demands on our time, and so we want to get the most for the least. In this talk I will present some low-barrier ways to get started, with a focus on getting pretty good quality videos without consuming too much of our valuable time. This talk is aimed at those with less experience (or no experience) creating video lectures, although those with lots of experience may also gain a few helpful tips. Most videos I have created can be previewed at www.youtube.com/marinstatlectures for those interested in seeing a sample of what I have done.

L'apprentissage mixte et en ligne devient de plus en plus courant. C'est pourquoi beaucoup d'entre nous créent (ou tentent de créer) des leçons sous forme vidéo pour les ajouter aux cours. L'idée de créer un «vidéo» peut sembler intimidante au début, mais il existe beaucoup de moyens d'introduction accessibles pour y parvenir. En tant qu'enseignants universitaires, nous sommes grandement sollicités. C'est pourquoi il faut optimiser l'utilisation de son temps. Lors de cet exposé, je présenterai quelques façons accessibles pour commencer, en me concentrant sur la bonne qualité des vidéos tout en évitant de gaspiller du temps précieux. Cette présentation est destinée à ceux ayant peu à pas d'expérience en création de leçons vidéo, même si ceux ayant de l'expérience pourraient bénéficier de certains conseils judicieux. Pour ceux souhaitant voir des échantillons de mon travail, vous trouverez des aperçus de la plupart des vidéos que j'ai créés sur www.youtube.com/marinstatlectures.

Chair/Président: Priscilla E. Greenwood

Organizer/Responsable: Priscilla E. Greenwood

Room/Salle: TB 202

Date: Monday May 29 / lundi 29 mai

Time/Heure: 10:20-11:50

Abstract/Résumé

[10:20-10:50]

Priscilla E. Greenwood (The University of British Columbia)

Desertification: Stochastic Models and Statistical Challenges

Désertification : modèles stochastiques et problèmes statistiques

Climate change is producing extreme weather events including serious droughts. The threat of formation and growth of deserts has serious consequences for world food supply. For example Iran has more than 75% arid and semiarid areas which are fragile and susceptible to desertification. Farmers in Colorado, Kansas, and several other states of the midwest USA are overexploiting the aquifer beneath. In this talk I will describe some stochastic models which have been introduced in response to this threat, preliminary results which increase understanding of the dynamics involved, and some remaining challenges.

Les changements climatiques entraînent des événements climatiques extrêmes, y compris de graves sécheresses. La formation et l'expansion des déserts est une menace qui a de sérieuses conséquences pour l'approvisionnement alimentaire mondial. Par exemple, 75 % de la superficie de l'Iran est formé de zones arides et semi-arides fragiles et susceptibles de désertification. Les fermiers du Colorado, du Kansas et de plusieurs autres États du Midwest américain surexploitent l'aquifère souterrain. Notre exposé décrit certains modèles stochastiques qui ont été proposés en réponse à cette menace, fournit des résultats préliminaires qui permettent de mieux comprendre la dynamique à l'œuvre et fait état de certains problèmes encore irrésolus.

[10:50-11:20]

Luis F. Gordillo (Utah State University) **Priscilla E. Greenwood** (University of British Columbia)

Intermittent Precipitation-Dependent Interactions, Encompassing Allee Effect, May Yield Vegetation Patterns in a Transitional Parameter Range

Des interactions intermittentes dépendant des précipitations, englobant l'effet Allee, peuvent donner lieu à des schémas de végétation dans une plage de paramètres transitoire

We present a spatial model that incorporates Allee-type and competition interactions for vegetation as an evolving random field of biomass density. The cumulative effect of close-range precipitation-dependent interactions is controlled by a parameter defining precipitation frequency. We identify a narrow parameter range in which the transition from persistence to extinction via an aggregation pattern is likely to occur. The aggregation pattern is tied to the initial configuration and arises differently from Turing's diffusion and differential flow patterns of other models. There is close agreement of our

Nous présentons un modèle spatial qui incorpore les interactions de la végétation de type Allee et de compétition comme un champ aléatoire évolutif de la densité de la biomasse. L'effet cumulatif des interactions à courte distance dépendant des précipitations est contrôlé par un paramètre définissant la fréquence des précipitations. Nous identifions une gamme étroite de paramètres dans laquelle la transition de la persistance à l'extinction via un modèle d'agrégation est susceptible de se produire. Le modèle d'agrégation est lié à la configuration initiale et se présente différemment des modèles de diffusion de Turing et de flux différentiel d'autres modèles. Notre plage de paramètres de transition critique concorde

critical transition parameter range with that of the corresponding evolving random mean-field model.

étroitement avec celle du modèle de champ moyen aléatoire évolutif correspondant.

[11:20-11:50]

Chunyi Gai (The University of British Columbia) **Yuxin Chen Theodore Kolokolnikov** (Dalhousie University) **Justin Tzou** (Macquarie University)

Patterned Vegetation, Tipping Points and the Rate of Climate Change

Végétation schématique, point de bascule et vitesse du changement climatique

When faced with slowly depleting resources, complex ecological systems are prone to sudden irreversible changes as the resource level dips below a tipping point. A possible coping mechanism is the formation of spatial patterns, which allows for concentration of sparse resources and the survival of the species within "ecological niches" even below the tipping point of the homogeneous vegetation state. However, if the change in resource availability is too sudden, the system may not have time to transition to the patterned state and will pass through the tipping point instead, leading to extinction. We argue that the deciding factors are the speed of resource depletion and the amount of background noise (seasonal climate changes) in the system. We illustrate this phenomenon on a model of patterned vegetation. Our analysis underscores the importance of, and the interplay between, the speed of climate change, heterogeneity of the environment, and the amount of seasonal variability.

Lorsque confrontés au lent épuisement des ressources, les systèmes écologiques complexes ont tendance à subir des changements rapides et irréversibles au fur et à mesure que le niveau de ressource plonge sous un point de bascule. Un des mécanismes d'adaptation possible serait la formation de schémas spatiaux, ce qui permettrait la concentration de ressources éparses et la survie des espèces à l'intérieur de ≪niches écologiques≫ même si l'on franchit le point de bascule de l'état de végétation homogène. Par contre, si l'épuisement de ressource progresse trop rapidement, le système risque de ne pas avoir le temps de réaliser sa transition vers un état schématique et dépassera plutôt le point de bascule, menant à son extinction. Nous soutenons que les facteurs déterminants sont la vitesse d'épuisement de ressource et la quantité de bruit de fond (changements climatiques saisonniers) dans le système. Nous illustrons ce phénomène sur un modèle de végétation schématique. Notre analyse souligne l'importance de la vitesse du changement climatique, l'hétérogénéité de l'environnement et la quantité de variabilité saisonnière, ainsi que les interactions entre ces facteurs.

Data Science in Industry: A Showcase
Science des données dans l'industrie : vitrine

Chair/Président: Matthew Greenberg

Organizer/Responsable: Matthew Greenberg

Room/Salle: AT 102

Date: Monday May 29 / lundi 29 mai

Time/Heure: 10:20-11:50

Abstract/Résumé

[10:20-10:50]

Jennifer Zhu (Amazon)

Data Science in Industry: A Showcase

Présentation de la science des données dans l'industrie

More and more companies are making machine learning (ML) an important part of their operations, but the path to ML success is not always straightforward. To foster more innovation and development in this area, data scientists work backwards and deliver a roadmap of ML use cases and their implementation. In this talk, I will introduce how we apply ML to real-life customer use cases, to achieve personalized in-vehicle thermal comfort and intelligent financial document understanding, respectively.

De plus en plus d'entreprises adoptent de façon importante l'apprentissage automatique (ML) dans leurs opérations, mais la voie menant au succès de la ML n'est pas toujours linéaire. Dans le but de générer plus d'innovation et de développement dans ce domaine, les scientifiques des données travaillent rétroactivement afin de fournir une feuille de route concernant les cas d'utilisation de la ML et de son implémentation. Lors de cet exposé, j'expliquerai comment appliquer la ML à des cas d'utilisation pour des clients réels, et comment obtenir un confort thermique personnalisé en voiture et une compréhension intelligente de documents financiers, respectivement.

[10:50-11:20]

Tingzhou Yu (University of Victoria)

Data Science and Optimization: An Industry Perspective

Point de vue de l'industrie sur la science et l'optimisation des données

Gurobi Optimization is a leader in the mathematical programming solver market and views optimization as a complementary tool to data science to extend predictive analytics to prescriptive. First, we'll see how machine learning and optimization worked together in industry to solve complex problems through two case studies. Next, we'll explore how Gurobi is integrating machine learning and optimization through two open-source Python packages. One is a wrapper library to connect Gurobi's modeling API to Pandas, which is the de-facto standard in Python for data manipulation and analysis. The second is an experimental package that allows users to add trained regressors from packages such as scikit-learn, TensorFlow/Keras, and PyTorch as constraints to a Gurobi model. Lastly, we will delve how

Gurobi Optimization est un chef de file sur le marché des solveurs de programmation mathématique et considère l'optimisation comme un outil complémentaire à la science des données pour étendre l'analyse prédictive à l'analyse prescriptive. Tout d'abord, nous présenterons deux études de cas illustrant la manière dont l'apprentissage automatique et l'optimisation ont permis de résoudre des problèmes complexes dans l'industrie. Ensuite, nous examinerons comment Gurobi intègre l'apprentissage automatique et l'optimisation au moyen de deux paquets Python à code source ouvert. L'une d'entre elles est une bibliothèque d'intégration permettant de connecter l'API de modélisation de Gurobi à Pandas, qui est la norme de facto de Python pour la manipulation et l'analyse des données. Le second est un paquet expérimental qui permet aux utilisateurs d'ajouter des régresseurs formés à partir de paquets tels que scikit-learn, TensorFlow Keras

people in academia and industry can add mathematical optimization to their skillset using Gurobi's learning material which include numerous modeling examples, videos, and games.

et PyTorch en tant que contraintes à un modèle Gurobi. Enfin, nous verrons la façon dont les universitaires et les industriels peuvent ajouter l'optimisation mathématique à leurs compétences avec le matériel d'apprentissage de Gurobi, qui comprend de nombreux exemples de modélisation, vidéos et jeux.

Stochastic Optimization in Insurance and Finance
Optimisation stochastique en assurance et finance

Chair/Président: Jean-François Renaud

Organizer/Responsable: Jean-François Renaud

Room/Salle: TB 208

Date: Monday May 29 / lundi 29 mai

Time/Heure: 10:20-11:50

Abstract/Résumé

[10:20-10:50]

Bin Li (University of Waterloo)

Equilibrium Pricing of Variable Annuities within a Principal-Agent Framework

Tarification d'équilibre des rentes variables dans un cadre mandant-mandataire

We propose a new equilibrium pricing approach of variable annuities within a principal-agent framework where an insurer (principal) is the contract provider and a policyholder (agent) is the follower having the surrender option. While the risk-neutral pricing approach adopted in the existing literature leads to significantly higher fees and more frequent surrendering than market observations, this new equilibrium pricing approach reconciles the misalignment between theoretical results and market observations. We also find that a surrender penalty or a lower insurance fee can make the insurer's expected profit more robust.

Nous proposons une nouvelle approche de tarification d'équilibre des rentes variables dans un cadre mandant-mandataire où l'assureur (le mandant) est prestataire d'un contrat, et un titulaire de contrat (le mandataire) est le demandeur ayant un droit d'abandon. Contrairement à l'approche de tarification à risque neutre adoptée dans la documentation actuelle qui mène à des frais considérablement élevés et des abandons plus fréquents que les observations du marché, notre nouvelle approche de tarification d'équilibre concilie les décalages entre les résultats théoriques et les observations du marché. Nous avons aussi découvert qu'une pénalité d'abandon ou des frais d'assurance réduits peuvent rendre les profits de l'assureur plus robustes.

[10:50-11:20]

Alexandre Roch (UQAM) **Simone Scotti** (Université Paris Diderot)

Optimal Ratcheting Dividends Policy with Resets

Politique optimale de dividendes de type cliquet avec réinitialisation

We study an optimal dividend problem under ratcheting and bankruptcy constraints. Firms face a tradeoff between the negative signal of a dividend reduction and a potential future bankruptcy. In contrast to previous works, the dividend rate is allowed to change negatively after incurring a reputation cost in accordance with a large empirical literature on the topic. We provide an explicit sufficient and necessary condition for the finiteness of the value function and prove the differentiability of the value function, and its characterization via viscosity solutions and the so-called super-contact property. The value function is given by the solution of a two dimensional PDE combining optimal stopping, singular and impulse controls. We show that the problem is

Nous étudions un problème de politiques de dividendes optimales sous des contraintes de type "cliquet". L'entreprise fait face à un compromis entre le signal négatif d'une réduction de dividende et une éventuelle faillite. Contrairement aux travaux précédents, le taux de dividende peut diminuer après avoir subi un coût réputationnel, de façon cohérente à une large littérature empirique sur le sujet. Nous obtenons une condition explicite qui soit nécessaire et suffisante sous laquelle la fonction de valeur est finie et démontrons la différentiabilité de la fonction de valeur et sa caractérisation via des solutions de viscosité et la propriété de type "super-contact". La fonction de valeur est donnée par la solution d'une équation différentielle bidimensionnelle combinant des temps d'arrêts optimaux, des contrôles singuliers et des contrôles de type impulsionnel. Nous démontrons

equivalent to a one-dimensional setup simplifying both mathematical and financial analysis. Finally, we detail the numerical study of the model and we provide a complete sensitivity analysis with respect to the model parameters.

que le problème est équivalent à un réduction unidimensionnelle simplifiant à la fois les analyses mathématique et financière. Enfin, nous complétons l'étude numérique du modèle et nous fournissons une analyse complète de la sensibilité par rapport aux paramètres du modèle.

[11:20-11:50]

Abel Cadenillas (University of Alberta) **Wenyue Liu** (University of Alberta)

Optimal Insurance Contracts for a Shot-Noise Cox Claim Process

Contrats d'assurance optimaux pour un processus de réclamations Cox à bruit quantique

We consider a continuous-time model in which an insurer proposes an insurance contract to a potential insured. Motivated by climate change and catastrophic events, we assume that the number of claims process is a shot-noise Cox process. The insurer selects the premium to be paid by the potential insured and the amount to be paid for each claim. In addition, the insurer can request some actions from the potential insured to reduce the number of claims. The insurer wants to maximize his expected total utility, while the potential insured signs the contract if his expected total utility for signing the contract is greater than or equal to his expected total utility when he does not sign the contract. We obtain an analytical solution for the optimal premium, the optimal amount to be paid for each claim, and the optimal actions of the insured.[Joint work with Wenyue Liu].

Nous présentons un modèle en temps continu dans lequel un assureur propose un contrat d'assurance à un assuré potentiel. Motivés par les changements climatiques et les événements catastrophiques, nous supposons que le processus du nombre de réclamations est un processus de Cox à bruit quantique. L'assureur sélectionne la prime à payer par l'assuré potentiel et le montant à payer pour chaque réclamation. De plus, l'assureur peut demander certaines actions à l'assuré potentiel pour réduire le nombre de réclamations. L'assureur veut maximiser son utilité totale espérée, tandis que l'assuré potentiel signe le contrat si son utilité totale espérée en signant le contrat est supérieure ou égale à son utilité totale espérée s'il ne le signe pas. Nous obtenons une solution analytique pour la prime optimale, le montant optimal à payer pour chaque réclamation et les actions optimales de l'assuré. Ce travail est fait en collaboration avec Wenyue Liu.

Chair/Président: Hugh Chipman

Room/Salle: TB 236

Date: Monday May 29 / lundi 29 mai

Time/Heure: 10:20-11:50

Abstract/Résumé

[10:20-10:35]

Larry Dong (University of Toronto: Dalla Lana School of Public Health) **David W. Johnson** (Department of Pediatrics, Faculty of Medicine, University of Calgary, Alberta Children's Hospital, Calgary, Canada) **Terry P. Klassen** (Manitoba Institute of Child Health, Winnipeg, Canada) **Anna Heath** (Child Evaluative Services, the Hospital for Sick Children, Toronto, Canada; Division of Biostatistics, Dalla Lana School of Public Health, University of Toronto, Toronto, Canada; Department of Statistical Science, University College London, London, UK) **Amy Plint** (CHEO Research Institute, Ottawa, Canada; Department of Emergency Medicine, University of Ottawa, Ottawa, Canada)

The Probability of Reducing Hospitalisation Rates for Bronchiolitis: A Bayesian Analysis

La probabilité de réduire les taux d'hospitalisation pour la bronchiolite : une analyse bayésienne

CanBEST is a randomised controlled trial that aimed to assess the effectiveness of epinephrine, dexamethasone and their combination (EPIDEX) in reducing 7-day hospitalization in infants due to bronchiolitis. Statistical significance was reached in the initial un-adjusted analysis but not after adjustment (p = 0.07). Hence, we conducted a Bayesian post-hoc analysis of CanBEST to see additional insights can be raised. We used Bayesian relative risk (RR) regression with a range of reference and data-driven priors for treatment effects. The Bayesian analysis showed that the 95% credible interval (CrI) for the EPIDEX RR has an upper bound smaller than 1 for all but the strongly skeptical priors. Using minimally informative priors provides comparable results to the frequentist analysis (RR = 0.67, 95% CrI = (0.45, 0.96)). This post-hoc analysis provides a framework for a personalized interpretation of the trial data by explicitly incorporating prior beliefs into the statistical modelling.

CanBEST est un essai contrôlé randomisé visant à évaluer l'efficacité de l'épinéphrine, de la dexaméthasone et de leur combinaison (EPIDEX) dans la réduction de l'hospitalisation de 7 jours chez les nourrissons atteints de bronchiolite. Un résultat statistiquement significatif a été atteint dans l'analyse non ajustée initiale, mais pas après l'ajustement (p = 0,07). Nous avons donc mené une analyse post-hoc bayésienne de CanBEST pour voir si on peut en extraire de nouvelles conclusions. Nous avons utilisé une régression de risque relatif (RR) bayésienne avec plusieurs distributions préalables dites "de référence" et basées sur des données antérieures pour les effets du traitement. L'analyse bayésienne a montré que l'intervalle de crédibilité à 95% (CrI) pour le RR EPIDEX a une limite supérieure de moins de 1 pour toutes les distributions préalables sauf une. L'utilisation de celle dite peu informative fournit des résultats comparables à l'analyse fréquentiste (RR = 0,67, CrI à 95% = (0,45, 0,96)). Cette analyse post-hoc fournit un cadre pour une interprétation personnalisée des résultats en incorporant explicitement les croyances antérieures dans la modélisation statistique.

[10:35-10:50]

Fatemeh Mahmoudi (University of Calgary) **Xuewen Lu** (University of Calgary)

Bi-Level Selection with the Sparse Group BAR for Multivariate Interval-Censored Data

Sélection à deux niveaux avec la régression ridge adaptative brisée (BAR) par groupe parcimonieux pour des données multivariées à intervalles censurés

Multivariate interval-censored data arise when individuals in the study are at the risk of experiencing multiple events of interest and the onset time of each event

Les données multivariées à intervalles censurés se produisent lorsque les participants à une étude sont à risque de connaître de multiples événements d'intérêt et que le temps d'apparition

is not observed precisely, but it is known to lie in a time window between two examinations. Regularized regression is a popular approach in the framework of variable selection problems. In this regard, we propose a new composite penalty to incorporate the oracle property and grouping effect property of the broken adaptive ridge (BAR) regression and a group bridge (GB) penalty, in contrast to the existing sparse group lasso, we term it as the sparse group BAR. This leads to a new bi-level variable selection method that selects variables at both individual and group levels. We propose a novel EM-based algorithm that encompasses data augmentation and iterative least square approximation to solve the optimization problem. Simulation studies and data analysis are conducted to demonstrate the performance of the proposed method.

de chaque événement n'est pas précisément observé, mais est connu pour arriver dans une fenêtre de temps entre deux examens. La régression régularisée est une approche populaire dans le contexte des problèmes de sélection de variables. À cet égard, nous proposons une nouvelle pénalité composite ; celle-ci incorpore la propriété oracle et la propriété des effets de regroupement de la régression ridge adaptative brisée (BAR) et une pénalité de pont de groupes (GB). Afin de différencier notre méthode de la méthode déjà existante du Group-lasso parcimonieux, nous l'appelons BAR par groupe parcimonieux. Nous arrivons ainsi à une nouvelle méthode de sélection à deux niveaux qui choisit des variables à la fois au niveau individuel et de groupe. Nous proposons également un nouvel algorithme EM qui englobe l'augmentation des données et une approximation par moindre carré itérative pour résoudre le problème d'optimisation. Des études par simulation et une analyse de données sont menées pour illustrer la performance de la méthode proposée.

[10:50-11:05]

Zixuan Zhao (Queen's University) **Yanglei Song** (Queen's University) **Wenyu Jiang** (Queen's University) **Dongsheng Tu** (Queen's University)

Consistent covariances estimation for stratum imbalances under marginal design for covariate adaptive randomization

Estimation convergente des covariances pour les déséquilibres de strate sous plan marginal pour la randomisation adaptative selon les covariables

Marginal design, also known as the minimization method, is a popular approach for covariate adaptive randomization when the number of strata is large in a clinical trial. Valid statistical inference with data collected under marginal design requires the knowledge of the limiting covariance matrix of within-stratum imbalances, whose existence is only recently established. In this work, we propose a bootstrap-based estimator for this limit and establish its consistency, in particular, by Le Cam's third lemma. As an application, we consider in simulation studies adjustments to existing robust tests for treatment effects with survival data by the proposed estimator. It shows that the adjusted tests achieve a size close to the nominal level, and unlike other designs, the robust tests without adjustment may have an asymptotic size inflation issue under the marginal design.

Le plan marginal, aussi appelé méthode de minimisation, est une approche populaire utilisée pour la randomisation adaptative aux covariables lorsqu'un essai clinique comporte un grand nombre de strates. Une inférence statistique valide avec des données recueillies dans le cadre d'un plan marginal requiert la connaissance de la matrice de covariance limite des déséquilibres intra-strates, cette limite ayant été établie récemment. Nous proposons un estimateur bootstrap pour cette limite et établissons sa convergence, en particulier par le troisième lemme de Le Cam. Sur le plan de l'application, nous considérons dans les études de simulation des ajustements par l'estimateur proposé aux tests robustes existants pour les effets du traitement avec des données de survie. La taille des tests ainsi ajustés approche le niveau nominal, et contrairement à d'autres plans, les tests robustes sans ajustement peuvent avoir un problème d'inflation asymptotique de taille sous le plan marginal.

[11:05-11:35]

Raphael Robert McDonald (Dalhousie University) **David Keith** (Fisheries and Oceans Canada) **Jessica Sameoto** (Fisheries and Oceans Canada) **Joanna Elizabeth Mills Flemming** (Dalhousie University)

Integrating habitat features into spatio-temporal biomass dynamics models for a better understanding of stock productivity: A case study of Sea Scallop in the Bay of Fundy

L'intégration de différents habitats à l'intérieur d'un modèle spatio-temporel de dynamique de la biomasse pour une meilleure compréhension de la productivité des stocks : Étude de cas sur le pétoncle de mer dans la Baie de Fundy

Spatio-temporal modelling frameworks have recently been developed to support the stock assessment of harvested marine species. Improved by harnessing state-space approaches, cutting-edge computational methods and geostatistics, these models have greatly improved our ability to provide reliable science advice for fisheries management. While some environmental covariates are often incorporated within these models, there is significant interest in incorporating other key biophysical characteristics such as habitat type. Here, we modify a recently developed spatio-temporal stock assessment model for sea scallops (Placopecten magellanicus) to explicitly incorporate the effect(s) of different habitats. Specifically, we allow for habitat-specific probabilities of zeroes through a combination of negative binomial and zero-inflated negative binomial distributions, and we also allow for habitat-specific catchability parameters. These catchability parameters are informed by sediment size within habitat categories. This approach provides more realistic estimates of population abundance, thereby improving the science advice provided for the purpose of managing the sea scallop population. Our approach is highly applicable to other stocks in spatially variable environments.

Plusieurs modèles spatio-temporels ont récemment été développés pour supporter l'évaluation des stocks d'espèces marines exploitées. Grâce aux développements d'approches espace-état, des méthodes informatiques de pointe et le domaine de la géostatistique, ces modèles ont grandement amélioré notre capacité de fournir un avis scientifique de haute qualité pour la gestion des pêches. Malgré le fait que des covariables environnementales sont souvent incluses à l'intérieur de ces modèles, il y a beaucoup d'intérêt pour l'inclusion de plusieurs autres caractéristiques biophysiques importantes tel que le type d'habitat. Nous modifions un modèle de gestion des stocks spatio-temporel récemment développé pour le pétoncle de mer (Placopecten magellanicus) pour inclure l'effet(s) de différents habitats sur leur abondance. Nous obtenons différentes probabilités d'obtenir un zéro d'un habitat à l'autre grâce à une combinaison de distributions binomiale négative et binomiale négative avec surreprésentation de zéros, ainsi que différentes capturabilités par habitat. Ces paramètres de capturabilité sont informés par la taille des sédiments à l'intérieur des différents types d'habitats. Cette nouvelle approche obtient des estimés d'abondance plus réalistes, améliorant par le fait-même les avis scientifiques qui sont fournis à partir de ce genre de modèle pour gérer la population de pétoncles de mer.

Chair/Président: Zhiyang Zhou

Room/Salle: TB 240

Date: Monday May 29 / lundi 29 mai

Time/Heure: 10:20-11:50

Abstract/Résumé

[10:20-10:35]

Wei Liang (University of Waterloo) **Changbao Wu** (University of Waterloo)

Propensity Score Weighting with Post-Treatment Survey Data

Pondération du score de propension avec des données d'enquête post-traitement

The propensity score (PS) weighting approach is widely used in observational studies to mitigate confounding bias when there are no unmeasured confounders. However, individualized treatment effects may vary across units, leading to issues with the generalizability of results from specific study samples to broader populations. In this paper, we consider drawing population-level causality from the post-treatment survey data, which has the feature that the treatment stage is prior to survey sampling. We propose a general framework for estimating the PS and making inferences about the population-level treatment effects in the presence of known sampling weights. When the survey sample is non-probability, we propose to estimate the sampling weights by combining multiple reference probability survey samples. We evaluate the proposed approach through simulation studies and a real case study of racial disparities in medical expenditure, demonstrating its finite-sample performance.

La pondération du score de propension (PS) est une approche largement utilisée dans les études observationnelles pour mitiger le biais de confusion lorsqu'il y a absence de facteurs de confusion non mesurés. Cependant, les effets de traitements individualisés peuvent varier dans l'ensemble des unités, ce qui entraîne des problèmes de généralisation des résultats d'une étude d'échantillons spécifiques à des populations plus larges. Dans cet article, nous tirons des conclusions de causalité au niveau de la population à partir de données d'enquête post-traitement, c'est-à-dire des données qui se caractérisent par une phase de traitement préalable à l'échantillonnage de l'enquête. Nous proposons un cadre général pour l'estimation du score de propension et faisons des inférences sur les effets du traitement au niveau de la population en présence de poids d'échantillonnage connus. Lorsque l'échantillon de l'étude est non probabiliste, nous proposons une estimation des poids de l'échantillonnage en combinant des échantillons d'étude probabilistes de multiples populations de référence. Nous évaluons aussi l'approche proposée à l'aide d'études de simulation et d'une étude de cas réelle des disparités raciales au chapitre des dépenses médicales, démontrant sa performance avec pour une taille d'échantillon finie.

[10:35-10:50]

Jia Ning Zhang (University of Ottawa) **David Haziza** (University of Ottawa) **Sixia Chen** (University of Oklahoma)

Robust imputation in the presence of influential units in surveys

Imputation robuste en présence d'unités influentes dans les enquêtes

Item nonresponse in surveys is often treated by some form of single imputation. In some cases, one faces the problem of influential units in the sample. This problem is especially acute in business surveys that collect economic variables whose distributions are highly skewed. In the presence of influential units, the classical imputed estimators are approximately unbiased if the first mo-

Dans les enquêtes, une non-réponse partielle est souvent traitée par une certaine forme d'imputation unique. Dans certains cas, on fait face au problème d'unités influentes dans l'échantillon. C'est un problème particulièrement aigu dans les enquêtes d'ordre commercial qui collectent des variables économiques dont les distributions sont hautement asymétriques. En présence d'unités influentes, les estimateurs imputés classiques sont approximative-

ment of the imputation model is correctly specified but may be very unstable. Therefore, it is desirable to develop a robust imputation method that produces biased but more stable imputed estimators, that is an estimator whose mean square error is smaller than that of the corresponding non-robust imputed estimator. To achieve the goal, we propose to use a robust regression based on the Huber function and an adaptative tuning constant. A simulation study is conducted to compare the empirical performance of the proposed method with non-robust methods for a wide class of distributions.

ment sans biais si le premier moment du modèle d'imputation est correctement spécifié mais ils peuvent être très instables. Par conséquent, il est désirable de développer une méthode d'imputation robuste qui produit des estimateurs imputés biaisés mais plus stables, c.-à-d. des estimateurs dont l'erreur quadratique moyenne est plus petite que celle des estimateurs imputés non robustes correspondants. Pour atteindre ce but, nous proposons d'utiliser une régression robuste basée sur la fonction de Huber et une constante de réglage adaptatif. Une étude par simulation est menée pour comparer la performance empirique de la méthode proposée à celle de méthodes non robustes pour une grande classe de distributions.

[10:50-11:05]

Caren Hasler (Université de Neuchâtel) **Esther Eustache** (Université de Neuchâtel)

Quasi-Model-Assisted Estimators under Nonresponse in Sample Surveys

Estimateurs quasi-assistés par modèle pour non-réponse dans les enquêtes

In the presence of auxiliary information, model-assisted estimators use a working model that links the variable of interest and the auxiliary variables in order to improve the Horvitz-Thompson estimator. The resulting estimators are asymptotically design unbiased and asymptotically more efficient than the Horvitz-Thompson estimator under some regularity conditions and for a wide range of working models. In this work, we adapt model-assisted total estimators to missing at random data building on the idea of nonresponse weighting adjustment. We consider nonresponse as a second phase of the survey and reweight the units in model-assisted estimators using the inverse of estimated response probabilities in order to compensate for the nonrespondents. We develop the asymptotic properties of our proposed estimators, study their double robustness, and discuss calibration of the resulting weights. We provide formulae for asymptotic variance and variance estimators. We conduct a simulation study that describes the behavior of the proposed estimators.

En présence d'information auxiliaire, les estimateurs assistés par modèle utilisent un modèle de travail mettant en correspondance la variable d'intérêt et les variables auxiliaires, ce qui permet d'améliorer l'estimateur d'Horvitz-Thompson. Les estimateurs ainsi obtenus sont asymptotiquement sans biais et plus efficaces que l'estimateur d'Horvitz-Thompson sous certaines conditions de régularité et ce pour un large éventail de modèles de travail. Dans ce travail, nous adaptons la technique assistée par modèle au contexte de la non-réponse aléatoire en utilisant la repondération. Nous voyons la non-réponse comme une deuxième phase d'échantillonnage et repondérons les répondants dans les estimateurs assistés par modèle au moyen de l'inverse de probabilités de réponse estimées. Nous décrivons les propriétés asymptotiques des estimateurs proposés, présentons leur double robustesse et montrons des cas dans lesquels les poids obtenus sont calés. Nous étudions la variance et l'estimation de variance des estimateurs proposés. Nous présentons les résultats d'une étude par simulation.

[11:05-11:20]

Beili Huang (University of Manitoba)

Statistical Models for Multilevel Data with "Don't know" category: Implications for program evaluation

Modèles statistiques pour des données multiniveaux avec la catégorie «ne sait pas» : conséquences sur l'évaluation de programme

"Don't know (DK)" category has been increasingly used in surveys of longitudinal research for incorporation of EDI (Equality, Diversity, and Inclusion) principles. Though reducing non-responses and guesswork to some extent, it creates unique challenges in program evalua-

On se sert de plus en plus du choix de réponse «ne sait pas» (NSP) dans les enquêtes de recherche longitudinales pour incorporer les principes d'EDI (égalité, diversité et inclusion). Elle réduit la non-réponse et la conjecture dans une certaine mesure, mais génère des défis uniques dans l'évaluation de programme. Dans le cadre

tion. In this study, a two-part mixed effects model was developed for multilevel longitudinal data with DKs to take both DKs and data correlations into account, with one part describing DK probability over time, the other part modeling the evolving patterns in longitudinal outcomes. The model also reveals the associations between DKs and outcomes by correlated random effects. Simulation studies showed that the proposed model outperformed other methods frequently used which consider DKs as missing under different DK mechanisms. The advantages of the proposed model in program evaluation were also illustrated with a real-world example PROJECT11 (a prevention program targeting Manitoba's Education outcomes).

de cette étude, nous avons conçu un modèle d'effets mixtes en deux parties pour les données longitudinales multiniveaux avec NSP pour tenir compte des corrélations de données et de NSP. L'une de ces parties décrit la probabilité de NSP dans le temps, tandis que l'autre modélise les schémas évolutifs des résultats longitudinaux. Le modèle révèle aussi les associations entre les NSP et les résultats par ses effets aléatoires corrélés. Des études de simulations démontrent que le modèle proposé surpasse d'autres méthodes couramment utilisées qui considèrent les NSP comme manquants selon différents mécanismes de NSP. Nous illustrons les avantages du modèle proposé en évaluation de programme avec un exemple tiré du monde réel : PROJECT11 (un programme de prévention ciblant les résultats en éducation du Manitoba).

Chair/Président: Simon J. Bonner

Room/Salle: TB 238

Date: Monday May 29 / lundi 29 mai

Time/Heure: 10:20-11:50

Abstract/Résumé

[10:20-10:35]

Jervis Gallanosa (University of Manitoba) **Yuliya V. Martsynyuk** (University of Manitoba) **Shivani Bhardwaj** (University of Manitoba)

Confidence Intervals for the Mean Based on the Functional CLT for a Weighted Student Process

Intervalles de confiance pour une moyenne basée sur le théorème de la limite centrale (CLT) fonctionnel dans un processus de Student pondéré

If a random variable X is in the domain of attraction of the normal law (DAN), an appropriately weighted Student process based on independent copies of X follows a functional central limit theorem (FCLT). Accordingly, appropriate functionals of such a process converge in distribution to the same functionals of the likewise weighted standard Wiener process. We use such a convergence for several integral functionals and their modifications, and derive asymptotic confidence intervals (CI) for the mean of X. We inspect the finite-sample coverage probabilities and expected lengths of our CI's and argue that they may be desirable alternatives to a CI based simply on the asymptotic normality of the Student t-statistic, especially for distributions with larger skewness and/or heavier tails in DAN. Moreover, our CI's are seen to be better alternatives to the latter CI than the CI's from the FCLT for the unweighted Student process that were previously obtained in the literature.

Si une variable aléatoire X est dans le domaine d'attraction de la loi normale (DAN), un processus de pondération adéquat de Student basé sur des copies indépendantes de X suit un CLT fonctionnel (FCLT). Par conséquent, les fonctionnelles appropriées d'un tel processus convergent en distribution aux mêmes fonctionnelles que le processus Wiener standard pareillement pondéré. Nous utilisons cette convergence pour plusieurs fonctionnelles intégrales et leurs modifications et dérivons des intervalles de confiance (IC) asymptotiques pour la moyenne de X. Nous vérifions les probabilités de couverture de l'échantillon fini et les longueurs attendues des IC et nous soutenons qu'elles peuvent être des solutions de rechange désirables à un IC simplement basé sur la normalité asymptotique de la statistique-t de Student, en particulier pour des distributions avec une plus grande asymétrie et (ou) des queues plus lourdes dans le DAN. De plus, nos IC sont perçus comme de meilleures solutions de rechange à ce dernier IC que les IC du FCLT pour le processus non pondéré de Student présenté auparavant dans la littérature.

[10:35-10:50]

Armin Hatefi (Memorial University of Newfoundland)

Bayesian Mixture Modeling with Ranked Set Samples

Modélisation de mélanges bayésiens avec plans d'échantillonnage d'ensemble classé

We consider the Bayesian estimation of the parameters of a finite mixture model from independent order statistics arising from imperfect ranked set sampling (RSS) designs. As a cost-effective method, ranked set sampling enables us to incorporate easily attainable characteristics, as ranking information, into data collection and Bayesian estimation. To handle the special structure of the ranked set samples, we develop a Bayesian estima-

Nous considérons l'estimation bayésienne des paramètres d'un modèle de mélange fini de statistiques d'ordre indépendantes découlant de plans d'échantillonnage d'ensemble classé imparfaits (RSS). Comme méthode rentable, le plan d'échantillonnage d'ensemble classé permet d'incorporer facilement des caractéristiques atteignables, comme information de classement, à une collection de données et une estimation bayésienne. Pour traiter la structure spéciale des plans d'échantillonnage d'ensemble classé,

tion approach exploiting the Expectation-Maximization (EM) algorithm in estimating the ranking parameters and Metropolis within Gibbs Sampling to estimate the parameters of the underlying mixture model. Our findings show that the proposed RSS-based Bayesian estimation method outperforms the commonly used Bayesian estimation method using simple random sampling. The developed method is finally applied to analyze the bone disorder status of women aged 50 and older.

nous développons une approche d'estimation bayésienne exploitant l'algorithme espérance-maximisation (EM) pour l'estimation des paramètres d'ordre et l'algorithme de Metropolis avec l'échantillonnage de Gibbs pour estimer les paramètres du modèle de mélange sous-jacent. Nos résultats montrent que la méthode d'estimation bayésienne basée sur le RSS que nous proposons surpasse la méthode d'estimation bayésienne couramment utilisée qui fait appel à un échantillonnage aléatoire simple. La méthode développée est appliquée à l'analyse de l'état des troubles osseux chez des femmes de 50 ans et plus.

[10:50-11:05]

Karim Barigou (Université Laval) **Pierre-Olivier Goffard** (Université de Strasbourg) **Stéphane Loisel** (Université Lyon 1) **Yahia Salhi** (Université Lyon 1)

Bayesian model averaging for mortality forecasting using leave-future-out validation

Modèle moyen bayésien pour prédire la mortalité par validation leave-future-out

Predicting mortality rates plays a key role for life insurance and pension funds. Relying on the prediction of only one mortality model can be too restrictive and can lead to drawbacks, such as model misspecification, parameter uncertainty, and overfitting. To address these issues we first consider mortality modeling in a Bayesian negative-binomial framework to account for overdispersion and parameter uncertainty. Model averaging techniques are then considered as a response to model misspecifications. In this paper, we propose two methods based on leave-future-out validation and compare them to standard Bayesian model averaging (BMA) based on marginal likelihood. An intensive numerical study is carried out to compare the performances of the proposed methodologies. An illustration is then proposed on real-life mortality datasets. Overall, we found that both methods based on an out-of-sample criterion outperform the standard BMA approach in terms of prediction performance and robustness.

Prédire la mortalité joue un rôle majeur pour les assurances vie et fonds de pension. Se baser sur la prédiction d'un seul modèle de mortalité est trop restrictif et mène à des problèmes d'erreur de spécification de modèle, d'incertitude de paramètres et de sur-paramétrisation. Pour résoudre cela, on considère un cadre bayésien binomiale négatif qui tient compte de la sur-dispersion et de l'incertitude sur les paramètres. Ceci est combiné à une approche de modèle moyen pour tenir compte de l'incertitude de modèles. Dans cet article, on propose deux méthodes de modèle moyen basées sur la validation leave-future-out qui sont comparées à l'approche standard de modèle moyen bayésien (BMA). Globalement, on montre que les deux méthodes basées sur un critère hors-échantillon surpassent l'approche standard BMA en termes de performance de prédictions et de robustesse.

[11:05-11:20]

Mufan Li (University of Toronto)

The Neural Covariance SDE: Shaped Infinite Depth-and-Width Networks at Initialization

La covariance neuronale SDE : réseaux à largeur-profondeur infinie modelés à l'initialisation

Recent work has shown that shaping the activation function as network depth grows large is necessary for this covariance matrix to be non-degenerate. However, the current infinite-width-style understanding of this shaping method is unsatisfactory for large depth: infinite-width analyses ignore the microscopic fluctuations from layer to layer, but these fluctuations accumulate over many layers. To overcome this shortcoming, we study

De récentes études ont démontré qu'il est nécessaire de modeler la fonction d'activation au fur et à mesure que la profondeur de réseau grandit pour que cette matrice de covariance soit non-dégénérée. Cependant, la connaissance actuelle du style à largeur infinie de cette méthode de modelage est insuffisante pour une grande profondeur ; c'est-à-dire que les analyses avec largeur infinie ne tiennent pas compte des fluctuations microscopiques d'une couche à l'autre, alors que ces fluctuations s'accumulent avec

the random covariance matrix in the shaped infinite-depth-and-width limit. We identify the precise scaling of the activation function necessary to arrive at a non-trivial limit, and show that the random covariance matrix is governed by a stochastic differential equation (SDE) that we call the Neural Covariance SDE. Using simulations, we show that the SDE closely matches the distribution of the random covariance matrix of finite networks.

chaque couche. Dans le but de corriger ce défaut, nous étudions la matrice de covariance aléatoire dans la limite à largeur et profondeur infinies. Nous identifions l'échelle précise de la fonction d'activation nécessaire pour arriver à une limite non triviale, et démontrons que la matrice de covariance aléatoire est gouvernée par une équation différentielle stochastique (SDE) que nous nommons SDE à covariance neuronale. Au moyen de simulations, nous démontrons que la SDE correspond de près à la distribution de la matrice de covariance de réseaux finis.

[11:20-11:35]

Xu (Sunny) Wang (Wilfrid Laurier University) **Yang Liu** (Wilfrid Laurier University) **Herteg Kohar** (Wilfrid Laurier University) **Doug Brown** (CIBC)

Enhancing Named Entity Recognition with Data Augmentation via Abstractive Summarization

Rehausser la reconnaissance d'entités nommées avec l'augmentation de données par l'entremise de synthèse par abstraction

Named Entity Recognition is an NLP task that requires token-level data for training. In some domains, little training data currently exists. Though some common tag sets and domains have plenty of data compiled by the public custom-named entity tags and niche domains often have little to no data available. Data augmentation is a technique that attempts to alleviate this. Through various approaches, individuals can create artificial data to help increase model performance. This paper examines various sentence-level approaches in low-resource domains wherein artificially produced training data is required for decent model training. We simulate a low-resource scenario for data augmentation, and a new-generation approach using abstractive summarization. In our experiments, we mimic four low-resource scenarios for examining model performance with abstractive summarization as compared to other augmentation methods and the original un-augmented data.

La reconnaissance d'entités nommées est une tâche de traitement automatique des langues (TALN) qui nécessite des données au niveau de l'unité lexicale pour l'éduquer. Dans certains domaines, il existe peu de données d'apprentissage. Bien que certains ensembles d'étiquettes et domaines courants possèdent amplement de données compilées par le public, les étiquettes d'entités nommées personnalisées et des domaines pointus n'offrent presque aucune donnée. L'augmentation de données est une technique qui tente d'atténuer le problème. À travers plusieurs approches, des individus peuvent créer des données artificielles afin d'améliorer la performance d'un modèle. Cette étude examine une variété d'approches au niveau de la phrase dans des domaines ayant peu de ressources et dans lesquels il est nécessaire de produire artificiellement des données d'apprentissage pour éduquer adéquatement un modèle. Nous simulons un scénario à faible ressource pour l'augmentation de données, et une approche nouvelle génération se servant de la synthèse par abstraction. Lors de nos expériences, nous imitons quatre scénarios à faible ressource pour examiner la performance du modèle avec la synthèse par abstraction par rapport aux autres méthodes d'augmentation et les données initiales non augmentées.

[11:35-11:50]

Ankana Dey (Université de Sherbrooke) **F. Guillaume Blanchet** (Université de Sherbrooke)

Modelling the Human Microbiome using Metacommunity Theory

Modélisation du microbiome humain via la théorie des métacommunautés

The human microbiota is composed of a diversity of bacteria, fungi, protists and viruses and it is dynamic; it changes based on what we eat, the medication we take or diseases we may have. In this respect, it is interesting to study how microbial species disperse and change across different parts of our body. To approach this problem,

Le microbiote humain est composé d'une diversité de bactéries, de champignons, protistes et de virus et, en plus il est dynamique ! Il change en fonction de ce que nous mangeons, des médicaments que nous prenons ou des maladies que nous pouvons avoir. À cet égard, il est intéressant d'étudier comment les espèces microbiennes se dispersent et changent dans les différentes parties de

we relied on metacommunity theory, which was originally developed to study macroecological systems but which has been suggested to study microbiomes. In our project we proposed to adapt metacommunity theory for the particularity of human system and thus developed a multivariate Lotka-Volterra Competition Diffusion model. The parameters of this new theoretical model will be estimated through Bayesian modelling using data from the Human Microbiome Project. In this presentation, the theoretical model and its associated statistical model will be presented as well as the procedure used to estimate the parameters, which relied on Markov chain Monte Carlo.

notre corps. Pour aborder ce problème, nous nous sommes appuyés sur la théorie des métacommunautés, qui a été développée à l'origine pour étudier les systèmes macroécologiques mais qui a récemment été suggérée pour étudier les microbiomes. Dans ce projet, nous avons proposé d'adapter la théorie des métacommunautés à la particularité du système humain et avons ainsi développé un modèle multivarié de compétition-diffusion de Lotka-Volterra. Les paramètres de ce nouveau modèle théorique seront estimés via une approche bayésienne en utilisant les données du Human Microbiome Project. Dans cette présentation, le modèle théorique et son modèle statistique associé seront présentés ainsi que la procédure utilisée pour estimer les paramètres, qui repose sur la méthode de Monte Carlo par chaîne de Markov.

Chair/Président: Sudhir Paul

Room/Salle: RB 1200

Date: Monday May 29 / lundi 29 mai

Time/Heure: 10:20-11:50

Abstract/Résumé

[10:20-10:35]

Samantha-Jo Caetano (University of Toronto) **Xiao Wu** (University of Toronto) **Wai Yu Amanda Ng** (University of Toronto) **Quynh Phuong Vu** (University of Toronto)

Development and implementation of a statistics consulting group

Développement et mise en place d'un groupe de consultation statistique

Statistics is ubiquitous to development in research [1]. As such there is a need for statistics and data science support to ensure researchers can suitably design their experiments and can employ appropriate statistics analyses [1]. Over the past year we have developed a "Data Sciences Café" at the University of Toronto [2], where we have a team of statisticians (faculty and students) meet weekly with different researchers who are seeking statistical advice/consultation. In this presentation, our team will present the working model of this consulting group, highlight its successes, discuss any limitations, and propose adaptations for future iterations. 1. Vance, Eric A. "Recent developments and their implications for the future of academic statistical consulting centers." The American Statistician 69.2 (2015): 127-137. 2. "Data Sciences Café - DSI", University of Toronto – Data Sciences Institute. https://datasciences.utoronto.ca/data-sciences-cafe/. Last Access February 16, 2023.

Les statistiques sont omniprésentes dans le développement de recherche [1]. C'est pourquoi il est nécessaire d'offrir du soutien en statistique et en science des données pour permettre aux chercheurs de concevoir adéquatement leurs expériences et employer les analyses statistiques appropriées [1]. Durant la dernière année, nous avons mis en place un ≪café des sciences des données≫ à l'université de Toronto [2], dans lequel nous avons une équipe de statisticiens (enseignants et étudiants) qui rencontre de façon hebdomadaire différents chercheurs en quête de conseils en statistiques. Dans cette présentation, notre équipe présentera le modèle de fonctionnement de ce groupe de consultation, soulignera ses réussites, discutera des limites et proposera des corrections pour les itérations à venir. 1. Vance, Eric A. ≪Recent developments and their implications for the future of academic statistical consulting centers.≫ The American Statistician 69.2 (2015) : 127-137. 2. "Data Sciences Café - DSI", University of Toronto – Data Sciences Institute. https ://datasciences.utoronto.ca/data-sciences-cafe/. Consulté le 16 février 2023.

[10:35-10:50]

Adel Ahmadi Nadi (University of Waterloo) **Stefan Steiner** (University of Waterloo) **Nathaniel T. Stevens** (University of Waterloo)

Assessing the Agreement Between Multi-Operator Measurement Systems Using the Probability of Agreement

Évaluation de la concordance entre systèmes de mesure multi-opérateurs à l'aide de la probabilité de concordance

Quantifying the agreement between different measurement systems (MS) is an important challenge in medical contexts. The probability of agreement (PoA) was recently introduced as an intuitive metric that assesses the agreement between an established MS and a new one. The PoA is simple to apply and easily interpreted and its interpretation does not change no matter how com-

La quantification de la concordance entre différents systèmes de mesure (SM) est un défi important dans les contextes médicaux. La probabilité de concordance (PoA) a été récemment introduite en tant que mesure intuitive qui évalue la concordance entre un système de mesure établi et un nouveau. La PoA est simple à appliquer et à interpréter, et son interprétation ne change pas, quelle que soit la complexité du modèle de données. La méthodologie

plex the data model is. The current PoA methodology is based on a one-factor random effects model that represents measurement variation from all sources with a single effect. However, in clinical analysis, an MS is often used by multiple operators to collect data. In this presentation, we develop the required methodology for applying PoA when the influence of operators is separately accounted for in the measurement error model. The proposed PoA methodology is illustrated on respiratory rates of Chronic Obstructive Pulmonary Disease (COPD) patients to assess the agreement between a gold standard device and a chest-band.

actuelle de la PoA est basée sur un modèle d'effets aléatoires à un facteur qui représente la variation de mesure de toutes les sources par un effet unique. Cependant, dans l'analyse clinique, un même SM est souvent utilisé par plusieurs opérateurs pour collecter des données. Dans cette présentation, nous développons la méthodologie requise pour appliquer la PoA lorsque l'influence des opérateurs est prise en compte séparément dans le modèle d'erreur de mesure. La méthodologie proposée est illustrée sur les taux respiratoires de la broncho-pneumopathie chronique obstructive (BPCO) afin d'évaluer la concordance entre un appareil de référence et une ceinture thoracique.

[10:50-11:05]

Zelalem Firisa Negeri (University of Waterloo)

Identifying and Accommodating Outlying Studies in Diagnostic Test Meta-Analyses: a Mixture Modelling Approach

Identification et prise en compte des études aberrantes dans les méta-analyses de tests diagnostiques : une approche de modélisation des mélanges

Outlying studies are common in meta-analyses of diagnostic test accuracy studies. However, appropriate statistical methods for detecting and down-weighting the effect of such studies appeared in the literature only recently. These methods dichotomize each study in the meta-analysis as outlying or non-outlying and focus on examining the effect of outlying studies on the summary sensitivity and specificity only. In this work, we develop a random-effects bivariate mixture model for meta-analyzing diagnostic test accuracy studies by accounting for both the within- and across-study heterogeneity in diagnostic test results. Instead of dichotomizing the studies in the meta-analysis, the proposed model generates the probability that each study is outlying and allows assessing the impact of outlying studies on the pooled sensitivity, specificity, and between-study heterogeneity. We illustrate the performance of the developed method on real-life and simulated meta-analytic data.

Les études aberrantes sont fréquentes dans les méta-analyses d'études sur la précision des tests de diagnostic. Toutefois, ce n'est que récemment que des méthodes statistiques appropriées permettant de détecter et de pondérer à la baisse l'effet de ces études sont apparues dans la littérature. Ces méthodes dichotomisent chaque étude de la méta-analyse comme étant aberrante ou non aberrante et se concentrent uniquement sur l'examen de l'effet des études aberrantes sur la sensibilité et la spécificité sommaires. Dans ce travail, nous développons un modèle de mélange bivarié à effets aléatoires pour la méta-analyse des études sur la précision des tests de diagnostic en tenant compte de l'hétérogénéité des résultats des tests de diagnostic à l'intérieur d'une même étude et d'une étude à l'autre. Au lieu de dichotomiser les études, le modèle proposé génère la probabilité que chaque étude soit aberrante et permet d'évaluer l'impact des études aberrantes sur la sensibilité, la spécificité et l'hétérogénéité entre les études. Nous illustrons les performances de la méthode développée sur des données méta-analytiques réelles et simulées.

[11:05-11:20]

Jairo Diaz-Rodriguez (York University)

Spatial Estimation of Virus Infection Propensity from GPS Spatio-temporal Locations

Estimation spatiale de la propension d'infection virale à partir d'emplacements spatiotemporels GPS

Identifying areas of higher probability of infection is a crucial step towards disease management. Data consist of GPS tracking of individuals moving in a fenced area over given time periods. A test is performed on each individual to detect infection at the beginning and at the end of its time period. Infection is indirectly trans-

L'identification de régions ayant une probabilité élevée d'infection est une étape cruciale dans la gestion thérapeutique. Les données comportent le déplacement d'individus repéré par GPS dans une région délimitée sur une période de temps donnée. Nous faisons passer un test à chaque individu afin de détecter une infection au début et à la fin d'une période de temps. L'infection est en-

mitted by a vector. The purpose is to estimate the spatial distribution of infection. We model the problem as a spatial binomial inverse problem. We basically perform an epidemiologic tomography of a region by reconstructing the spatial disease propensity using non-invasive measurements. To enforce sparsity, maximum likelihood estimation is regularized by a total variation penalty. The method has the advantage of not employing cross-validation to select the penalization parameter. It can also test whether the propensity is constant over the region or not. We applied our method to simulated and real data showing better performance than current methodologies.

suite indirectement transmise par un vecteur. L'objectif est d'estimer la distribution spatiale de l'infection. Nous modélisons donc le problème en tant que loi binomiale spatiale inversée. En gros, nous réalisons une tomographie épidémiologique d'une région en reconstruisant la propension spatiale de maladie au moyen de mesures non intrusives. Afin de renforcer l'éparsité, l'estimation du maximum de vraisemblance est régulée par une pénalité de variation totale. Cette méthode à l'avantage de n'employer aucune validation croisée pour sélectionner le paramètre de pénalisation. Elle peut aussi tester si la propension est constante ou non dans la région. En appliquant notre méthode à des données réelles et simulées, nous avons démontré sa performance supérieure par rapport aux méthodologies courantes.

[11:20-11:35]

Yuan Fang (Binghamton University) **Jiachen Chen** (Boston University) **Joanne Murabito** (Boston University) **Kathryn Lunetta** (Boston University)

Modeling heterogeneity in cognitive trajectories in the Framingham Heart Study

Modélisation de l'hétérogénéité des trajectoires cognitives dans la Framingham Heart Study (FHS)

Heterogeneity exists in the rate of cognitive decline across different cognitive domains. Harmonized factor scores measuring global cognition for the memory, executive function, and language domains have been created in the Framingham Heart Study (FHS). We identified FHS participants with 2 or more factor scores after age 60 and fitted latent class mixed models (LCMM) to cluster trajectories from all three domains. Non-linear shapes of trajectories were modeled piecewise linearly followed by a stepwise selection of cluster-specific change points. We identified different latent classes of participants with early cognitive decline, compared to late decliners, across different domains. Stable subgroupings were yielded from 10-fold cross-validation. Our findings show differential patterns in cognitive aging in the FHS. Future associations between the identified subclasses, cognitive outcomes, and physical/biological markers may advance the knowledge of multiple pathways of cognitive aging.

L'hétérogénéité existe dans le taux de déclin cognitif dans divers domaines de la cognition. Des scores factoriels harmonisés mesurant la cognition globale dans les domaines de la mémoire, de la fonction exécutive et du langage ont été créés dans le cadre de l'étude Framingham Heart Study (FHS). Nous avons identifié des participants à l'étude avec au moins deux scores factoriels après l'âge de 60 ans et avons ajusté des modèles mixtes à classes latentes (LCMM) pour mettre en grappes les trajectoires dans les trois domaines. Des formes non linéaires de trajectoires ont été modélisées de façon linéaire par morceau suivies d'une sélection pas à pas des points de changement spécifiques aux grappes. Nous avons identifié diverses classes latentes de participants présentant un déclin cognitif précoce, comparativement à d'autres dont le déclin est tardif, dans différents domaines. Une validation croisée 10 fois a fourni des sous-groupes stables. Nos résultats montrent des motifs différentiels du vieillissement cognitif dans l'étude FHS. Des associations ultérieures entre les sous-classes identifiées, les résultats sur le plan cognitif et les marqueurs physiques/biologiques pourront faire progresser la connaissance des voies multiples du vieillissement cognitif.

[11:35-11:50]

Kunasekaran Nirmalkanna (Fisheries and Marine Institute of Memorial University of Newfoundland) **Noel Cadigan** (Fisheries and Marine Institute of Memorial University of Newfoundland) **S. J. W. W. M. M. P Weerasekera** (Fisheries and Marine Institute of Memorial University of Newfoundland)

Spatiotemporal relationship between weight and length of Atlantic cod on the southern Grand Bank of Newfoundland

Relation spatiotemporelle entre le poids et la longueur de la morue de l'Atlantique dans le sud des Grands Bancs de Terre-Neuve

The condition of a fish may be measured as its weight

L'état d'un poisson peut être mesuré par son poids divisé par le

divided by the expected weight at that size. Condition is assessed through the relationship between weight and length. We develop a spatiotemporal model for the weight–length relationship that is useful for predicting condition. Our main objective is to model how condition changes spatially, temporally (between- and within-years), and with length. We introduce an adjusted Gaussian AR(1) process that can account for cyclic seasonal changes in condition from higher values in the summer and fall following improved feeding conditions, to lower values in the winter and spring when prey are scarce. This process is assumed to be the same each year. We adjust the AR(1) process so that the correlation between condition at the end of the year, and condition at the beginning of the year, is the same as the correlation for any other consecutive days. We develop the model based on data for Atlantic cod from DFO spring and fall research surveys on the southern Grand Bank of Newfoundland. We use our model to infer spatiotemporal trends in starvation mortality, which is important information for sustainable fisheries management.

poids attendu selon la longueur. Son état est évalué par le lien entre le poids et la longueur. Nous développons un modèle spatiotemporel pour la relation poids-longueur qui est pratique pour la prédiction de l'état. Notre principal objectif est de modéliser la façon dont l'état varie dans l'espace, le temps (selon les années) et la longueur. Nous présentons un processus gaussien ajusté AR(1) qui peut tenir compte des changements saisonniers cycliques sur l'état, des valeurs supérieures en été et en automne en raison d'une forte alimentation, jusqu'aux valeurs inférieures en hiver et au printemps lorsque les proies se font rares. On suppose que ce processus reste le même chaque année. Nous ajustons le processus AR(1) pour faire en sorte que la corrélation entre l'état à la fin de l'année et au début de l'année soit la même que la corrélation pour tout autre jour consécutif. Nous développons le modèle à partir de données sur la morue de l'Atlantique tirées des enquêtes de recherche du printemps et de l'automne du MPO dans le sud des Grands Bancs de Terre-Neuve. Nous employons notre modèle pour inférer les schémas spatiotemporels de mortalité par inanition, qui représentent des renseignements importants pour la gestion de pêche durable.

Chair/Président: David Haziza

Organizer/Responsable: W. John Braun

Room/Salle: TB 210

Date: Monday May 29 / lundi 29 mai

Time/Heure: 10:20-11:50

Abstract/Résumé

[10:20-11:50]

Adele Ngi-Song (NSERC) **Rachel Desrochers** (Natural Sciences and Engineering Research Council of Canada) **David Haziza** (University of Ottawa)

NSERC Discovery Grant Application Assistance

Aide à la soumission d'une demande de subvention à la découverte du CRSNG

This workshop will be presented by NSERC Research Grants staff and will cover the NOI (Notification of Intent to Apply) 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.

Chair/Président: Wanhua Su

Room/Salle: RB 1201

Date: Monday May 29 / lundi 29 mai

Time/Heure: 10:20-11:50

Abstract/Résumé

[10:20-10:35]

Johanna de Haan-Ward (University of Western Ontario) **Douglas G. Woolford** (University of Western Ontario) **Simon J. Bonner** (University of Western Ontario)

Modelling Predicted Probabilities of Rare Events Using Stratified Sampling, with Application to Human-Caused Wildland Fire

Modélisation des probabilités prédites d'événements rares à l'aide d'un échantillonnage stratifié, avec application aux incendies de forêt d'origine humaine

Datasets on the occurrence of rare events, like wildland fires, are often large and severely imbalanced, where less than 1% of all observations are event occurrences (cases). The typical approach to make modelling more computationally feasible is to generate balanced datasets by randomly sampling both cases and controls. An offset or sampling weights are then added to the logistic regression model to obtain unbiased parameter estimates of the regression parameters. We propose a method that uses a stratified sampling design in place of random sampling, where we preferentially sample observations with conditions characteristic of the rare event by stratifying according to a key covariate. To then account for the stratified sampling design in the model, there is a separate estimate of the intercept for each stratum which are then combined into a single intercept, along with sampling weights. This method is illustrated via a fine-scale model for human-caused wildland fires in Ontario.

Les ensembles de données sur l'occurrence d'événements rares, tels que les incendies de forêt, sont souvent volumineux et fortement déséquilibrés, les occurrences d'événements (cas) représentant moins de 1 % de toutes les observations. L'approche classique pour rendre la modélisation plus facile à calculer consiste à générer des ensembles de données équilibrés en échantillonnant de manière aléatoire les cas et les témoins. On ajoute ensuite un décalage ou des poids d'échantillonnage au modèle de régression logistique afin d'obtenir des estimations non biaisées des paramètres de régression. Nous proposons une méthode qui utilise un plan d'échantillonnage stratifié au lieu d'un échantillonnage aléatoire, où nous échantillonnons de préférence les observations présentant des conditions caractéristiques de l'événement rare en les stratifiant en fonction d'une covariable clé. Pour tenir compte du plan d'échantillonnage stratifié dans le modèle, nous avons une estimation distincte de l'ordonnée à l'origine pour chaque strate, qui sont ensuite combinées en une seule ordonnée à l'origine, avec des poids d'échantillonnage. Nous illustrons la méthode par un modèle à échelle fine pour les incendies de forêt d'origine humaine en Ontario.

[10:35-10:50]

Meixi Chen (University of Waterloo) **Martin Lysy** (University of Waterloo) **Reza Ramezan** (University of Waterloo)

Decoding Neural Population Dynamics Through Continuous-Time Latent Factor Models

Décodage de la dynamique d'une population neuronale à l'aide de modèles de facteurs latents en temps continu

The human brain contains some hundred billion neurons which communicate through electrochemical waves called spikes. A sequence of consecutive spikes from a neuron is called a spike train, which encodes information about the stimuli to which the brain has been exposed. Statistical models play an important role in ex-

Le cerveau humain renferme des centaines de milliards de neurones qui communiquent par vagues électrochimiques appelées pics. Une séquence de pics consécutifs d'un neurone est appelée train de pics et elle encode l'information sur les stimuli auxquels le cerveau a été exposé. Les modèles statistiques jouent un rôle important dans l'exploration des phénomènes neuroscienti-

ploring neuroscientific phenomena such as the interactions among multiple neurons over time. We introduce a novel continuous-time latent factor model for studying neuronal interactions based on simultaneously recorded spike trains. The activities of a neural population are described by a multivariate Wiener process, which itself depends on a small number of dynamic latent factors determining the neuronal clustering. We show that our model is highly scalable and can accurately recover neuronal clusters in simulations. Applying our model to real data from rats' medial prefrontal cortex, we found different levels of coordinated activities in different brain areas.

fiques, comme les interactions entre les neurones avec le temps. Nous présentons un nouveau modèle de facteurs latents en temps continu pour l'étude d'une population neuronale basée sur les trains de pics enregistrés simultanément. Les activités d'une population neuronale sont décrites par une processus de Wiener multivarié, qui lui-même dépend d'un petit nombre de facteurs latents dynamiques déterminant le regroupement neuronal. Nous montrons que notre modèle est très évolutif et peut récupérer avec exactitude les regroupements neuronaux dans les études de simulation. En appliquant notre modèle à des données réelles sur le cortex préfrontal médian de rats, nous avons trouvé divers degrés d'activités coordonnées dans différentes zones du cerveau.

[10:50-11:05]

Jianchu Chen (University of Waterloo) **Richard J. Cook** (University of Waterloo)

A Copula Model for Recurrent and Terminal Events

Modèle de copule pour événement récurrent et terminal

We present a new model for joint analysis of recurrent event and terminal event processes subject to right-censoring. A copula function links a random effect modulating a recurrent event rate function with a terminal event. Conditional on the random effect, recurrent events are governed by a Poisson process with a multiplicative rate function; a Cox model can be used for the terminal event process. A direct optimization of the observed data likelihood as well as an EM algorithm for complete data likelihood are used for parameter estimation with variance estimates derived according to Louis (1982). A computationally more convenient two-stage estimation procedure is also investigated. Simulation studies demonstrate satisfactory finite sample performance under simultaneous and two-stage estimation, and an application is given to a motivating study.

Nous présentons un nouveau modèle d'analyse conjointe de processus d'événement récurrent et d'événement terminal assujettis à une censure à droite. Une fonction de copule lie un effet aléatoire modulant une fonction de taux d'événements récurrents à un événement terminal. Conditionnellement à l'effet aléatoire, les événements récurrents sont régis par un processus Poisson avec une fonction de taux multiplicateur ; un modèle de Cox peut être utilisé pour le traitement de l'événement terminal. Une optimisation directe de la vraisemblance des données observées de même qu'un algorithme EM pour la vraisemblance des données complètes sont utilisés pour l'estimation des paramètres avec des estimations de la variance dérivées selon Thomas A. Louis (1982). Une procédure d'estimation en deux étapes plus pratique sur le plan computationnel est aussi étudiée. Des études de simulation illustrent la performance satisfaisante pour échantillon fini tant pour l'estimation simultanée qu'en deux étapes et une application est fournie pour une étude motivante.

[11:05-11:20]

Fabiha Binte Farooq (Simon Fraser University) **Christina Nieuwoudt** (Simon Fraser University)

Statistical prioritization of sequencing variants in disease-enriched families

Priorisation statistique des variables de séquençage dans des familles avec surreprésentation d'une maladie

A BC Cancer study collected families with multiple cases. Affected relatives had their exomes sequenced for shared genetic variants. Since different families harbor different causal variants and each family harbors many variants, the goal was to prioritize shared variants for further investigation. The transmission-disequilibrium test prioritizes variants based on departures from Mendelian transmission in parent-child trios.

Une étude sur le cancer menée en Colombie-Britannique a recensé des familles avec des cas multiples. On a procédé au séquençage des exomes de parents affectés pour déterminer des variables génétiques partagées. Comme les variables causales diffèrent en fonction des familles et que chacune a plusieurs variables, l'objectif était de prioriser les variables partagées pour approfondir l'étude. Le test de déséquilibre de transmission accorde la priorité aux variables basées sur les écarts de la transmission mendélienne

We extend this idea to families and allow for disease subtypes and other characteristics of the individual or variant. Global approaches condition on a variant being observed in the study and include an unknown nuisance parameter. By contrast, local approaches condition on a variant being observed in specific families to eliminate the nuisance parameter. Simulation results indicate that global approaches are robust to misspecification of the nuisance parameter and prioritize better than local approaches even when misspecified.

dans les trios parents-enfant. Nous étendons ce concept aux familles et intégrons des sous-types de la maladie et d'autres caractéristiques de l'individu ou de la variable. Les approches globales reposent sur l'observation d'une variable dans le cadre de l'étude et comprennent un paramètre de nuisance inconnu. Par contraste, les approches locales reposent sur l'observation d'une variable dans des familles précises afin d'éliminer le paramètre de nuisance. Des résultats d'étude de simulation indiquent que les approches globales sont robustes aux erreurs de spécification du paramètre de nuisance et que même incorrectement spécifiées, elles priorisent mieux que les approches locales.

[11:20-11:35]

Jingyue Huang (University of Waterloo) **Leilei Zeng** (University of Waterloo) **Changbao Wu** (University of Waterloo)

Sample Empirical Likelihood Approaches for Causal Inference

Approches de vraisemblance empirique fondée sur l'échantillon pour une inférence causale

The causal inference has attracted increased interest in recent years. In this talk, we explore two different approaches using the sample empirical likelihood (SEL) for the estimation and inference of the average treatment effect (ATE). For the first approach, we use the standard empirical likelihood function and introduce a propensity score (PS) calibrated constraint in addition to the standard model-calibration and parameter constraints. For the second approach, the propensity scores are used to form weighted versions of the model-calibration and parameter constraints. Both approaches involve a two-sample empirical likelihood function. Theoretical properties of the two approaches are studied, and asymptotic results are established. We also discuss the bootstrap-calibrated confidence intervals for both approaches. Finite sample performances of the methods are investigated through simulation studies. This is a joint work with my Ph.D. supervisors, Dr. Leilei Zeng and Dr. Changbao Wu.

L'intérêt pour l'inférence causale s'est accru ces dernières années. Notre exposé porte sur deux approches différentes misant sur la vraisemblance empirique fondée sur l'échantillon (SEL) pour l'estimation et l'inférence de l'effet moyen du traitement (ATE). La première approche fait appel à la fonction de vraisemblance empirique standard et introduit une contrainte calibrée du score de propension (PS), en plus des contraintes standards de calibrage des modèles et des paramètres. La deuxième approche utilise les scores de propension pour former des versions pondérées des contraintes de calibrage des modèles et des paramètres. Les deux approches impliquent une fonction de vraisemblance empirique fondée sur deux échantillons. Nous étudions les propriétés théoriques des deux approches et établissons des résultats asymptotiques. Nous abordons également les intervalles de confiance calibrés par bootstrap pour les deux approches. Des études de simulation permettent d'examiner les performances des échantillons de population finie des méthodes. Ce travail est fait en collaboration avec mes directeurs de thèse, la Dre Leilei Zeng et le Dr Changbao Wu.

[11:35-11:50]

Ashani N. Wickramasinghe (University of Manitoba) **Saman Muthukumarana** (University of Manitoba) **Surajith N. Wanasundara** (University of Manitoba) **Matt Schaubroeck** (ioAirFlow)

Identifying Locations with Abnormal Behavior of Temperature in School Buildings Using Anomaly Detection

Repérage de régions où la température se comporte anormalement dans les établissements scolaires au moyen de détection d'anomalie

Time series data collected using wireless sensors can provide insights into a building's heating, ventilation, and air conditioning (HVAC) system. Anomalies of these sensor measurements can be used to identify poorly designed or maintained locations of a building. In this study, we developed a scoring method to identify

Les données de séries temporelles recueillies par capteurs sans fil peuvent fournir des aperçus concernant le système de chauffage, de ventilation et de climatisation (CVC) d'un immeuble. Les anomalies parmi ces mesures peuvent servir à identifier les régions mal entretenues ou mal conçues d'un bâtiment. Dans cette étude, nous avons développé une méthode de pointage afin de repérer

sensors that shows collective anomalies due to environmental issues. The Dynamic Time Warping (DTW) based anomaly detection method was applied to identify collective anomalies. Then, a score for each sensor was obtained by taking the weighted sum of the selected features of anomalies. The weights were optimized using a well-defined simulation study with a grid search algorithm. Finally, we used two case studies to evaluate the performance of our method. In conclusion, this newly developed scoring method successfully detects collective anomalies even with data collected over one week, compared to the machine learning models which need more data to train themselves.

les capteurs qui démontrent des anomalies collectives en raison de problèmes environnementaux. Nous avons repéré les anomalies collectives par l'application de la méthode de détection d'anomalie basée sur l'alignement temporel dynamique (DTW). Ensuite, on obtient un score pour chaque capteur en utilisant la somme pondérée des variables sélectionnées à partir des anomalies. Les poids ont été optimisés à l'aide d'une étude de simulations bien définie avec un algorithme de recherche par quadrillage. Enfin, nous avons réalisé deux études de cas pour évaluer la performance de notre méthode. En conclusion, cette nouvelle méthode de pointage réussit à détecter les anomalies collectives mêmes avec des données recueillies sur une semaine, contrairement aux modèles d'apprentissage automatique qui ont besoin de plus de données pour se former.

Statistical design of Bayesian adaptive and platform trials
Conception statistique d'essais adaptatifs et de plates-formes bayésiennes

Chair/Président: Luke Hagar

Organizer/Responsable: Shirin Golchi

Room/Salle: AT 102

Date: Monday May 29 / lundi 29 mai

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-14:00]

Anna Heath (The Hospital for Sick Children)

Efficient Design of Platform Trials with Complex Outcomes

Conception efficace d'études plateformes avec résultats complexes

Adaptive platform trials answer multiple research questions for a disease using a common infrastructure and can add research questions during the trial. These trials provide an efficient and flexible approach to screen many interventions, crucial for efficiently evaluating treatments. However, this flexibility requires us to repeatedly analyse the study data to react to accumulating evidence. Thus, the characteristics of the proposed trial design must be evaluated using simulations to understand the impact of these analyses. These simulations are challenging when the proposed analysis requires non-conjugate Bayesian analysis. To overcome this barrier, we used the Integrated Nested Laplace Approximations (INLA) algorithm, an efficient approximation method for the posterior distribution in hierarchical generalised linear models. Thanks to INLA, we designed studies with ordinal, binary and survival outcomes that use hierarchical models to account for treatment effect heterogeneity.

Les études cliniques plateformes sont innovantes et permettent de tester plusieurs médicaments simultanément. Ces études sont souples et visent à évaluer des interventions en manière efficace et ajouter des sous-études à tout moment. Afin de réagir vite, il faut analyser les données souvent pendant l'étude. Donc, il faut assurer les propretés statistiques de ces analyses fréquentes en utilisant la méthode de simulation. Malheureusement, la méthode de simulation pose un défi dans le cas où les analyses bayésiennes ne sont pas conjuguées. Pour réduire la complexité de la simulation, on a utilisé l'algorithme « Intergrated Nested Laplace Approximation » (INLA) qui calcule une approximation de la distribution a posteriori pour les modèles linéaires généralisés hiérarchiques. Grâce à INLA, on a évalué des études qui collectent les données de survie, ordinales et binaires et utilisent les modèles hiérarchiques pour tenir compte d'hétérogénéité des effets de traitement.

[14:00-14:30]

Steffen Ventz (Univrsity of Minnesota)

Bayesian Multi-Arm De-Intensification Designs

Plans bayésiens de désintensification à plusieurs bras

In recent years new cancer treatments improved survival in multiple histologies. Some of these therapeutics, and in particular treatment combinations, are often associated with severe treatment-related adverse events (AEs). It is therefore important to identify alternative de-intensified therapies, for example dose-reduced therapies, with reduced AEs and similar efficacy. We intro-

Ces dernières années, de nouveaux traitements anticancéreux ont permis d'améliorer la survie dans de multiples histologies. Certaines de ces thérapies, et en particulier les combinaisons de traitements, sont souvent associées à des effets indésirables graves liés au traitement. Il est donc important d'identifier des thérapies alternatives désintensifiées, par exemple des thérapies à dose réduite, avec des effets indésirables réduits et une effica-

duce a sequential design for multi-arm de-intensification studies. The design evaluates multiple de-intensified therapies at different dose levels, one at the time, based on modeling of toxicity and efficacy endpoints. We study the utility of the design in oropharynx cancer de-intensification studies. We use a Bayesian nonparametric model for efficacy and toxicity outcomes to define decision rules at interim and final analyses. Interim decisions include early termination of the study due to inferior survival of experimental arms compared to the standard of care (SOC), and transitions from one de-intensified treatment arm to another with a further reduced dose when there is sufficient evidence of non-inferior survival. We evaluate the operating characteristics of the design using data from recent de-intensification studies in human papillomavirus (HPV)-associated oropharynx cancer.

cité similaire. Nous présentons un plan séquentiel pour les études de désintensification à plusieurs bras. Ce plan évalue plusieurs thérapies désintensifiées à différents niveaux de dose, une à la fois, sur la base d'une modélisation de la toxicité et des critères d'efficacité. Nous étudions l'utilité de ce modèle dans les études de désintensification du cancer de l'oropharynx. Nous utilisons un modèle bayésien non paramétrique pour les résultats d'efficacité et de toxicité afin de définir des règles de décision lors des analyses intermédiaires et finales. Les décisions intermédiaires comprennent l'arrêt prématuré de l'étude en raison de la survie inférieure des bras expérimentaux par rapport à la norme de soins, et les transitions d'un bras de traitement désintensifié à un autre avec une dose réduite davantage lorsqu'il y a suffisamment de preuves d'une survie non inférieure. Nous évaluons les caractéristiques de fonctionnement de ce modèle à l'aide de données provenant d'études récentes de désintensification du cancer de l'oropharynx associé au virus du papillome humain (VPH).

[14:30-15:00]

Shirin Golchi (McGill University) **James Willard** (McGill University)

Estimating the Sampling Distribution of a Bayesian Probability Statement with Application to Clinical Trials

Estimation de la distribution échantillonnale d'un énoncé de probabilité bayésien avec application à des essais cliniques

Bayesian posterior probability statements are commonly used as test statistics in Bayesian clinical trials. Despite the popular use of Bayesian inference in clinical trials, the design needs to be assessed with respect to frequentist operating characteristics (power and type I error rate) that are defined according to the sampling distribution of the test statistic. Unlike in classical hypothesis tests used in randomized clinical trials, the sampling distribution of the "Bayesian test statistic" is not available analytically and is obtained via simulations. In this talk we propose an approach for modelling the sampling distribution of the test statistic taking advantage of the known behaviour of the Bayesian probability statements as a function of the effect and sample sizes. The proposed approach significantly reduces the computational burden in design of Bayesian clinical trials thereby enabling exploration of a larger variety of designs with complex analysis models.

Les énoncés de probabilité a posteriori bayésienne servent couramment de statistique de test dans les essais cliniques bayésiens. Malgré la popularité de l'inférence bayésienne dans les essais cliniques, le devis nécessite une évaluation quant aux caractéristiques fréquentistes (puissance et taux d'erreur de type 1) définies en fonction de la distribution échantillonnale de la statistique de test. Contrairement aux tests d'hypothèse classiques utilisés dans les essais cliniques randomisés, la distribution échantillonnale de la ≪ statistique de test bayésienne ≫ n'est pas disponible de façon analytique et s'obtient par des simulations. Notre exposé porte sur une approche pour la modélisation de la distribution échantillonnale de la statistique de test qui prend en compte le comportement connu des énoncés de probabilité bayésienne comme fonction des tailles d'effets et d'échantillons. L'approche proposée réduit sensiblement le fardeau computationnel dans la conception d'essais cliniques bayésiens, ce qui permet une exploration d'une plus grande diversité de devis avec des modèles d'analyse complexe.

In celebration of 50 years of research in Stochastics at Carleton University: Session in Honour of Miklós Csörgő
Célébrations de 50 ans de recherche en stochastique à l'Université Carleton : Session en l'honneur de Miklós Csörgö

Chair/Président: Barbara Szyszkowicz

Organizer/Responsable: Barbara Szyszkowicz

Room/Salle: AT 101

Date: Monday May 29 / lundi 29 mai

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-14:00]

Barbara Szyszkowicz (Carleton University)
Revisiting ICAMPS'97, 8-13 July 1997, Ottawa.
Revisiter ICAMPS' 97, du 8 au 13 juillet 1997, Ottawa.

I will present, with some comments, a part of a video from ICAMPS'97 -International Conference on Asymptotic Methods in Probability and Statistics, that I have organized in honour of Miklós Csörgõ on the occasion of his 65th Birthday. This video contains the talk by Don Dawson about Miklós' research. One can also see Miklós talking for a few minutes about his joint work with Pál Révész (who couldn't be there, but is visible on the screen with registered greetings and introduction of his paper). The video concludes with the introduction of Sándor Csörgõ's presentation.

Je présenterai, avec certains commentaires, une partie d'une vidéo tirée de ICAMPS' 97, la conférence internationale sur les méthodes asymptotiques en probabilité et statistiques, que j'ai organisée en honneur de Miklós Csörgõ à l'occasion de son 65e anniversaire. Cette vidéo contient l'exposé de Don Dawson à propos de la recherche de Miklós. On peut aussi voir Miklós parler quelques minutes au sujet de son travail conjoint avec Pál Révész (qui n'a pas pu être présent, mais est visible sur l'écran avec les salutations et l'introduction de son article). La vidéo conclut avec l'introduction de la présentation de Sándor Csörgõ's.

[14:00-14:30]

Yifan Li (University of Western Ontario) **Reg Kulperger** (University of Western Ontario) **Hao Yu** (University of Western Ontario)
Do we still have a nice CLT under general model uncertainty?
Avons-nous encore un bon théorème central limite en cas d'incertitude générale du modèle ?

The long-existing concern of model uncertainty becomes more notable in the current era of big data and machine learning, where both data and models are becoming more complicated. However, the limiting theorems (e.g. CLTs) under classical probabilities hold a fixed-model view, requiring information on the variance. In response to model uncertainty, it is fundamental to extend the CLT accordingly. Meanwhile, it is important to extend it in a nice way, that preserves these properties: (1) A kind of "normal" as the asymptotic object, (2) Multivariate consistency, (3) Exchangeability. The existing attempts include the extension based on nonlinear expectations, such as the G-expectation and an equivalent view from the nonlinear PDE. However,

Le problème de l'incertitude du modèle, qui existe depuis longtemps, prend de l'ampleur à l'ère des mégadonnées et de l'apprentissage automatique, où les données et les modèles sont de plus en plus complexes. Toutefois, les théorèmes limites (p. ex., les théorèmes centraux limites) dans le cadre des probabilités classiques s'appuient sur un modèle fixe, nécessitant de l'information sur la variance. En réponse à l'incertitude du modèle, il est essentiel d'étendre le théorème central limite en conséquence. Entretemps, il est important de l'étendre de manière à préserver les propriétés suivantes : 1) Une sorte de « normale » comme objet asymptotique, 2) la cohérence multivariée, 3) l'échangeabilité. Les tentatives actuelles comprennent l'extension basée sur les espérances non linéaires, telles que la g-espérance et une vue équivalente de l'équation aux dérivées partielles non linéaire. Tou-

these attempts do not satisfy (2) and (3). We prove a CLT (called the semi-G-CLT) under model uncertainty satisfying (1)-(3) with temporal or spatial symmetry. As an application, we provide a nonparametric inference on the model structure of noise data.

tefois, ces tentatives ne préservent pas les propriétés 2) et 3). Nous proposons un théorème central limite (appelé théorème central limite semi-G) en cas d'incertitude du modèle préservant les propriétés 1) et 3) avec une symétrie temporelle ou spatiale. En guise d'application, nous fournissons une inférence non paramétrique sur la structure du modèle des données de bruit.

[14:30-15:00]

Masoud Nasari (Bank of Canada and Carleton University)

A new inferential framework for the mean

Nouveau cadre inférentiel pour la moyenne

A new inferential framework for a vast class of weakly stationary data will be introduced. The framework exploits the information contained in a given sample by incorporating a set of exogenous randomizing agents generated from a stable distribution such as Cauchy. The framework allows for replication to smooth out the effect of randomization.

Nous présenterons un nouveau cadre inférentiel pour une vaste catégorie de données faiblement stationnaires. Ce cadre exploite les données contenues dans un échantillon donné en incorporant un ensemble d'agents aléatoires exogènes générés à partir d'une distribution stable telle que celle de Cauchy. Ce cadre permet la réplication afin d'atténuer l'effet de la randomisation.

Chair/Président: Jun Cai

Organizer/Responsable: Hong Li

Room/Salle: TB 210

Date: Monday May 29 / lundi 29 mai

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-14:00]

Jianxi Su (PURDUE UNIVERSITY)

Some statistical properties of the multivariate truncated normal distributions with actuarial applications in view

Quelques propriétés statistiques des distributions normales tronquées multivariées en vue d'applications actuarielles

The multivariate normal distributions have been widely advocated as an elegant yet flexible model, which uses a simple covariance matrix parameter to capture the intricate dependence involved in high-dimensional data. However, insurance loss random variables are often assumed to be non-negative. Thereby, the multivariate normal distributions must be properly truncated to be adopted in insurance applications. In this presentation, we are to going to review some fundamental statistics properties of the multivariate truncated normal distributions, including their independence, non-steepness and maximum likelihood estimation properties. For actuarial applications, we propose an efficient numeric scheme to compute the tail conditional allocation for the multivariate truncated normal distributions, which hinges on the holonomic gradient method for computing the probability content of a simplex structure.

Les distributions normales multivariées ont été largement préconisées en tant que modèle élégant et souple, qui utilise un simple paramètre de matrice de covariance pour capturer la dépendance complexe que présentent les données en haute dimension. Cependant, les variables aléatoires des sinistres d'assurance sont souvent supposées non négatives. Par conséquent, les distributions normales multivariées doivent être correctement tronquées pour être adoptées dans les applications d'assurance. Dans cette présentation, nous allons passer en revue certaines propriétés statistiques fondamentales des distributions normales tronquées multivariées, y compris leurs propriétés d'indépendance, d'absence de pente et d'estimation du maximum de vraisemblance. Pour les applications actuarielles, nous proposons un schéma numérique efficace pour calculer l'allocation conditionnelle de la queue pour les distributions normales tronquées multivariées, qui s'articule autour de la méthode du gradient holonomique pour calculer le contenu de probabilité d'une structure simplexe.

[14:00-14:30]

Kenneth Q. Zhou (Arizona State University) **Hong Li** (University of Guelph) **Ze Chen** (Renmin University of China) **Yu Mao** (University of International Business and Economics)

Modeling and Managing Pandemic Mortality Risk via a Stochastic Compartmental Model

Modélisation et gestion du risque de mortalité lié à une pandémie à l'aide d'un modèle compartimental stochastique

In this paper, we present a novel approach to model the excess death rate of pandemics using a stochastic Susceptible-Infected-Recovered-Deceased (SIRD) model. The proposed SIRD model not only accounts for the stochastic nature of pandemics, but also takes parameter uncertainty into consideration. Based on the modeling results, we further propose a pandemic (PAN) bond that can serve as a financial instrument for mitigat-

Notre exposé présente une nouvelle approche de modélisation du taux de surmortalité en pandémie qui utilise un modèle stochastique SIRD (susceptibilité, infection, récupération, décès). En plus de prendre en compte la nature stochastique de la pandémie, le modèle SIRD proposé considère également l'incertitude des paramètres. Basé sur les résultats de la modélisation, nous proposons aussi une obligation pandémique (PAN) qui peut servir d'instrument financier pour atténuer l'impact économique de la pandémie.

ing the economic impact of pandemics. The PAN bond is designed to enable governments and insurance companies to appropriately redistribute the risk associated with pandemics and provide a stable source of funding for pandemic response efforts. Our numerical results show that the stochastic SIRD model provides a reliable estimate of the excess death rate and the PAN bond is a valuable tool for managing the financial consequences of pandemics to a life insurance company.

L'obligation PAN est conçue pour permettre aux gouvernements et aux compagnies d'assurance de redistribuer adéquatement le risque associé à la pandémie et de fournir une source stable de financement des efforts pour répondre à la pandémie. Nos résultats numériques montrent que le modèle stochastique SIRD fournit une estimation fiable du taux de surmortalité et que l'obligation PAN est un outil de gestion valable des conséquences financières de la pandémie pour une compagnie d'assurance-vie.

[14:30-15:00]

Hong Li (University of Guleph) **Jianxi Su** (Purdue University)

Mitigating Wildfire Losses via Insurance-Linked Securities: Modeling and Risk Management Perspectives

Mitiger les pertes relatives aux feux incontrôlés grâce aux contrats d'assurance titrisés : perspectives de modélisation et de gestion de risques

This paper considers the use of catastrophe (CAT) bonds for managing wildfire risks. A class of Bayesian dynamic models is proposed to model wildfire losses, with which the pricing and hedge effectiveness of wildfire CAT bonds are investigated. The proposed wildfire model is able to capture stylized features of wildfire data including zero inflation, over-dispersion, time-varying patterns, and spatial dependence. Through extensive quantitative analyses with US data, we find that using CAT bonds as a hedging tool can substantially reduce the variability risk and tail risk of the insurer's liability. Furthermore, the hedge effectiveness using index-based CAT bonds, whose payoffs are determined by aggregate wildfire losses in a larger area than that the insurer operates in, is still satisfying. Our study suggests that wildfire CAT bonds could serve as an effective complement to traditional reinsurance for insurers to manage their wildfire risk.

Cet article aborde l'usage des obligations catastrophes (CAT) pour la gestion de risques de feux incontrôlés. Nous proposons une classe de modèles dynamiques bayésiens pour modéliser les pertes relatives aux feux incontrôlés, à partir desquels nous évaluons le prix et l'efficacité de couverture des obligations CAT pour les feux incontrôlés. Le modèle de feu incontrôlé proposé réussi à saisir les traits stylisés des données sur les feux comprenant l'inflation nulle, la surdispersion, les schémas variants dans le temps et la dépendance spatiale. Au moyen d'analyses quantitatives complètes avec des données américaines, nous découvrons que l'adoption des obligations CAT en guise d'instrument de couverture peut réduire substantiellement le risque de variabilité et le risque extrême de la responsabilité de l'assureur. De plus, l'efficacité de la couverture au moyen des obligations CAT basées sur indice reste adéquate, même si les remboursements sont déterminés par les pertes agrégées relatives aux feux incontrôlés dans un secteur plus grand que celui dans lequel l'assureur travaille. Notre étude suggère que les obligations de feux CAT pourraient servir d'ajout efficace à la réassurance traditionnelle pour les assureurs afin de gérer les risques de feux incontrôlés.

Early Career Mentoring through the SSC Accreditation Program
Mentorat de début de carrière dans le programme d'accréditation de la SSC

Chair/Président: Darcy C. Pickard

Organizer/Responsable: Fabrice Larribe, Beatrice D. Baribeau

Room/Salle: TB 238

Date: Monday May 29 / lundi 29 mai

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-14:00]

Judy-Anne W. Chapman (Queen's University (retired))
Premise for Professional Mentoring of A.Stat. by P.Stat

Présentation de mentorat professionnel pour statisticiens associés (A. Stat) par des statisticiens professionnels (P.Stat.)

This talk will provide an overview of the SSC Mentorship program and how it enhances the development of Canadian professional practice, through the sharing of knowledge and experience between Professional Statisticians (P.Stat.) and members entering as Associate Statisticians (A.Stat.). I will discuss the role of the Accreditation Mentoring Facilitator, the Mentor's role, and the Mentee's role. As well, the benefits to the Mentee, Mentor, and Employer will be outlined. Decades of personal experience as Facilitator, Mentee, and Mentor will be used to describe the importance of the SSC Accreditation Mentoring Program.

Cet exposé offre un aperçu du programme de mentorat de la SSC et démontre comment il améliore la pratique professionnelle de la statistique au Canada par le partage de connaissances et d'expérience entre praticiens professionnels (P.Stat.) et membres nouvellement qualifiés comme statisticiens associés (A.Stat.). Je discuterai du rôle du facilitateur à l'accréditation au mentorat, du rôle du mentor et de celui de la personne mentorée. Sans oublier la présentation des avantages pour les mentors, les personnes mentorées et les employeurs. Je me sers de décennies d'expériences personnelles provenant de facilitateurs, mentors et personnes mentorées pour illustrer l'importance du programme d'accréditation au mentorat de la SSC.

[14:00-14:30]

Kevonn Morgan (RBC - Manager, Global AML Insights & Analytics)
Accreditation Mentoring from the perspective of an A.Stat

Accréditation au mentorat selon un A.Stat

This presentation will provide an overview of accreditation through the lens of an A.Stat. I will discuss the benefits and advantages of obtaining accreditation, my experience of having a mentor through the mentorship program, and the steps to obtain the A.Stat accreditation. Additionally, I will explain the impact of the accreditation on my career and how it has enabled me to confidently pursue new opportunities.

Cette présentation procurera un aperçu de l'accréditation selon la perspective d'un A.Stat. J'aborderai les avantages d'obtenir une accréditation, mon expérience après avoir été mentoré par l'entremise du programme, et les étapes pour obtenir une accréditation de A.Stat. En outre, j'expliquerai l'impact de l'accréditation sur ma carrière, et comment elle m'a permis de rechercher de nouvelles opportunités en toute confiance.

[14:30-15:00]

May Raad (HDR Corporation)
Accreditation Mentoring from the perspective of a P.Stat.

Mentorat au programme d'accréditation dans la perspective d'un statut de statisticien professionnel (P.Stat.)

Career choices for young statisticians and their paths for professional development are diverse. How does one

Pour les jeunes statisticiens, autant les choix de carrière que les avenues de perfectionnement professionnel sont divers. Comment

plan career development which may include growth opportunities within his or her place of employment, investing in continuing education, or working towards a P.Stat. designation? With over 30 years of work experience as a professional statistician in the public and private sectors, I see many benefits of becoming a mentor in the accreditation program, both for the mentor and mentee. Sharing lessons learned or career decisions steps from my own career is personally rewarding, especially if it can guide the mentee to move forward with goals and push boundaries. The dialogue between mentor and mentee is invigorating as new ideas are exchanged and clear career progression paths naturally unfold. For this session, I will discuss my recent experience as an accreditation mentor and bring in other perspectives from my role as a leadership mentor within my current organization.

planifier l'évolution d'une carrière qui peut comprendre des occasions de croissance chez un employeur, investir dans la formation continue ou travailler à devenir statisticien professionnel (P.Stat.) accrédité ? Fort d'une expérience de plus de 30 ans à titre de statisticien professionnel dans les secteurs public et privé, je vois de nombreux avantages à devenir mentor dans un programme d'accréditation, autant pour le mentor que pour le mentoré. Je trouve personnellement gratifiant de partager les leçons apprises ou les prises de décision en carrière à partir de ma propre expérience, en particulier si je peux guider le mentoré à progresser en se fixant des objectifs et à repousser ses limites. L'échange de nouvelles idées et les possibilités de progression de carrière qui se dessinent naturellement rendent le dialogue entre mentor et mentoré très stimulant. Au cours de cette séance, j'aborderai mon expérience récente de mentor dans un programme d'accréditation et présenterai d'autres perspectives à titre de mentor en matière de leadership au sein de mon organisation actuelle.

Chair/Président: Dehan Kong

Organizer/Responsable: Dehan Kong

Room/Salle: AT 301

Date: Monday May 29 / lundi 29 mai

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-14:00]

Oscar Hernan Madrid Padilla (UCLA) **Hangjian Li** (University of California, Los Angeles) **Qing Zhou** (University of California, Los Angeles)

Learning Gaussian DAGs from Network Data

Apprentissage de DAG gaussiens à partir de données de réseau

Structural learning of directed acyclic graphs (DAGs) or Bayesian networks has been studied extensively under the assumption that data are independent. We propose a new Gaussian DAG model for dependent data which assumes the observations are correlated according to an undirected network. Under this model, we develop a method to estimate the DAG structure given a topological ordering of the nodes. The proposed method jointly estimates the Bayesian network and the correlations among observations by optimizing a scoring function based on penalized likelihood. We show that under some mild conditions, the proposed method produces consistent estimators after one iteration. Extensive numerical experiments also demonstrate that by jointly estimating the DAG structure and the sample correlation, our method achieves much higher accuracy in structure learning. When the node ordering is unknown, through experiments on synthetic and real data, we show that our algorithm can be used to estimate the correlations between samples, with which we can de-correlate the dependent data to significantly improve the performance of classical DAG learning methods.

L'apprentissage structurel des graphes orientés acycliques (DAG) ou des réseaux bayésiens a été largement étudié dans l'hypothèse où les données sont indépendantes. Nous proposons un nouveau modèle de DAG gaussien pour données dépendantes qui suppose que les observations sont corrélées selon un réseau non orienté. Dans le cadre de ce modèle, nous développons une méthode permettant d'estimer la structure du DAG en fonction de l'ordre topologique des nœuds. La méthode proposée estime conjointement le réseau bayésien et les corrélations entre les observations en optimisant une fonction de notation basée sur la vraisemblance pénalisée. Nous montrons que sous certaines conditions légères, la méthode proposée produit des estimateurs cohérents après une itération. Des expériences numériques approfondies démontrent également qu'en estimant conjointement la structure du DAG et la corrélation de l'échantillon, notre méthode atteint une précision beaucoup plus élevée dans l'apprentissage de la structure. Lorsque l'ordre des nœuds est inconnu, nous montrons, grâce à des expériences sur des données synthétiques et réelles, que notre algorithme peut être utilisé pour estimer les corrélations entre les échantillons, ce qui nous permet de décorréler les données dépendantes et d'améliorer de manière significative les performances des méthodes classiques d'apprentissage de DAG.

[14:00-14:30]

Peijun Sang (University of Waterloo)

Nonlinear function-on-function Regression by RKHS

Régression de fonction sur fonction non linéaire à partir du RKHS

We propose a nonlinear function-on-function regression model where both the covariate and the response are

Nous proposons un modèle de régression fonction sur fonction non linéaire dans lequel les deux covariables et la réponse sont des

random functions. The nonlinear regression is carried out in two steps: we first construct Hilbert spaces to accommodate the functional covariate and the functional response, and then build a second-layer Hilbert space for the covariate to capture nonlinearity. The second-layer space is assumed to be a reproducing kernel Hilbert space, which is generated by a positive kernel determined by the inner product of the first-layer Hilbert space for X–this structure is known as the nested Hilbert spaces. We develop estimation procedures to implement the proposed method, which allows the functional data to be observed at different time points for different subjects. Furthermore, we establish the convergence rate of our estimator as well as the weak convergence of the predicted response in the Hilbert space. Numerical studies including both simulations and a data application are conducted to investigate the finite-sample performance of our estimator.

fonctions aléatoires. La régression non linéaire est réalisée en deux étapes : tout d'abord, il faut construire des espaces de Hilbert pour accommoder les covariables fonctionnelles et la réponse fonctionnelle ; puis construire un espace de Hilbert de seconde couche pour les covariables afin de saisir la non-linéarité. On suppose que l'espace de seconde couche est un espace de Hilbert à noyau reproduisant, étant généré par un noyau positif déterminé par le produit intérieur de l'espace de Hilbert de première couche pour X. Cette structure est connue sous le nom d'espaces de Hilbert imbriqués. Nous concevons des processus d'estimation afin d'implanter la méthode proposée, qui permet l'observation des données fonctionnelles à différents points dans le temps et pour différents sujets. De plus, nous établissons le taux de convergence de notre estimateur ainsi que la faible convergence de la réponse prédite dans l'espace de Hilbert. Nous menons des études numériques en simulations et appliquées à des données dans le but d'évaluer la performance d'échantillon fini de notre estimateur.

[14:30-15:00]

Archer Yang (McGill University) **Yue Zhao** (University of York) **Yi Lian** (University of Pennsylvania) **Jun Fan** (McGill University) **Yuwen Gu** (University of Connecticut)

Flexible Regularized Estimating Equations: Some New Perspectives

Équations d'estimation régularisées flexibles : Quelques nouvelles perspectives

We make some observations about the equivalences between regularized estimating equations, fixed-point problems and variational inequalities: (a) A regularized estimating equation is equivalent to a fixed-point problem, specified via the proximal operator of the corresponding penalty. (b) A regularized estimating equation is equivalent to a (generalized) variational inequality. Both equivalences extend to any estimating equations with convex penalty functions. To solve large-scale regularized estimating equations, it is worth pursuing computation by exploiting these connections. While fast computational algorithms are less developed for regularized estimating equation, there are many efficient solvers for fixed-point problems and variational inequalities. In this regard, we apply some efficient and scalable solvers which can deliver hundred-fold speed improvement. These connections can lead to further research in both computational and theoretical aspects of the regularized estimating equations.

Nous faisons quelques observations sur les équivalences entre les équations d'estimation régularisées, les problèmes à point fixe et les inégalités variationnelles : (a) Une équation d'estimation régularisée est équivalente à un problème à point fixe, spécifié par l'opérateur proximal de la pénalité correspondante. (b) Une équation d'estimation régularisée est équivalente à une inégalité variationnelle (généralisée). Les deux équivalences s'étendent à toutes les équations d'estimation avec des fonctions de pénalité convexes. Pour résoudre des équations d'estimation régularisées à grande échelle, il est intéressant de poursuivre le calcul en exploitant ces connexions. Bien que les algorithmes de calcul rapide soient moins développés pour les équations d'estimation régularisées, il existe de nombreux solveurs efficaces pour les problèmes de point fixe et les inégalités variationnelles. À cet égard, nous appliquons certains solveurs efficaces et évolutifs qui peuvent améliorer la vitesse de cent fois. Ces liens peuvent conduire à des recherches plus poussées sur les aspects informatiques et théoriques des équations d'estimation régularisées.

Chair/Président: Jean-François Beaumont

Organizer/Responsable: Jean-François Beaumont

Room/Salle: TB 240

Date: Monday May 29 / lundi 29 mai

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-14:00]

Keven Bosa (Statistics Canada) **Jean-François Beaumont** (Statistics Canada)

How to allocate the sample to maximize benefits from small area estimation techniques?

Comment répartir l'échantillon pour maximiser les avantages des techniques d'estimation sur petits domaines ?

The use of small area estimation techniques is common practice nowadays for National Statistics Offices to obtain reliable statistics for more domains while maintaining a reasonable cost. Typically, small area estimation is not considered when building the sampling design. It is rather used at the estimation stage to produce reliable estimates for more domains than with standard design-based methods. Sample allocation is not a new problem to survey statisticians, and many papers were published on that topic in the design-based framework. However, the literature on efficient sampling for small area estimation is sparse. During this talk, we present research results on a new sample allocation method assuming the common Fay-Herriot model is used at the estimation stage. Our proposed method is compared to that of Longford (2006) and other classical allocation methods through a simulation study where the domains coincide with the strata.

L'utilisation de techniques d'estimation sur petits domaines est une pratique courante de nos jours pour les organismes statistiques nationaux afin d'obtenir des statistiques fiables pour un plus grand nombre de domaines tout en maintenant un coût raisonnable. Généralement, l'estimation sur petits domaines n'est pas prise en compte lors de l'élaboration du plan d'échantillonnage. Elle est plutôt utilisée à l'étape d'estimation pour produire des estimations fiables pour plus de domaines qu'avec les méthodes classiques. La répartition de l'échantillon n'est pas un nouveau problème pour les statisticiens, et de nombreux articles ont été publiés sur ce sujet. Cependant, la littérature sur l'échantillonnage pour l'estimation sur petits domaines est rare. Au cours de cet exposé, nous présentons des résultats de recherche sur une nouvelle méthode de répartition de l'échantillon en supposant que le modèle de Fay-Herriot est utilisé à l'étape d'estimation. Notre méthode est comparée à celle de Longford (2006) et à d'autres méthodes classiques à travers une étude de simulation où les domaines coïncident avec les strates.

[14:00-14:30]

Abel C. Dasylva (Statistics Canada) **Jean-François Beaumont** (Statistics Canada) **Keven Bosa** (Statistics Canada) **Guillaume Maranda** (Statistics Canada)

Measuring the accuracy of a prediction for a finite population total

Mesure de l'exactitude d'une prédiction du total d'une population finie

For timeliness, a statistical agency may predict a total by modeling future responses based on past responses and fixed covariates, e.g., predicting the yield of crops based on remote sensing and agro-climatic data. Such a prediction may resort to different methodologies, from simple linear predictors to complex predictors based on machine learning techniques, including random forests,

Pour la rapidité, une agence statistique peut prédire un total en modélisant les réponses futures selon des réponses antérieures et des covariables fixes, par exemple en prédisant le rendement des cultures à l'aide de données de télédétection et d'ordre agro-climatique. Une telle prédiction peut faire appel à diverses méthodologies, des prédicteurs linéaires simples jusqu'aux prédicteurs complexes, basés sur des techniques d'apprentissage

boosted trees or deep learning. When doing so, the loss can be measured by the mean-square error to compare different predictors or report the prediction accuracy. However estimating this mean square error is challenging with complex predictors. To address this issue without being tied to a specific prediction methodology, it is proposed to bootstrap the past and future responses based on the residuals while holding the covariates fixed.

automatique, dont des forêts aléatoires, des arbres de décision optimisés ou l'apprentissage profond. De cette façon, on peut mesurer la perte par l'erreur quadratique moyenne afin de comparer différents prédicteurs ou rapporter l'exactitude de la prédiction. L'estimation de l'erreur quadratique moyenne est cependant difficile avec des prédicteurs complexes. Pour répondre à ce problème sans se contraindre à une méthodologie de prédiction spécifique, nous proposons un bootstrap des réponses passées et futures basé sur les résidus, tout en conservant les covariables fixes.

[14:30-15:00]

Marie-Hélène Toupin (Statistics Canada)

Improving the coverage of confidence intervals using degrees of freedom

Améliorer la couverture des intervalles de confiance à l'aide des degrés de liberté

For the 2021 Canadian Census of the Population, the accuracy of the estimates from the long-form questionnaire is conveyed by confidence intervals. The length of a confidence interval faithfully reflects the accuracy of the estimate provided the targeted coverage is achieved. To this end, the number of degrees of freedom must be computed as precisely as possible. The degrees of freedom linked to the variance estimation method used for the census are obtained using Satterthwaite's approximation. They are used to compute Modified Wilson confidence intervals for counts of small domains. The main goal of this presentation is to determine how computing the precise number of degrees of freedom improves the coverage of confidence intervals in comparison with the rule of thumb currently used. As we shall see, the formula obtained for the degrees of freedom provides a deeper understanding of the factors that come into play.

Pour le Recensement de la population du Canada de 2021, la précision des estimations issues du questionnaire long est véhiculée par des intervalles de confiance. La longueur d'un intervalle de confiance témoigne de la précision de l'estimation en autant que la couverture annoncée est respectée. À cette fin, le nombre de degrés de liberté associés à sa détermination en pratique doit être le plus précis possible. Les degrés de liberté reliés à la méthode d'estimation de la variance utilisée pour le recensement sont obtenus en utilisant l'approximation de Satterthwaite. Ceux-ci sont utilisés pour le calcul des intervalles de confiance de Wilson modifié pour des comptes de petits domaines. L'objectif principal de cette présentation est de déterminer dans quelle mesure l'utilisation du nombre précis de degrés de liberté améliore la couverture des intervalles de confiance en comparaison à la règle du pouce autrement employée. Comme on le verra, la formule des degrés de liberté permet d'identifier les facteurs qui sont en jeu.

Advances in spatial epidemiology
Progrès en épidémiologie spatiale

Chair/Président: Rob Deardon

Organizer/Responsable: Rob Deardon

Room/Salle: TB 208

Date: Monday May 29 / lundi 29 mai

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-13:52]

Patrick Brown (University of Toronto)

Daily air pollution and short-term health effects

Pollution atmosphérique quotidienne et effets à court terme sur la santé

Quantifying the relationship between daily changes in air pollution and short-term effects on mortality and hospitalisations is a complex task with many steps. Air pollution data contains missing values and outliers. Uncertainty in exposures should be reflected in uncertainty in effect sizes. Inference must account for the multitude of factors unrelated to air pollution that can influence mortality. Exposure-response effects are non-linear. Analyses across multiple cities and regions must take into account possible city-level variation in effects. This talk will describe ongoing work undertaken in collaboration with from Health Canada to estimate health effects from pollution in 50 Canadian cities. A key feature of the project is the case-crossover model, a form of partial likelihood which adjusts for most changes in time using control days.

Quantifier la relation entre les variations quotidiennes de la pollution de l'air et les effets à court terme sur la mortalité et les hospitalisations sont une tâche complexe avec plusieurs étapes. Les données de la pollution contiennent des valeurs manquantes ou aberrantes. L'incertitude concernant les facteurs d'exposition devrait se refléter dans l'incertitude des tailles d'effet. L'inférence doit tenir compte de la multitude de facteurs sans rapport avec la pollution pouvant influencer la mortalité. Les effets exposition-réponse sont non-linéaire. Les analyses sur plusieurs villes et régions doivent tenir compte la variation possible des effets au niveau de la ville. Cette présentation décrira les travaux en cours entrepris en collaboration avec Santé Canada pour estimer les effets sur la santé de la pollution dans 50 villes canadiennes. Une caractéristique clé du projet est le modèle de croisement de cas, une forme de probabilité qui s'ajuste à la plupart des changements de temps en utilisant des jours de contrôle.

[13:52-14:15]

Madeline Ward (University of Calgary) **Rob Deardon** (University of Calgary) **Lorna Deeth** (University of Guelph) **Caitlin Ward** (University of Minnesota)

Incorporating Behavioural Change into Infectious Disease Transmission Models

Incorporation de changements comportementaux dans les modèles de transmission de maladies infectieuses

During an epidemic, people will often modify their behaviour over time as their perceived risk of contracting or spreading the disease changes. This can have a substantial impact on the trajectory of the epidemic, however most infectious disease models assume stable population behaviour due to the challenges of modelling these changes. We will present a class of infectious disease transmission models where behavioural change is incorporated by modelling it as a function of

Pendant une épidémie, les gens modifient souvent leur comportement au fil du temps, à mesure que leur perception du risque de contracter ou propager la maladie change. Ce facteur peut avoir un impact important sur la trajectoire de l'épidémie ; cependant, la plupart des modèles de maladie infectieuse supposent un comportement stable de la population en raison de la difficulté de modéliser ces changements. Nous présentons une classe de modèles de transmission de maladie infectieuse dans lesquels le changement comportemental est incorporé en le modélisant

a direct epidemic metric, such as prevalence. We consider several different mechanisms by which the effect of behaviour changes may vary according to these metrics, and results from modelling on both the population-averaged and individual levels will be shown. Model fitting in a Bayesian framework will be illustrated through simulation studies and applications to data from foot and mouth disease and COVID-19.

comme une fonction d'une mesure directe de l'épidémie, telle que la prévalence. Nous considérons plusieurs mécanismes différents par lesquels l'effet des changements de comportement peut varier en fonction de ces mesures et montrons aussi les résultats de la modélisation au niveau de la moyenne de la population et au niveau individuel. L'ajustement du modèle dans un cadre bayésien sera illustré par des études par simulation et des applications à des données sur la fièvre aphteuse et la COVID-19.

[14:15-14:38]

Dirk Douwes-Schultz (McGill) **Alexandra Schmidt** (McGill University) **Yannan Shen** (McGill University) **David Buckeridge** (McGill University)

A Three-state Coupled Markov Switching Model for COVID-19 Outbreaks across Quebec based on Hospital Admissions

Un modèle de Markov à changement entre trois états pour étudier les épidémies de COVID-19 au Québec mesurées par les hospitalisations

Recurrent COVID-19 outbreaks have placed immense strain on the hospital system in Quebec. We develop a Bayesian three-state coupled Markov switching model to analyze COVID-19 outbreaks across Quebec based on admissions in the 30 largest hospitals. Within each catchment area we assume the existence of three states for the disease: absence, a new state meant to account for many zeroes in some of the smaller areas, endemic and outbreak. Then we assume the disease switches between the three states in each area through a series of coupled nonhomogeneous hidden Markov chains. Unlike previous approaches, the transition probabilities may depend on covariates and the occurrence of outbreaks in neighboring areas, to account for geographical outbreak spread. We make some interesting findings such as that mobility in retail and recreation venues had a strong positive association with the development and persistence of new COVID-19 outbreaks in Quebec. Based on model comparison our contributions show promise in improving state estimation retrospectively and in real-time, especially when there are smaller areas and highly spatially synchronized outbreaks, and they offer new and interesting epidemiological interpretations.

Au Québec, les épidémies récurrentes de COVID-19 ont mis à rude épreuve le système hospitalier. Nous développons un modèle bayésien de Markov avec changement entre trois états pour analyser les épidémies de COVID-19 à travers le Québec en fonction des hospitalisations dans les 30 plus grands hôpitaux. Dans chaque circonscription (région), nous supposons que trois états de la maladie sont possibles : absence, un nouvel état censé représenter les nombreux zéros dans certaines des régions les plus petites, endémique et épidémique. Ensuite, nous supposons que la maladie bascule entre les trois états dans chaque région via une série de chaînes de Markov cachées couplées non-homogènes. Contrairement aux approches précédentes, les probabilités de transition peuvent dépendre de covariables et de l'apparition d'épidémies dans les régions voisines, afin de tenir compte de la propagation géographique des épidémies. Nous notons que la mobilité dans les lieux de loisir et de vente au détail est fortement associée à l'apparition et à la persistance de nouvelles épidémies de COVID-19 au Québec. Selon une comparaison de performances de modèles, nos contributions semblent prometteuses pour l'amélioration de l'estimation de l'état de la maladie, à la fois rétrospectivement et en temps réel.

[14:38-15:00]

Longhai Li (University of Saskatchewan)

Cross-validatory Residual Diagnostics for Bayesian Spatial Models

Diagnostics résiduels à validation croisée pour les modèles spatiaux bayésiens

Residual diagnostics is very important for frequentist normal regression modelling. We can check the overall goodness-of-fit (GOF) of a fitted model by checking the normality of residuals with QQ plots and statis-

Les diagnostics résiduels sont très importants pour la modélisation de régression normale fréquentiste. On peut vérifier l'adéquation (GOF) générale d'un modèle ajusté en vérifiant la normalité des résidus à l'aide de diagrammes QQ et de tests statistiques. Ce-

tical tests. However, residual diagnostic tools are not available for Bayesian models. In this talk, we propose a method to define residuals for checking Bayesian models, which is called Z-residual. The Z-residual is transformed from the cross-validatory randomized predictive p-values (RSP) with the normal quantile function. We show that the RSP has a uniform distribution on $(0, 1)$ when the distributions for the likelihood and the prior are correctly specified. Due to the uniformity of the RSPs, Z-residuals are normally distributed under the true model. Therefore, we can use the Z-residual to conduct residual diagnostics for Bayesian models as for normal regression. Applying Z-residual diagnostics to a Poisson disease mapping model with spatial effects, we demonstrate that the graphical diagnostics based on Z-residuals can effectively identify the mis-specification in the distributional family for the response variable. We also investigate the sizes and powers of statistical tests based on the Z-residual with large-scale simulation studies.

pendant, les outils de diagnostic résiduel ne sont pas disponibles pour les modèles bayésiens. Dans le cadre de cet exposé, nous proposons une méthode pour définir les résidus afin de vérifier les modèles bayésiens, que l'on appelle ≪résidu Z≫. Le résidu Z est transformé à partir des valeurs P prédictives randomisées à validation croisée (RSP) au moyen de la fonction quantile normale. Nous démontrons que le RSP est doté d'une distribution uniforme sur $(0, 1)$ lorsque les distributions pour la vraisemblance et l'a priori sont spécifiées correctement. En raison de l'uniformité des RSP, les résidus Z sont distribués normalement sous le vrai modèle. Conséquemment, nous pouvons utiliser le résidu Z pour mener des diagnostics résiduels pour les modèles bayésiens tout comme pour une régression normale. En appliquant le diagnostic de résidu Z à un modèle de cartographie de maladie de Poisson avec effets spatiaux, nous démontrons que le diagnostic graphique basé sur les résidus Z peut repérer efficacement les erreurs de spécification dans la famille distributionnelle pour la variable de réponse. Nous étudions aussi les tailles et les puissances des tests statistiques basés sur le résidu Z à travers des études de simulations de grande envergure.

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

Chair/Président: Wesley S. Burr

Organizer/Responsable: Wesley S. Burr

Room/Salle: RB 2200

Date: Monday May 29 / lundi 29 mai

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-15:00]

Alison L. Gibbs (University of Toronto)

Teaching Statistics through Changing Times: Lessons from my mother and others

Enseigner la statistique d'une époque à une autre : enseignements de ma mère et d'autres

From the emergence of data science to the widespread availability of generative artificial intelligence, it often feels like change is accelerating. What does this mean for teachers of statistics? And how do we prepare our students for what's coming next? I'll share some retrospective, prospective and introspective thinking on these questions. While much is changing, there is much that persists.

De l'émergence de la science des données à la disponibilité généralisée de l'intelligence artificielle générative, on a souvent l'impression que le changement s'accélère. Qu'est-ce que cela signifie pour les enseignants en statistique ? Et comment préparer nos étudiants à ce qui les attend ? Je vais partager avec vous quelques réflexions rétrospectives, prospectives et introspectives sur ces questions. Beaucoup de choses ont changé, mais bien d'autres persistent.

Advanced Models in Finance and Insurance
Modèles avancés en finance et en assurance

Chair/Président: Hanna K Jankowski

Room/Salle: TB 206

Date: Monday May 29 / lundi 29 mai

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-13:45]

Ismael Afolabi ASSANI (Université de Montréal) **Maciej Augustyniak** (Université de Montréal)

Optimal Quadratic Hedging with Basis Risk under Bivariate Q-Affine GARCH Models

Couverture quadratique sous risque de base dans des modèles GARCH Q-affines

This paper investigates quadratic hedging of financial derivatives under basis risk within a bivariate class of Q-affine GARCH models. The hedging strategy is constructed using an asset whose return and volatility dynamics are correlated with those of the underlying security in the derivative product. We derive a semi-explicit risk-minimization hedging formula that is applicable to European derivatives, such as calls and puts, with payoff functions that admit an inverse Laplace transform representation. We conduct numerical experiments to assess the effectiveness of our formula and the performance of our hedging strategy. Our findings shed light on the potential of our approach for hedging derivatives under basis risk.

Cet article étudie la couverture quadratique de produits dérivés financiers sous risque de base dans une classe de modèles Q-affine GARCH bivariée. La stratégie de couverture est construite à l'aide d'un actif dont les dynamiques du rendement et de la volatilité sont corrélées avec celles de la sécurité sous-jacente du produit dérivé. Nous dérivons une formule de couverture semi-explicite de minimisation du risque quadratique applicable aux dérivés européens, tels que les options d'achat et de vente, dont les versements admettent une représentation en forme de transformée inverse de Laplace. Des expériences numériques ont été menées pour évaluer l'efficacité et les performances de notre stratégie de couverture. Nos résultats mettent en lumière le potentiel de notre approche pour la couverture de produits dérivés sous risque de base.

[13:45-14:00]

Juan-Sebastian Yanez (Université du Québec à Montréal) **Mathieu Pigeon** (Université du Québec à Montréal)

Dependence in Claim Processing Time: A Frailty Analysis Perspective

Dépendance dans le temps de traitement des réclamations : une perspective basée sur l'analyse de la fragilité

When a micro-level loss reserving model is considered, instead of modelling the aggregated data from a loss triangle, the development of each claim is modelled directly. Often, these models suppose that each claim is independent from one another, however for a variety of reasons a data set could contain dependent claims. We take an interest in claims that affect various distinct coverages with their own development and even their own covariate information. When considered separately the issue of dependence arises because of their shared origin. We take into account of this dependence for the duration of claims by considering a shared frailty random effect model. Furthermore, the aforementioned model is ideal in a loss-reserving context because the conditional

Lorsqu'un modèle de provisionnement de pertes au niveau micro est considéré, au lieu de modéliser les données agrégées à partir d'un triangle de perte, l'évolution de chaque réclamation est modélisée directement. Souvent, ces modèles supposent que les réclamations sont indépendantes, mais pour diverses raisons, une base de données peut contenir des réclamations dépendantes. Nous nous intéressons aux réclamations qui affectent diverses couvertures distinctes avec leur propre développement et même leur propre information au niveau des covariables. Lorsqu'elles sont considérées séparément, la question de la dépendance se pose en raison de leur origine commune. Nous tenons compte de cette dépendance sur la durée des réclamations en considérant un modèle à effets aléatoires de fragilité partagée. De plus, le modèle susmentionné est idéal dans un contexte de provisionne-

distribution of claim duration given its past development can be found in a closed form. We also evaluate our model by fitting it into a data set, while also discussing estimation procedures and analyzing the results

ment de pertes car la distribution conditionnelle de la durée des réclamations compte tenu de son évolution passée peut se trouver sous une forme fermée. Nous évaluons également notre modèle en l'ajustant à un ensemble de données, tout en discutant des procédures d'estimation et en analysant les résultats.

[14:00-14:15]

Hashan Savinda Peiris Kalugama Gardige (Simon Fraser University)

Integration of Traditional and Telematics data for Efficient Insurance Claims Prediction

Intégration de données télématiques et traditionnelles pour une prédiction efficace des 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 concern or adverse selection compared to the data points with traditional features. To handle this issue, we propose a data integration technique based on calibration weights. It is shown that the proposed technique can efficiently integrate the so-called traditional data and telematics data and also cope with possible adverse selection issues on the availability of telematics data. Our findings are supported by a simulation study and empirical analysis in a synthetic telematics data set. Keywords: Adverse selection, Automobile insurance, Data integration, Driver telematics, Missing data analysis

Alors que la télématique pour les conducteurs a attiré de l'attention concernant la classification de risque dans l'assurance automobile, le manque d'observations relatives aux caractéristiques télématiques pose problème. Les enjeux de confidentialité ou une antisélection par rapport au point de données avec les caractéristiques traditionnelles pourraient expliquer cette lacune. Pour aborder le problème, nous proposons une technique d'intégration de données basée sur les poids de calibration. Nous démontrons que la technique proposée peut intégrer efficacement les données traditionnelles et celles télématiques et peut aussi gérer les problèmes potentiels d'antisélection dans la disponibilité des données télématiques. Nos découvertes sont soutenues par une étude de simulation et des analyses empiriques dans un jeu de données télématiques synthétiques. Mots clés : antisélection, assurance automobile, intégration de données, télématique pour les conducteurs, analyse de données manquantes.

[14:15-14:30]

Philipp Ratz (Université du Québec à Montréal)

Solving censored regression problems using a multitask approach

Données censurées sous l'angle d'une regression multitâche

Censored observations arise in many disciplines such as economics and medicine. There exists a stack of well-established parametric estimators that work well on regular and aggregated data but less so on individual prediction problems. More recently, algorithms for predictive survival problems were proposed, but they often favour predictive accuracy over interpretability and calibration. To bridge the gap, we show that common survival estimators can be reformulated into problems that involve solving multiple related "tasks". This reformulation then allows to leverage techniques from the field of machine learning which work well in high dimensions, without sacrificing the interpretability of the results completely. Deriving the estimators from known survival models also permits the use of existing techniques to deal with common issues such as dependent censoring. We present two applications where we es-

Les observations censurées apparaissent dans de nombreuses disciplines telles que l'économie et la médecine. Un nombre d'estimateurs paramétriques, bien établis dans la littérature, existent dans la boîte à outils statistique. Ces méthodes sont généralement performantes sur les données agrégées et régulières, mais moins sur les prédictions individuelles. Plus récemment, des algorithmes de l'apprentissage statistique ont été proposés pour les problèmes de prévision, mais souvent ils favorisent la précision prédictive au détriment de l'interprétabilité et de la calibration. Pour combler cette lacune, nous montrons que des estimateurs de survie courants peuvent être reformulés en problèmes qui tentent de résoudre plusieurs tâches reliées à la fois. Cette reformulation permet ensuite de se servir des techniques du domaine de l'apprentissage statistique qui sont performantes en haute dimension, sans sacrifier complètement l'interprétabilité des résultats. Nous présentons deux applications dans lesquelles nous estimons les courbes de survie individuelles à partir de données censurées et prédisons les

timate individual survival curves and predict survival quantiles under censoring.

quantiles de survie sous censure.

[14:30-14:45]

Sulalitha Bowala Mudiyanselage (University of Manitoba)

Daily and Hourly Fuzzy Forecasts of Electricity Demand Using Neural Network Dynamic Regression Models

Prévisions horaires et quotidiennes floues de la demande en électricité à l'aide de modèles de régression dynamique par réseaux neuronaux

The goal of accurate electricity demand forecasts is to fulfill the electricity load requirement of users. In the literature, both statistical and machine learning techniques have been applied to obtain accurate electricity demand forecasts by incorporating features such as temperature and calendar characteristics. The novelty of this paper is to obtain daily and hourly point and fuzzy forecasts by extending the recently proposed hybrid forecast models (dynamic regression models with the Seasonal Autoregressive Integrated Moving Average errors (DRSARIMA) and the hybrid neural network dynamic regression (NNDR) model). The demand forecasting scheme incorporates several features, such as temperature, cloud cover, pressure, wind direction, humidity, and day type, to improve the efficiency of the forecasts. Long-term point forecasts and innovations are used to obtain two classes of fuzzy prediction intervals (PIs) using data-driven probabilistic innovation distribution and bootstrapping for DRSARIMA and NNDR. The experimental results show that fuzzy forecasts are better using NNDR for daily data, whereas, for hourly data, fuzzy forecasts are better using DRSARIMA. Furthermore, the data-driven approaches using nonlinear adaptive fuzzy numbers lead to high coverage probabilities under both models.

Le but de prévoir avec exactitude la demande en électricité est de satisfaire à l'exigence de charge électrique des utilisateurs. La littérature fait état de l'application à la fois de techniques d'apprentissage statistique et automatique pour obtenir des prévisions exactes de la demande en électricité en incorporant des éléments comme des caractéristiques de température et de calendrier. La nouveauté dans cet article est d'obtenir des prévisions journalières et quotidiennes ponctuelles et floues en étendant les modèles hybrides de prévisions proposés récemment (les modèles de régression dynamique avec erreurs de la moyenne mobile intégrée saisonnière autorégressive (DRSARIMA) et le modèle hybride de régression dynamique par réseaux neuronaux (NNDR)). Afin d'améliorer l'efficacité des prévisions, le système de prévision de la demande incorpore plusieurs éléments, comme la température, la couverture nuageuse, la pression, la direction du vent, l'humidité et le type de journée. Des prévisions ponctuelles et innovations à long terme sont utilisées pour obtenir deux classes d'intervalles de prédictions floues (PI) à l'aide d'une distribution probabiliste de l'innovation et d'une méthode bootstrap fondées sur les données pour les DRSARIMA et le NNDR. Les résultats expérimentaux montrent que l'on obtient de meilleures prévisions floues avec le modèle NNDR pour les données quotidiennes et avec la DRSARIMA pour les données horaires. Par ailleurs, les approches basées sur les données utilisant des nombres non linéaires adaptatifs flous donnent des probabilités de couverture élevées sous les deux modèles.

Missing Data and Patient-Focused Statistical Methods
Données manquantes et méthodes statistiques axées sur le patient

Chair/Président: Janie Coulombe

Room/Salle: TB 202

Date: Monday May 29 / lundi 29 mai

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-13:45]

Abdoulaye Dioni (Université Laval)

Sensitivity analysis in the presence of a missing not at random categorical explanatory variable

Analyse de sensibilité en présence d'une variable explicative, catégorielle manquant non aléatoirement

Background: missing data are common in health care and their analysis depends on the mechanism by which they occur. The most problematic mechanism is MNAR (Missing Not At Random) because it requires an additional assumption not contained in the data. Objective: the aim is to propose a sensitivity analysis method based on multiple imputation that describes the possible deviation from MAR (Missing At Random) in the presence of an ordinal explanatory variable for which some categories tend to be missing. Method: the proposed approach consists of four steps. 1) performing a multiple imputation using existing packages. 2) performing a latent variable transformation to introduce the additional information 3) performing the final analysis on the modified data. 4) applying Rubin's rule. Application and Results: the application is done on the simulated and real data. The real data was extracted form the trauma registry of Canada between 2013 and 2018 and includes patients over 16 years old.

Contexte : les données manquantes sont courantes dans la santé et l'analyse dépend du mécanisme par lequel elles apparaissent. Le plus problématique est MNAR (Missing Not At Random) car il nécessite une hypothèse supplémentaire non contenue dans les données. Objectif : le but est de proposer une méthode d'analyse de sensibilité basée sur l'imputation multiple qui décrit la possible déviation du MAR (Missing At random) en présence d'une variable explicative ordinale dont certaines catégories ont tendance à être manquantes. Méthode : l'approche comprend quatre étapes. 1) faire une imputation multiple grâce aux modules existants. 2) faire une transformation de variable latente pour introduire l'information supplémentaire. 3) faire l'analyse finale sur les données modifiées. 4) appliquer la règle de Rubin. Application et Résultats : l'application est faite sur les données simulées et réelles. Les données réelles portent sur des patients âgés de plus de 16 ans, hospitalisés suite à un traumatisme au Canada entre 2013 et 2018.

[13:45-14:00]

Md. Erfanul Hoque (Thompson Rivers Univeristy) **Mahmoud Torabi** (University of Manitoba) **Elif Acar** (University of Manitoba)

A Heterogeneous Random Effects Covariance Matrix in Longitudinal Data with Missing Responses and Mismeasured Covariates

Matrice de covariance des effets aléatoires hétérogènes pour données longitudinales avec des réponses manquantes et des covariables mal mesurées

Longitudinal data often contain missing responses and mismeasured covariates. While generalized linear mixed model (GLMM) frameworks are commonly adopted to simultaneously analyze such data, the random effect covariance matrix in these models is often assumed constant across the subjects, and is restricted

Les données longitudinales contiennent souvent des réponses manquantes et des covariables mal mesurées. Alors que les modèles linéaires mixtes généralisés (GLMM) sont couramment adoptés pour analyser simultanément de telles données, la matrice de covariance des effets aléatoires dans ces modèles est souvent supposée constante entre les sujets, et est restreinte en raison de

because of its high dimensionality and its positive definite constraint. However, the random effect covariance matrix may vary by measured covariates in many situations and ignoring this heterogeneity can lead to biased estimates of model parameters. In this talk, we present an approach to incorporate potential heterogeneity in the random effects covariance matrix. The proposed approach uses a modified Cholesky decomposition and allows the random effects covariance matrix to depend on covariates. The performance of the proposed approach is evaluated through simulation studies, and is demonstrated using longitudinal data from Framingham Heart Study

sa haute dimensionnalité et de sa contrainte de définition positive. Cependant, cette matrice peut varier en fonction des covariables mesurées dans de nombreuses situations et le fait d'ignorer cette hétérogénéité peut conduire à des estimations biaisées des paramètres du modèle. Dans cet exposé, nous présentons une approche permettant d'intégrer l'hétérogénéité potentielle de la matrice de covariance des effets aléatoires. L'approche proposée utilise une décomposition de Cholesky modifiée et permet à la matrice de dépendre des covariables. Nous évaluons la performance de l'approche proposée par des études de simulation et la démontrons à l'aide de données longitudinales tirées de l'étude Framingham Heart Study.

[14:00-14:15]

Chendong Li (University of Manitoba) **Depeng Jiang** (University of Manitoba) **Robert Tate** (University of Manitoba) **Philip St. John** (University of Manitoba)

Modeling developmental trajectories with nonrandomly missing data: Investigating trajectories of frailty using data from the Manitoba Follow-up Study

Modélisation de trajectoires de développement avec des données manquantes de manière non aléatoire : enquête sur les trajectoires de fragilité à l'aide de données de l'étude Manitoba Follow-up Study

Frailty trajectory analysis studies the change in frailty over time and relies on longitudinal data. However, missing data is a major challenge, especially when it is non-ignorable. The conventional group-based trajectory analysis (GBTA) can handle missing data when it is missing completely at random or missing at random. However, when data is missing not at random, the extended GBTA can estimate attrition probability and provide unbiased estimates. This thesis aims to compare the performance of conventional and extended GBTA in different missing data patterns through simulations and apply them to frailty trajectory analysis using data from the Manitoba Follow-up Study (MFUS). The MFUS data will be used to analyze frailty trajectory and its association with living arrangements and marital status. This study will contribute to applied researchers by providing a solution to handle nonrandom missing data and help better understand the heterogeneity of frailty over time.

L'analyse de trajectoires de fragilité permet d'étudier le changement de la fragilité au fil du temps et s'appuie sur des données longitudinales. Cependant, les données manquantes posent un problème important, en particulier si elles sont non ignorables. L'analyse conventionnelle de trajectoires basées sur le groupe (GBTA) peut traiter les données manquantes si elles sont manquantes de manière complètement aléatoire ou manquantes de manière aléatoire. Cependant, s'il s'agit de données manquantes de manière non aléatoire, la GBTA étendue peut estimer la probabilité d'attrition et fournir des estimations sans biais. Cette thèse vise à comparer les performances d'une GBTA conventionnelle à celle d'une GBTA étendue dans divers modèles de données manquantes à l'aide de simulations et de les appliquer à une analyse de trajectoires de fragilité en utilisant les données de l'étude Manitoba Follow-up Study (MFUS). Les données de cette étude seront utilisées pour l'analyse de trajectoires de fragilité et leur association avec le mode de vie et l'état matrimonial. Cette étude servira en recherche appliquée en apportant une solution pour traiter les données manquantes de manière non aléatoire et aidera à mieux comprendre l'hétérogénéité de la fragilité au fil du temps.

[14:15-14:30]

Marc Angelo Parsons (McGill University) **Andrea Benedetti** (McGill University) **Russell Steele** (McGill University)

Flexible Methods for Trajectory Estimation in Individual Participant Data Meta-Analyses

Méthodes flexibles pour l'estimation de la trajectoire dans les méta-analyses de données individuelles sur les patients

Individual participant data meta-analysis (IPDMA) involves meta-analyzing primary patient data gathered

La méta-analyse des données individuelles des participants (MADIP) consiste à méta-analyser les données primaires des pa-

from multiple studies in a systematic review. These analyses may be of higher quality and more reliable than standard aggregate meta-analysis (MA). Various methods have been proposed to estimate trajectories of a continuous outcome along an exposure in standard MA. Fewer, however, have been proposed in an IPDMA context. These include both one- and two-stage approaches using a variety of flexible regression methods such as fractional polynomials or splines. In this study, the results of flexible regression approaches applied to an individual patient dataset (N=9 primary studies) examining depression scores in pregnant and postpartum will be presented. The goal of the analysis is to estimate trajectories of depression during the perinatal period. Extensions of existing MA methods to IPDMA will be presented and discussed. Measures of model fit will be used to compare between methods.

tients recueillies à partir de plusieurs études dans une revue systématique. Ces analyses peuvent être de meilleure qualité et plus fiables que la méta-analyse agrégée (MA). Diverses méthodes ont été proposées pour estimer les trajectoires d'une variable en MA. Cependant, moins de méthodes ont été proposées dans le contexte MADIP. Celles-ci incluent une variété de méthodes de régression flexibles telles que des polynômes fractionnaires ou des splines. Dans cette étude, les approches de régression flexibles appliquées à une base de données provenant de neuf études primaires sur la dépression chez les femmes seront comparées. Le but de l'analyse est d'estimer les trajectoires de la dépression au cours de la période périnatale. Des extensions des méthodes existantes seront présentées et discutées. Des mesures de l'ajustement du modèle seront utilisées pour comparer les méthodes.

[14:30-14:45]

William Ruth (Simon Fraser University) **Richard Lockhart** (Simon Fraser University)

A Monte Carlo EM Analysis of COVID-19 Outbreaks in Long-Term Healthcare Facilities

Analyse espérance-maximisation (EM) Monte-Carlo des éclosions de la COVID-19 dans les centres de soins de longue durée

We analyze a dataset of daily case counts from COVID-19 outbreaks in long-term care facilities across BC, Canada. We treat infection durations as missing data and apply the EM algorithm to estimate the probability of transmission and mean infection duration, as well as a third parameter governing the effectiveness of outbreak management. We use the Monte Carlo EM (MCEM) algorithm to approximate difficult conditional expectations with Monte Carlo integrals. Importance sampling is used to generate the necessary Monte Carlo samples. We propose a novel stopping rule for the MCEM algorithm based on a well-known relationship between the score vectors of the observed and complete data models. Finally, we investigate how outbreak parameters vary across facilities based on the values of some facility-level covariates.

Nous analysons un ensemble de données sur les nombres quotidiens de cas d'infection causés par les éclosions de la COVID-19 dans des centres de soins de longue durée en Colombie-Britannique, au Canada. Nous traitons la durée d'infection comme une donnée manquante et appliquons l'algorithme EM pour estimer la probabilité de transmission, la durée moyenne d'infection et un troisième paramètre régissant l'efficacité de la gestion des éclosions. L'algorithme EM Monte-Carlo nous sert aussi à faire une approximation des espérances conditionnelles difficiles à faire avec l'intégration de Monte-Carlo. L'échantillonnage préférentiel est utilisé pour générer les échantillons Monte-Carlo nécessaires. Nous proposons une nouvelle règle d'arrêt pour l'algorithme EM Monte-Carlo basée sur une relation bien connue entre les vecteurs de score dans les modèles de données observées et complètes. Enfin, nous étudions les variations des paramètres d'éclosion en fonction des valeurs de certaines covariables au niveau des établissements.

[14:45-15:00]

Hensley Hubert Mariathas (Memorial University of Newfoundland) **Shabnam Asghari** (Memorial University of Newfoundland) **Oliver Hurley** (Memorial University of Newfoundland)

Sample size calculation for stepped wedge cluster randomized trial when the number of clusters available is limited to small

Calcul de la taille d'échantillon pour les essais randomisés en grappes avec permutation séquentielle lorsque le nombre de grappes est limité ou faible

The stepped-wedge cluster randomized trial (SW-CRT)

L'essai randomisé en grappes avec permutation séquentielle (SW-

is a novel study design that is increasingly being utilized to evaluate healthcare services. This design typically requires a relatively small number of clusters. However, in certain situations, the number of clusters available for randomization is often limited, which can create uncertainty regarding the statistical power necessary to draw conclusions about the intervention effect. In this presentation, we analyzed the effectiveness of three multilevel model assumptions, which included: a model that accounted for variation between clusters and times within clusters, a model that assumed multiple measurement periods outside roll-out, and a model that had no observations outside roll-out. The last two approaches assume no within-cluster time effect. These methods differ in the choice of within and between cluster correlation structures. We compared the performance of these three approaches for multiple measurement periods in terms of intra-cluster correlation coefficients (ICC), where we assumed observations are equally correlated within clusters. The approaches are illustrated for normally distributed continuous outcomes using two different scenarios, depending on whether an equal number of observations are measured at different time periods or steps where different participants are observed.

CRT) est un nouveau plan d'étude qui est de plus en plus employé pour évaluer les services de soins de santé. Ce plan demande généralement un nombre relativement petit de grappes. Cependant, dans certaines situations, le nombre de grappes disponibles pour la randomisation est souvent limité, ce qui peut causer de l'incertitude concernant la puissance statistique nécessaire pour obtenir des conclusions relatives à l'effet d'intervention. Dans cette présentation, nous avons analysé l'efficacité de trois hypothèses de modèles multiniveaux : un modèle qui tient compte de la variation entre les grappes et le temps à l'intérieur des grappes, un modèle qui suppose de multiples périodes de mesure dehors du déploiement, et un modèle qui n'a pas d'observation en dehors du déploiement. Les deux dernières approches supposent une absence d'effet temporel à l'intérieure des grappes. Ces méthodes diffèrent dans le choix des structures de corrélation intragrappes et intergrappes. Nous avons comparé la performance de ces trois approches pour plusieurs périodes de mesure en termes de coefficients de corrélation intragrappes (ICC), où nous avons supposé que les observations seraient corrélées de façon identique à l'intérieur des grappes. Les approches sont illustrées pour des résultats continus distribués normalement à l'aide de deux différents scénarios, en fonction qu'un nombre égal d'observations soit mesuré à différentes périodes de temps ou étapes dans lesquelles on observe différents participants.

Recent Developments in models for event-history, case-cohort and mixed-type trait data
Développements récents dans les modèles pour données d'histoire d'événements, de cohortes de cas et de traits de type mixte

Chair/Président: Celia M.T. Greenwood

Room/Salle: TB 236

Date: Monday May 29 / lundi 29 mai

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-13:45]

Marcus Hlady (University of Manitoba) **Yuliya V. Martsynyuk** (University of Manitoba)

Importance of Reliability Ratio for Properties of Asymptotic Confidence Intervals in Linear Structural Errors-in-Variables Models

Importance du rapport de fiabilité pour les propriétés des intervalles de confiance asymptotiques dans les modèles structurels linéaires d'erreurs dans les variables

We revisit linear structural errors-in-variables models with univariate observations and study finite-sample properties of several asymptotic confidence intervals (CI) for the slope and intercept that are based on modified least squares estimators under some most general moment conditions on the explanatory variables and measurement errors. Assessment and comparison of finite-sample properties of these CI's are done via computing their empirical finite-sample coverage probabilities and expected lengths for a variety of distributions of the explanatory and error variables, different values of the reliability ratio and several sample sizes. We conclude that larger values of the reliability ratio generally yield better finite-sample performance of these CI's. Special attention is devoted to the case of the reliability ratio taking its maximal value of one, when the explanatory variables are in the domain of attraction of the normal law with an infinite variance.

Nous revoyons les modèles structurels linéaires d'erreurs dans les variables avec des observations multivariées et étudions les propriétés pour des échantillons de taille finie de plusieurs intervalles de confiance (CI) pour une pente et une ordonnée à l'origine basées sur des estimateurs des moindres carrés modifiés sous certaines conditions de moments les plus généralisés sur les variables explicatives et les erreurs de mesure. L'évaluation et la comparaison des propriétés pour des échantillons de taille finie de ces CI se font par computation de leur probabilité empirique de couverture et de leur longueur espérée pour diverses distributions des variables explicatives et des erreurs, pour différentes valeurs du rapport de fiabilité et pour plusieurs tailles d'échantillon. Nous concluons que de grandes valeurs du rapport de fiabilité fournissent généralement une meilleure performance pour des échantillons de taille finie de ces intervalles de confiance. Nous portons une attention particulière au cas où le rapport de fiabilité prend sa valeur maximale, qui est égale à un, lorsque les variables explicatives sont dans le domaine d'attraction de la loi normale avec une variance infinie.

[13:45-14:00]

Jacqueline A. May (University of Guelph) **Zeny Feng** (University of Guelph) **Sarah J. Adamowicz** (University of Guelph)

A real data-driven simulation strategy to select an imputation method for mixed-type trait data

Stratégie de simulation pilotée par des données réelles pour la sélection d'une méthode d'imputation pour des données de traits de type mixte

Missing values in trait datasets pose an obstacle for analyses in myriad biological disciplines. We invoked a real data-driven simulation strategy to select an imputation method for a given mixed-type trait dataset. Using a reptile trait dataset, a near complete-case dataset was formed for method selection. Values were removed un-

Les valeurs manquantes dans les ensembles de données de traits constituent un obstacle pour les analyses dans une multitude de disciplines biologiques. Nous faisons appel à une stratégie de simulation pilotée par des données réelles pour la sélection d'une méthode d'imputation pour un ensemble déterminé de données de traits de type mixte. À l'aide d'un ensemble de données de

der different missing mechanisms: missing completely at random, missing at random, and missing not at random. Method combinations with and without phylogeny were used to impute the missing values. A random forest method with nuclear phylogeny performed best and was used to impute missing values in the original dataset. Imputed data better reflected the original data characteristics compared to complete case. Caution should be taken when imputing trait data as phylogeny did not improve performance in every scenario. These results support use of a real data-driven simulation strategy for selecting an imputation method for a given mixed-type trait dataset.

traits de reptile, un ensemble de données presque complet a été formé pour la sélection d'une méthode. Des valeurs ont été retirées sous différents mécanismes d'omission : omission de manière complètement aléatoire, omission de manière aléatoire et omission de manière non aléatoire. Des combinaisons de méthodes avec ou sans phylogénie ont été utilisées pour l'imputation des valeurs manquantes. La meilleure performance a été celle de la méthode de forêt aléatoire avec une phylogénie nucléaire, laquelle a été utilisée pour l'imputation des valeurs manquantes dans l'ensemble de données original. Les données imputées ont mieux reflété les caractéristiques des données originales par comparaison avec celles par cas complet. La prudence est de mise dans l'imputation des données de trait, car la phylogénie n'a pas amélioré la performance dans tous les scénarios. Ces résultats appuient l'utilisation d'une stratégie de simulation pilotée par des données réelles pour la sélection d'une méthode d'imputation pour un ensemble déterminé de données de traits de type mixte.

[14:00-14:15]

Hameed A. Jimoh (University of Saskatchewan) **Shahedul Khan** (University of Saskatchewan)

The proportional hazards assumption for joint models of longitudinal and time-to-event data

Hypothèse de risques proportionnels pour les modèles conjoints de données longitudinales et de temps avant l'événement

In follow-up studies, both fixed and time-dependent covariates are often available along with an observation on the time to an event of interest. A time-dependent covariate can also be internal (generated by the unit under study and measured longitudinally over time) or external (changes its value due to external characteristics). The joint modeling approach works in two steps: (a) modeling the internal covariate taking into account measurement error to construct its complete history, and (b) quantifying the effects of the covariates on time-to-event taking into account the history as constructed in (a). Combining the two processes enables us to mutually borrow information from each process and gain efficiency in statistical inference. The standard approach for joint modeling is to assume a proportional hazards (PH) model to describe the event process, where the covariates act multiplicatively on the hazard function. However, the assumption of PH is strong, and it is important that it be checked. We propose a method to check the PH assumption for joint models, developed in a Bayesian framework. The proposed methodology is demonstrated using both simulated and real-world data.

Dans les études de suivi, les covariables fixes et celles dépendantes du temps accompagnent souvent une observation sur le temps avant un événement d'intérêt. Une covariable dépendante du temps peut aussi être interne (générée par l'unité étudiée et mesurée longitudinalement avec le temps) ou externe (des caractéristiques externes en changent la valeur). L'approche de modélisation conjointe fonctionne en deux étapes : (a) la modélisation de la covariable interne en prenant en compte l'erreur de mesure pour construire son histoire intégrale, et (b) la quantification des effets des covariables sur le temps avant l'événement en prenant en compte l'histoire telle qu'elle est construite à l'étape (a). La combinaison des deux processus permet l'emprunt mutuel d'information de chacun des deux et une plus grande efficacité de l'inférence statistique. L'approche normative de la modélisation conjointe suppose un modèle à risques proportionnels (RP) pour décrire le processus événementiel, tandis que l'action des covariables sur la fonction de risque est multiplicative. Cependant, l'hypothèse de RP est forte et il est important de la vérifier. Nous proposons une méthode de vérification de l'hypothèse de risques proportionnels dans les modèles conjoints développée dans un cadre bayésien. Des données simulées et réelles servent à illustrer la méthodologie proposée.

[14:15-14:30]

Christian Chan (University of Calgary) **Xuewen Lu** (University of Calgary) **Xiaotian Dai** (University of Calgary) **Thierry Chekouo** (University of Minnesota) **Quan Long** (University of Calgary)

Broken Adaptive Ridge Method for Variable Selection in Logistic Partly Linear Models with application to CATHGEN data

Méthode de crête adaptative brisée pour la sélection de variables dans les modèles logistiques partiellement linéaires avec application aux données CATHGEN

We develop a new method for estimation and variable selection in logistic partly linear models for high-dimensional data. The method is named the broken adaptive ridge estimator, which approximates the L0-regularization by iteratively performing reweighted squared L2-regularization. The logistic partly linear model extends the logistic linear model by adding a nonparametric component to it and becomes a flexible model for modeling various types of covariate effects. We use the Bernstein polynomials as the sieve space to approximate the nonparametric functions of some low-dimensional covariates. Our method is implemented easily using existing R packages. Extensive simulation studies suggest the proposed method outperforms other commonly used penalty-based variable selection methods. We apply the method to a coronary artery disease study, which motivated our research, and report new findings in both high-dimensional genetic and low-dimensional non-genetic covariates.

Nous développons une nouvelle méthode d'estimation et de sélection de variables dans les modèles logistiques partiellement linéaires pour des données de grande dimension. La méthode s'appelle estimateur de crête adaptative brisée, et elle estime la régularisation L0 en réalisant itérativement une régularisation L2 quadratique repondérée. Le modèle logistique partiellement linéaire élargit le modèle logistique linéaire en lui ajoutant un composant non paramétrique et le rend assez flexible pour modéliser plusieurs types d'effets de covariables. Nous utilisons les polynômes de Bernstein en guise d'espace de criblage afin d'estimer les fonctions non paramétriques de certaines covariables de faible dimension. Notre méthode s'implante facilement grâce aux bibliothèques R existantes. Des études de simulations approfondies suggèrent que la méthode proposée surpasse les autres méthodes de sélection de variables à base de pénalité couramment employées. Nous appliquons la méthode à une étude en coronaropathie (ce qui est d'ailleurs la raison de notre recherche) et rapportons de nouvelles découvertes concernant les covariables génétiques de grande dimension et celles non génétiques de faible dimension.

[14:30-14:45]

Yan Liu (University of Montreal) **Mireille Schnitzer** (Université de Montréal) **Ronald Herrera** (Université de Montréal; Bayer AG, Berlin, Germany) **Iván Díaz** (NYU Langone Health) **Jennifer O'Loughlin** (Université de Montréal) **Marie-Pierre Sylvestre** (Université de Montréal)

The application of target trials with longitudinal targeted maximum likelihood estimation to assess the effect of alcohol consumption in adolescence on depressive symptoms in adulthood

Application d'essais ciblés avec une estimation par maximum de vraisemblance ciblée longitudinale pour évaluer l'effet de la consommation d'alcool à l'adolescence sur les symptômes de dépression à l'âge adulte

Time-varying confounder adjustment is usually limited by data sparsity, weight instability and computational issues. The Nicotine Dependence in Teens study is a prospective cohort study involving 25 cycles in 1294 high school students. We apply the target trials approach to estimate associations between the alcohol initiation time and the cumulative duration of alcohol use on depression symptoms in adulthood. We define intention-to-treat and as-treated parameters in a marginal structural model with sex as a potential effect modifier and use a double robust estimator, pooled longitudinal target maximum likelihood estimation. We describe strategies for dealing with high-dimensional potential drinking patterns and practical positivity violations due to a long follow-up time, including modifying the effect of interest by removing sparsely observed drinking patterns from the loss function and applying longitudinal

L'ajustement des facteurs de confusion variant dans le temps est généralement limité par les problèmes de rareté des données, d'instabilité du poids et de computation. L'étude Nicotine Dependence in Teens (dépendance à la nicotine chez les adolescents) est une étude de cohorte prospective de 25 cycles d'enquête avec 1 294 étudiants du secondaire. Nous appliquons l'approche par essais ciblés pour estimer les associations entre le moment d'initiation à l'alcool et la durée cumulative de consommation sur les symptômes de dépression à l'âge adulte. Nous définissons les paramètres "intention de traiter" et "tel que traité" dans un modèle structurel marginal avec le sexe comme modificateur potentiel de l'effet et utilisons un estimateur doublement robuste, conjointement à une estimation longitudinale ciblée du maximum de vraisemblance. Nous décrivons des stratégies pour traiter les modèles à haute dimension des potentiels de consommation et les violations en pratique de l'hypothèse de positivité en raison d'un longue période de suivi, y compris en modifiant l'effet d'intérêt par le re-

modified treatment policies to represent the effect of discouraging drinking.

trait des modèles de consommation rarement observés de la fonction de perte et en appliquant des politiques de traitement modifiés longitudinaux pour représenter l'effet de décourager la consommation d'alcool.

Chair/Président: Nicholas Mitsakakis

Room/Salle: RB 1201

Date: Monday May 29 / lundi 29 mai

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-13:45]

Myoungjin Ko (Department of Statistics and Actuarial Science, University of Waterloo)

Adaptive Particle Markov Chain Monte Carlo for Jump Diffusion Models

MCMC à particules adaptatif pour modèles de diffusion à sauts

Jump diffusion models are often used to describe the stochastic and intermittently discontinuous nature of many processes in finance and physics. As these models are solutions to partial differential equations, the likelihood function is rarely available in closed form, thus rendering calibration of model parameters a challenging task. In a Bayesian framework, a popular approach is to use particle filtering within a Markov chain Monte Carlo (MCMC) algorithm to sample from the parameter posteriors. However, such particle MCMC (pMCMC) algorithms are computationally intensive and have many tuning parameters which are difficult to optimize by hand. In this work, we examine various adaptive pMCMC methods, each specifically designed for jump diffusions, and with tuning parameters optimized adaptively in an automated way. The implementation is given in JAX, a high-performance numerical computing package in Python.

Les modèles de diffusion à sauts sont souvent utilisés pour décrire la nature stochastique et discontinue par intermittence de nombreux processus en finance et en physique. Comme ces modèles sont des solutions d'équations différentielles partielles, la fonction de vraisemblance est rarement disponible sous forme fermée, ce qui rend l'étalonnage des paramètres du modèle difficile. Dans un cadre bayésien, une approche populaire consiste à utiliser le filtrage particulaire dans un algorithme de Monte Carlo par chaînes de Markov (MCMC) pour échantillonner les probabilités a posteriori des paramètres. Toutefois, ces algorithmes de MCMC à particules (pMCMC) sont très gourmands en ressources informatiques et comportent de nombreux paramètres de réglage qu'il est difficile d'optimiser à la main. Dans ce travail, nous examinons diverses méthodes pMCMC adaptatives, chacune spécifiquement conçue pour les diffusions à sauts, et dont les paramètres de réglage sont optimisés de manière automatisée et adaptative. L'implémentation est donnée dans JAX, un package de calcul numérique à haute performance sous Python.

[13:45-14:00]

Menelaos Konstantinidis (University of Toronto: Dalla Lana School of Public Health) **Myanca Rodriques** (Department of Health Research Methods, Evidence, and Impact, McMaster University) **Sofia Tsokani** (Department of Primary Education, School of Education, University of Ioannina) **Ian R. White** (Medical Research Council Clinical Trials Unit at UCL, Institute of Clinical Trials and Methodology, University College London) **Julian P. T. Higgins** (Population Health Sciences, NIHR Applied Research Collaboration West (ARC West), University Hospitals Bristol and Weston NHS Foundation Trust) **Guido Schwarzer** (Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center, University of Freidburg) **Dimitris Mavridis** (Department of Primary Education, School of Education, University of Ioannina) **Andrea Tricco** (Division of Epidemiology & Institute of Health Policy, Management and Evaluation, University of Toronto. Li Ka Shing Knowledge Institute, St. Michael's Hospital, Unity Health Toronto) **Areti-Angeliki Veroniki** (Institute of Health Policy, Management and Evaluation, University of Toronto. Li Ka Shing Knowledge Institute, St. Michael's Hospital, Unity Health Toronto)

Assessing network meta-analysis inconsistency in the design-by-treatment interaction model: A simulation study

Évaluation de l'incohérence dans les méta-analyses en réseau dans le modèle d'interaction conçu par traitement (DBT) : une étude par simulation

Network meta-analysis (NMA) is a standard method to simultaneously synthesize evidence from multiple studies comparing multiple interventions. The method allows inferences based on direct and indirect comparisons in a network. However, NMA results are reliable only when assumptions are met. The consistency assumption requires that results from direct and indirect evidence coincide. This work presents a simulation to assess the Power and Type I error of the inconsistency estimator from the design-by-treatment (DBT) interaction model (considered the best approach to date). We will present results over various scenarios (i.e., varying numbers of interventions, studies, and loops in the network). Past simulation studies have focused only on networks of up to 4 interventions, significantly less than what is observed in practice. Thus, by exploring larger networks with varying degrees of connectivity and sparseness, we will evaluate the DBT approach more holistically.

La méta-analyse en réseau (NMA) est une méthode normative pour synthétiser simultanément les résultats de nombreuses études comparant des interventions multiples. Le méthode permet des inférences basées sur des comparaisons directes et indirectes dans un réseau. Cependant les résultats d'une NMA sont fiables seulement si les hypothèses sont respectées. La supposition de cohérence requiert que les résultats des preuves directes et indirectes coïncident. Cette étude présente une simulation pour l'évaluation de la puissance et de l'erreur de type I de l'estimateur d'incohérence du modèle d'interaction conçu par traitement (considérée comme la meilleure approche à ce jour). Nous présentons les résultats selon divers scénarios (c.-à-d. différents nombres d'interventions, d'études et de boucles dans le réseau). Des études de simulation antérieures étaient axées seulement sur des réseaux allant jusqu'à quatre interventions, ce qui est sensiblement inférieur à ce qui est observé en pratique. Par l'exploration de réseaux plus étendus avec divers degrés de connectivité et de parcimonie, nous évaluons l'approche DBT de façon plus holistique.

[14:00-14:15]

Sidi Wu (Simon Fraser University) **Cédric Beaulac** (Université du Québec à Montréal) **Jiguo Cao** (Simon Fraser University)

Autoencoder for Discrete Functional Data Representation Learning and Smoothing

Auto-encodeur pour l'apprentissage et le lissage de représentations de données fonctionnelles discrètes

A common pipeline in functional data analysis is to firstly convert the discretely observed data to smooth functions, and then represent the functions by a finite-dimensional vector of coefficients. Existing methods for smoothing and dimensional reduction mainly focus on learning the linear mappings from the data space to the representation space, which, however, may not be sufficient. In this study, we propose to learn the nonlinear representations of functional data using neural network autoencoders. We design the encoders to employ a feature layer taking the weighted sum of the inner product of the functional data and functional weights over the observed timestamp, and the decoders to apply a coefficient layer that maps the representations extracted back to functional space using basis functions. The developed method compresses the discrete functional data to representations and then outputs smooth functions, and can be applied for both regularly and irregularly spaced data.

Un pipeline courant dans l'analyse de données fonctionnelles est est de convertir les données observées de façon discrète en des fonctions lisses, puis de représenter ces fonctions par un vecteur de coefficients de dimension finie. Les méthodes existantes de lissage et de réduction de dimension sont surtout axées sur l'apprentissage de mappages linéaires à partir de l'espace de données à l'espace de représentation qui peuvent cependant ne pas être suffisants. Nous proposons dans cette étude d'apprendre les représentations non linéaires de données fonctionnelles en utilisant des auto-encodeurs de réseaux neuronaux. Nous concevons les encodeurs pour employer une couche de caractéristiques qui transpose la somme pondérée des données fonctionnelles et poids fonctionnels du produit intérieur à un horodateur observé et les décodeurs pour appliquer une couche de coefficients qui cartographie les représentations extraites de l'espace fonctionnel en utilisant des fonctions de base. La méthode développée comprime les données fonctionnelles discrètes à des représentations et sort ensuite des fonctions lisses ; la méthode peut s'appliquer à des données régulièrement et irrégulièrement observées.

[14:15-14:30]

Luis Ledesma (University of Toronto: Dalla Lana School of Public Health) **Eleanor M. Pullenayegum** (Hospital for Sick Children)

Intercept estimation of semi-parametric joint models in the context of longitudinal data subject to irregular observations

Estimation de l'ordonnée à l'orgine de modèles conjoints semi-paramétriques dans un contexte de données longitudinales

faisant l'objet d'observations irrégulières

Longitudinal data is often subject to irregular visiting times. While there are methods developed to deal with certain irregular visiting distributions, it may be the case that the visiting times are informative, that is, the current value of the outcome may influence the visit intensity at that time. One model that allows for such dependence provided there is conditional independence given a random effect is the Sun multiplicative model, which is one of the few semi-parametric joint models that is suitable for count data using categorical predictors. We extend this model by additionally estimating the intercept term which was formerly treated as a nuisance parameter. This approach enables us to assess prognosis by examining the evolution of the outcome variable over time, rather than just the relative effects of model covariates. We evaluate performance with a simulation study, and an application to a dataset of CD4 counts for a longitudinal cohort of HIV patients will be presented.

Les données longitudinales sont souvent assujetties à des temps de visite irréguliers. Même si des méthodes sont développées pour traiter certaines distributions de temps de visite irréguliers, il peut arriver que les temps de visite soient informatifs, c.-à-d. que la valeur actuelle d'un résultat peut influer sur l'intensité de visites à ce moment. Un modèle qui permet une telle dépendance sous réserve d'une indépendance conditionnelle sachant un effet aléatoire est le modèle multiplicatif de Sun, l'un des rares modèles conjoints semi-paramétriques qui convient aux données de comptage utilisant des prédicteurs catégoriques. Nous étendons ce modèle à l'estimation additionnelle du terme de l'ordonnée à l'orgine qui était auparavant traité comme un paramètre de nuisance. Cette approche permet d'évaluer le pronostic en examinant l'évolution de la variable du résultat avec le temps, plutôt que seulement les effets relatifs des covariables du modèle. Nous présentons une étude par simulation qui permet d'évaluer sa performance et une application à un ensemble de données de comptage CD4 pour une cohorte longitudinale de patients atteints du sida.

[14:30-14:45]

Chi Zhang (University of Waterloo) **Peijun Sang** (University of Waterloo) **Yingli Qin** (University of Waterloo)

Two-sample inferences for sparse functional data

Inférences à deux échantillons pour des données fonctionnelles éparses

We propose a novel procedure to test the difference in mean functions between two groups using penalized smoothing splines. The method addresses sparse and irregularly sampled functional data where observations are made randomly in the domain. Unlike conventional functional principal components analysis approaches, which usually assume the same covariance structures for the two groups, we directly derive a pointwise asymptotic distribution of a mean estimator separately for each group. To do this, we utilize the functional Bahadur representation to obtain a first-order approximation for the mean estimator. The proposed method is conceptually simple and easy to implement. We conduct Monte Carlo simulations to demonstrate finite sample performance under various settings.

Nous proposons une nouvelle procédure pour tester la différence des fonctions de moyennes entre deux groupes à l'aide de splines de lisage pénalisées. La méthode cible les données fonctionnelles éparses et échantillonnées irrégulièrement où les observations sont faites aléatoirement dans le domaine. Contrairement aux approches d'analyse en composantes principales fonctionnelle conventionnelles, qui supposent généralement les mêmes structures de variance pour les deux groupes, nous dérivons directement un estimateur ponctuel de la distribution d'une moyenne séparément pour chaque groupe. Pour ce faire, nous utilisons la représentation fonctionnelle de Bahadur pour obtenir une approximation de premier ordre pour l'estimateur de moyenne. La méthode proposée est conceptuellement simple et facile à implanter. Nous menons des simulations de Monte Carlo afin de démontrer la performance en échantillon fini dans des cadres variés.

[14:45-15:00]

Chi-Kuang Yeh (University of Waterloo) **Gregory Rice** (University of Waterloo) **Joel A. Dubin** (University of Waterloo)

Functional Spherical Autocorrelation: Robust Autocorrelation Estimation of a Functional Time Series

Autocorrélation sphérique fonctionnelle : estimation d'autocorrélation robuste de séries temporelles fonctionnelles

We propose a new autocorrelation measure for functional time series that we term spherical autocorrelation. It is based on measuring the average angle be-

Nous proposons une nouvelle mesure d'autocorrélation pour les séries temporelles fonctionnelles que nous appelons autocorrélation sphérique. Elle est basée sur la mesure de l'angle

tween lagged pairs of series after having been projected onto the unit sphere. This new measure enjoys several complimentary advantages compared to existing autocorrelation measures for functional data, since it both 1) describes a notion of sign or direction of serial dependence in the series, and 2) is more robust to outliers. The asymptotic properties of estimators of the spherical autocorrelation are established, and are used to construct confidence intervals and portmanteau white noise tests. These confidence intervals and tests are shown to be effective in simulation experiments, and demonstrated in applications to model selection for daily electricity price curves, and measuring the volatility in densely observed asset price data.

moyen entre les paires décalées de séries après avoir été projetées sur la sphère unité. Cette nouvelle mesure bénéficie de plusieurs avantages par rapport aux mesures d'autocorrélations actuelles pour les données fonctionnelles : elle décrit une notion de signe ou direction de la dépendance sérielle dans les séries, et elle est plus robuste contre les valeurs aberrantes. Les propriétés asymptotiques des estimateurs de l'autocorrélation sphérique sont établies et exploitées pour construire des intervalles de confiance et des tests portemanteaux de bruit blanc. L'efficacité de ces intervalles de confiance et de ces tests a été démontrée dans des expériences en simulations et dans des applications de sélection de modèles pour des courbes du prix de l'électricité quotidien, et pour la mesure de la volatilité dans les données du prix des actifs densément observé.

Data science methods and applications I
Méthodes et applications de la science des données I

Chair/Président: Jinko Graham

Room/Salle: RB 1200

Date: Monday May 29 / lundi 29 mai

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-13:45]

Rose Garrett (University of Toronto) **Eleanor M. Pullenayegum** (Hospital for Sick Children)

The Role of Sensitivity Analysis in Handling Untestable Assumptions in Analyses of EHR Data

Rôle de l'analyse de sensibilité dans le traitement des hypothèses non vérifiables dans les analyses de données de DSE

EHRs are an important data source for research. However, EHR data are not collected for research purposes, and this creates many analytic challenges. The key issue we address is irregularity of the measurements; the patient's health status influences how often their record is updated, with observations on sicker patients being overrepresented. Failing to properly account for this in the analysis could lead to biased results. Although several analytic methods for irregular observation have been established, Bayesian approaches have been limited. We aim to formulate a novel Bayesian joint model of the disease outcome and irregular visit process that captures complex features of the disease course, incorporates a range of prior clinical beliefs into the modelling framework, and relaxes some of the restrictive and often unrealistic simplifying assumptions about the visit process made by existing methods. Using this model we will develop a novel approach for Bayesian sensitivity analysis.

Les DSE constituent une source de données importante pour la recherche. Cependant, les données des DSE ne sont pas collectées à des fins de recherche, ce qui pose de nombreux défis analytiques. Le principal problème que nous abordons est l'irrégularité des mesures ; l'état de santé du patient influence la fréquence de mise à jour de son dossier, les observations concernant les patients les plus malades étant surreprésentées. Le fait de ne pas tenir compte de ce facteur dans l'analyse pourrait fausser les résultats. Bien que plusieurs méthodes d'analyse des observations irrégulières aient été établies, les approches bayésiennes sont limitées à ce jour. Notre objectif est de formuler un nouveau modèle bayésien conjoint de l'issue de la maladie et du processus de visite irrégulière qui capture les caractéristiques complexes de l'évolution de la maladie, intègre une série de croyances cliniques préalables dans le cadre de modélisation et assouplit certaines des hypothèses simplificatrices restrictives et souvent irréalistes sur le processus de visite formulées par les méthodes existantes. En utilisant ce modèle, nous développerons une nouvelle approche pour l'analyse de sensibilité bayésienne.

[13:45-14:00]

Preethi Ravikumar (McGill University)

Detecting Influential Points in High-Dimensional MRI Data Using Random Projections and Kernel Density Estimation

Détection de points influents dans des données IRM en haute dimension à l'aide de projections aléatoires et de l'estimation de la densité du noyau

Outlier detection in high-dimensional data has gained much attention in recent statistical research. MRI data consists of hundreds of thousands of measurements for every subject at hundreds of time points, and influential observations may be difficult to identify since conventional leverage metrics likely fail for the high dimensional setup. Here, we use a combination of random projection and kernel density estimation to mea-

Récemment, la détection des valeurs aberrantes dans les données en haute dimension a fait l'objet d'une grande attention dans la recherche statistique. Les données IRM se composent de centaines de milliers de mesures pour chaque sujet à des centaines de points dans le temps, et les observations influentes peuvent être difficiles à identifier car il est probable que les mesures conventionnelles de l'effet de levier échouent pour la configuration en haute dimension. Nous utilisons ici une combinaison de projections aléatoires

sure influential points in a high-dimensional MRI data (ABIDE). We take features from every parcellated region of the brain and project it onto a one-dimensional subspace to create a projection matrix. We then use kernel density estimation (KDE) to estimate the underlying distribution of the points in the reduced dataset. We select the top K points with the lowest probability as the outliers. Our approach is computationally efficient and is a global approach since it uses all the data.

et d'estimation de la densité du noyau pour mesurer les points influents dans les données IRM de haute dimension. Nous prenons les caractéristiques de chaque région parcellaire du cerveau et les projetons sur un sous-espace unidimensionnel pour créer une matrice de projection. Nous utilisons ensuite l'estimation de la densité du noyau pour estimer la distribution sous-jacente des points dans l'ensemble de données ainsi réduit. Nous sélectionnons les K premiers points ayant la probabilité la plus faible en tant que valeurs aberrantes. Notre approche est efficace d'un point de vue informatique et est une approche globale puisqu'elle utilise toutes les données.

[14:00-14:15]

Jérémie Boudreault (Institut National de la Recherche Scientifique) **Celine Campagna** (Institut national de santé publique du Québec (INSPQ)) **Fateh Chebana** (Institut national de la recherche scientifique (INRS))

Does machine and deep learning outperform statistical models in predicting the health effects of extreme heat?

L'apprentissage automatique et profond est-il meilleur que les modèles statistiques pour prédire les effets de la chaleur extrême sur la santé ?

Extreme heat events have important health impacts that will be amplified by climate change. To assess these impacts, statistical models have traditionally been used. In this study, we introduced various machine and deep learning approaches to predict the adverse health effects of extreme heat. By allowing complex interactions between environmental predictors, we hypothesized that the predictive power could thus be improved over statistical models. The results showed that, depending on the evaluated performance criteria e.g., with all predictions vs during extreme heat events only, the best models were not always the most complex ones from machine and deep learning. For example, non-linear statistical models seemed to more adequately reproduce the mortality peak observed during heatwaves. This comparison between machine/deep learning and statistical models for heat-health relationships showed that both approaches can be relevant depending on the end goal of the model.

La chaleur extrême a des conséquences importantes sur la santé qui seront amplifiées par les changements climatiques. Pour estimer ces impacts, des modèles statistiques sont traditionnellement utilisés. Dans cette étude, plusieurs approches d'apprentissage automatique et profond sont aussi utilisées, permettant de considérer des interactions complexes entre les variables environnementales. L'hypothèse est que leur performance sera améliorée par rapport aux modèles statistiques. Les résultats montent que, selon la façon d'évaluer la performance, les meilleurs modèles ne sont pas toujours ceux de l'apprentissage automatique et profond. Par exemple, les modèles statistiques non linéaires semblent mieux représenter les pics de mortalité observés lors des vagues de chaleur. Cette comparaison entre l'apprentissage automatique/profond et les modèles statistiques montre que les deux approches peuvent être pertinentes dépendamment de l'objectif visé par le modèle.

[14:15-14:30]

Ziqian Zhuang (University of Toronto: Dalla Lana School of Public Health) **Wei Xu** (Princess Margaret Cancer Centre)

A Supervised Distance-based Feature Selection Algorithm for Multiple Mixed Outcomes

Un algorithme de sélection de caractéristiques basé sur la distance supervisée pour des résultats multiples mixtes

Feature selection is crucial in high-dimensional data analysis for reducing complexity and enhancing performance. However, joint models that simultaneously analyze multiple outcomes pose a challenge for traditional feature selection methods. Traditional feature selection methods relying on specific feature and model metrics may not be suitable in this scenario because of

La sélection de caractéristiques est cruciale dans l'analyse de données de haute dimension pour réduire la complexité et améliorer les performances. Cependant, les modèles conjoints qui analysent simultanément plusieurs résultats posent un défi pour les méthodes traditionnelles de sélection de caractéristiques. Les méthodes traditionnelles de sélection de caractéristiques, qui reposent sur des mesures spécifiques de caractéristiques et de modèles, peuvent

the absence of uniform measurements. A new distance-based supervised learning method is proposed to address these issues. This method performs feature selection by integrating model performance on subsets of features from bootstrap samples in a performance matrix. Gower distance is used to measure predictions of multiple outcomes and the method reduces dimensionality. The method is also compatible with existing model techniques for mixed outcomes. Simulation and real data studies have shown its improvement in feature selection accuracy and prediction performance compared to single-outcome feature selection methods.

ne pas être adaptées dans ce scénario en raison de l'absence de mesures uniformes. Une nouvelle méthode d'apprentissage supervisé basée sur la distance est proposée pour résoudre ces problèmes. Cette méthode effectue la sélection de caractéristiques en intégrant les performances du modèle sur des sous-ensembles de caractéristiques à partir d'échantillons bootstrap dans une matrice de performance. La distance de Gower est utilisée pour mesurer les prédictions de plusieurs résultats et la méthode réduit la dimensionnalité. La méthode est également compatible avec les techniques de modèles existantes pour les résultats mixtes. Des études de simulation et de données réelles ont montré une amélioration de la précision de la sélection de caractéristiques et des performances de prédiction par rapport aux méthodes de sélection de caractéristiques pour un seul résultat.

[14:30-14:45]

Jasper Zhongyuan Zhang (University of Toronto) **Wei Xu** (University of Toronto, Princess Margaret Cancer Centre) **Pingzhao Hu** (Western University)

Integrating Multiomics Data using Deep Tensor Factorization for Survival Outcomes Prediction

Intégration de données multiomiques en utilisant la factorisation de tenseur profond pour les résultats de survie dans le cancer

Motivation: Multiomics cancer profiles provide essential signals for predicting cancer patient survival, but it is challenging to reveal complex patterns from multiple types of data and link them to survival outcomes. We develop a new deep learning-based algorithm to integrate three high-dimensional omics data measured on the same individuals to better understand cancer survival outcomes. Results: Our approach utilizes a three-dimension tensor that integrates multi-omics cancer data and factorizes it into two-dimensional matrices of latent factors, which are fed into neural networks-based survival models. We applied the new algorithm to breast cancer and colon/rectal cancer data from TCGA and compared it to other multi-omics and individual genomic survival analysis methods. Our results demonstrate that the proposed tight integration framework performs better survival prediction than the models using single-type genomic data and other conventional data integration methods.

Motivation : Les profils de cancer multiomiques fournissent des signaux essentiels pour prédire la survie des patients atteints du cancer, mais il est difficile de révéler des motifs complexes intrinsèques aux données et de relier ceux-là aux résultats de survie. À cette fin, nous développons un nouvel algorithme basé sur l'apprentissage profond pour combiner trois types de données omiques à haute dimension mesurées sur les mêmes individus afin de mieux comprendre les résultats de survie. Résultats : Notre approche utilise un tenseur à trois dimensions qui combine les données de cancer multiomiques et les factorise en matrices à deux dimensions de facteurs latents qui sont alimentées dans des modèles de survie basés sur des réseaux de neurones. Nous avons appliqué l'algorithme sur des données de cancer du sein et de cancer colorectal de TCGA et l'avons comparé à d'autres méthodes. Nos résultats montrent que notre méthode possède une meilleure performance de prédiction de survie que les modèles d'intégration de données plus conventionnels.

[14:45-15:00]

Quanhan Xi (The University of British Columbia) **Bloem-Reddy Benjamin** (University of British Columbia)

Indeterminacy in Generative Models: Characterization and Strong Identifiability

Indétermination dans les modèles génératifs : Caractérisation et identifiabilité forte

Generative models posit that latent variables describe the statistical properties of the observed data via a deterministic mapping to the observables, known as the

Les modèles génératifs partent du principe que les variables latentes décrivent les propriétés statistiques des données observées par le biais d'une correspondance déterministe avec les obser-

generator. Modern non-linear generative models are capable of accurately modelling and generating highly complex data, with vast applications in science and industry. Unfortunately, most generative models are non-identifiable, limiting their use for scientific discovery. Our research builds a comprehensive theory of a general class of generative models, both classical and modern, characterizing the precise mathematical form of the indeterminacies that can arise from a given model design. First, identifiable models are proposed by combining classical ideas from factor analysis with modern elements, a connection made possible by the framework developed. Second, examples of downstream tasks that are robust to certain indeterminacies (i.e., weak identifiability) are given, which are safe to use under certain model designs.

vables, appelée générateur. Les modèles génératifs non linéaires modernes sont capables de modéliser et de générer avec précision des données très complexes, avec de vastes applications dans les domaines de la science et de l'industrie. Malheureusement, la plupart des modèles génératifs ne sont pas identifiables, ce qui limite leur utilisation à des fins de découverte scientifique. Notre recherche construit une théorie complète d'une classe générale de modèles génératifs, à la fois classiques et modernes, caractérisant la forme mathématique précise des indéterminations qui peuvent survenir à partir d'une conception de modèle donnée. Tout d'abord, nous proposons des modèles identifiables en combinant des idées classiques de l'analyse factorielle avec des éléments modernes, une connexion rendue possible par le cadre développé. Ensuite, nous donnons des exemples de tâches en aval qui sont robustes à certaines indéterminations (c'est-à-dire une faible identifiabilité), qui peuvent être utilisés en toute sécurité sous certaines conceptions de modèles.

Chair/Président: Kevin McGregor

Organizer/Responsable: Kevin McGregor

Room/Salle: AT 301

Date: Monday May 29 / lundi 29 mai

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-16:00]

Jingyi Cao (York University) **Dongchen Li** (Brock University) **Virginia R. Young** (University of Michigan) **Bin Zou** (University of Connecticut)

Reinsurance Games with Two Reinsurers: Tree Versus Chain

Jeux de réassurance avec deux réassureurs : arborescence versus chaîne

We study reinsurance contracting and competition in a continuous-time model with ambiguity. The market consists of one insurer and two reinsurers, who apply the expected-value premium principle and the variance premium principle to price reinsurance contracts, respectively. The reinsurance contracting problems between the insurer and reinsurers are resolved by Stackelberg differential games, and the reinsurance competition between two reinsurers is settled by a non-cooperative Nash game. We obtain the closed-form equilibrium strategies for all three players under both a tree structure and a chain structure. A detailed comparison study reveals that the tree structure is preferred to the chain structure from a social planner's perspective, and the tree structure is generally preferred from the insurer's perspective.

Nous étudions la passation de contrat et la compétition en réassurance dans un modèle à temps continu avec ambiguïté. Le marché consiste en un assureur et deux réassureurs, qui appliquent respectivement le principe de prime de valeur attendue et de prime de risque de variance aux prix des contrats de réassurance. Les problèmes liés à la passation d'un contrat de réassurance entre l'assureur et les réassureurs sont résolus par les jeux différentiels de Stackelberg et la compétition en matière de réassurance entre les deux réassureurs est réglée par un jeu non coopératif de Nash. Nous obtenons des stratégies d'équilibrage de forme fermée pour les trois acteurs autant sous une structure d'arborescence que de chaîne. Une étude comparative détaillée révèle que la structure arborescente est préférée à la structure de chaîne du point de vue d'un planificateur social, et aussi de façon générale du point de vue de l'assureur.

[16:00-16:30]

Pratheepa Jeganathan (McMaster University) **Shiheng Huang** (McMaster University) **Jamie McNicol** (McMaster University)

Probabilistic Topic Models and Tessellation to Uncover Spatial Compartmentalization in Spatial Omics

Modèles probabilistes de sujets et tessellation pour découvrir la compartimentation spatiale dans les données omiques spatiales

Spatial omics, especially at the single-cell level, uncover substantial heterogeneity, and applying integrative analyses to those data is key to understanding biological complexity. As a result, statistical methods have been developed to address replicated point pattern data, technical variability, spatial dependence, heterogeneity, and sparsity. One of the important goals of spatial omics is

Les données omiques spatiales, en particulier au niveau de la cellule unique, révèlent une hétérogénéité substantielle, et l'application d'analyses intégratives à ces données est essentielle pour comprendre la complexité biologique. En conséquence, des méthodes statistiques ont été développées pour traiter les données de modèles ponctuels répliqués, la variabilité technique, la dépendance spatiale, l'hétérogénéité et l'éparsité. L'un des ob-

to characterize cellular neighborhoods. Recent studies have shown that probabilistic topic models can uncover cellular organization as a mixture of cell types into topics. In this talk, we propose to use tessellation to identify spatial compartmentalization based on the distribution of topics. Finally, we will demonstrate the proposed method to learn the tumor micro-environment in triple-negative breast cancer tumors.

jectifs importants de l'omique spatiale est de caractériser les voisinages cellulaires. Des études récentes ont montré que les modèles probabilistes de sujets peuvent révéler l'organisation cellulaire comme un mélange de types de cellules en sujets. Dans cet exposé, nous proposons d'utiliser la tessellation pour identifier la compartimentation spatiale sur la base de la distribution des sujets. Enfin, nous démontrerons la méthode proposée pour apprendre le micro-environnement tumoral dans les tumeurs du cancer du sein triple négatif.

[16:30-17:00]

Mohamed Belalia (University of Windsor) **Guanjie Lyu** (University of Windsor)

Nonparametric hypotheses testing in copula models: Some recent avenues.

Test d'hypothèses non paramétrique dans des modèles de copule : quelques avenues récentes

Testing hypotheses in copula models is central to modern statistical analysis. This talk presents some recent advances in nonparametric hypothesis testing, which include testing independence, the equality of the dependence structure in two samples and testing the symmetry of copula models. The asymptotic properties of the proposed tests statistics will be established under certain regularity conditions. The performance of the presented approaches will be discussed through simulation studies. Furthermore, some real data will be analyzed using the aforementioned methodologies.

Tester des hypothèses dans des modèles de copule est au cœur de l'analyse statistique moderne. Nous abordons certaines avancées récentes en matière de tests d'hypothèses non paramétriques, qui comprennent les tests d'indépendance, d'égalité de la structure de dépendance dans deux échantillons et de symétrie des modèles de copule. Les propriétés asymptotiques des statistiques de test proposées sont établies sous certaines conditions de régularité. À l'aide d'études de simulation, nous discuterons de la performance des approches présentées. De plus, les méthodologies susmentionnées serviront à l'analyse de certaines données réelles.

Nonlinear and Nonparametric Regression Models in Environmental and Social Health
Modèles de régression non linéaires et non paramétriques en santé environnementale et sociale

Chair/Président: Alex Stringer

Organizer/Responsable: Alex Stringer

Room/Salle: TB 210

Date: Monday May 29 / lundi 29 mai

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-16:00]

Glen McGee (University of Waterloo)

Testing for Non-Additive Interaction in Flexible Bayesian Models for Multi-Pollutant Mixtures

Test pour l'interaction non additive dans des modèles bayésiens flexibles pour des mélanges de multi-polluants

An important goal of environmental health research is to assess the risk posed by mixtures of multiple environmental exposures. In these mixtures analyses, flexible models like Bayesian kernel machine regression are appealing because they allow for non-linear associations and non-additive interactions among mixture components. However, this flexibility comes at the cost of low power and difficult interpretation. We propose a flexible framework that allows for separate selection of additive and non-additive effects within the kernel machine framework, and it extends easily to multiple index models and distributed lag interaction models. The proposed approach yields increased power and simpler interpretation when there is little evidence of interaction. Further, it allows users to specify separate priors for additive and non-additive effects, and allows for tests of non-additive interaction.

Un objectif important de la recherche en santé environnementale est l'évaluation du risque posé par les mélanges d'expositions environnementales multiples. Dans les analyses de ces mélanges, des modèles flexibles comme la régression bayésienne à machines de noyau sont attrayants parce qu'ils permettent des associations non linéaires et des interactions non additives parmi les composantes du mélange. Cette flexibilité vient cependant au prix d'une faible puissance et d'une interprétation difficile. Nous proposons un cadre flexible qui permet une sélection séparée d'effets additifs et non additifs dans un cadre de machines de noyau et qui s'étend facilement aux modèles multi-indices et aux modèles d'interaction à retard distribué. L'approche proposée accroît la puissance et simplifie l'interprétation lorsqu'il existe peu de preuve d'interaction. De plus, elle permet aux utilisateurs de spécifier des a priori séparés pour les effets additifs et non additifs, tout en permettant les tests d'interaction non additive.

[16:00-16:30]

Michael Y.C. Chong (University of Toronto) **Monica Alexander** (University of Toronto)

Estimating the Timing of Stillbirths in Countries Worldwide Using a Bayesian Hierarchical Penalized Splines Regression Model

Estimation des périodes de mortinatalité dans les pays du monde entier à l'aide d'un modèle de régression hiérarchique bayésien à splines pénalisées

Reducing the global burden of stillbirths is important to improving child and maternal health. Of interest are patterns in the timing of stillbirths – that is, whether they occur in the intra- or antepartum period – because stillbirths that occur intrapartum are largely preventable. However, data availability on the timing of stillbirths is highly variable, with low- and middle-income countries

Il est important de réduire le fardeau mondial de la mortinatalité pour améliorer la santé maternelle et infantile. Il est intéressant de noter les tendances concernant le moment où survient la mortinaissance, c'est-à-dire s'il s'agit d'une mortinaissance intranatale ou anténatale, car les mortinaissances intrapartum peuvent être en grande partie évitées. Cependant, la disponibilité des données sur le moment de la naissance d'un enfant mort-né est très variable,

generally having few reliable observations. In this paper, we develop a Bayesian penalized splines regression framework to estimate the proportion of stillbirths that are intrapartum for all countries worldwide. A weighting procedure is proposed to account for unrepresentative subnational data.

les pays à revenu faible ou intermédiaire disposant généralement de peu d'observations fiables. Dans cet article, nous élaborons un cadre de régression bayésienne par splines pénalisées pour estimer la proportion de mortinaissances intrapartum dans le monde entier. Nous proposons une procédure de pondération pour tenir compte des données infranationales non représentatives.

[16:30-17:00]

Liqun Diao (University of Waterloo) **Grace Y. Yi** (University of Western Ontario) **Haoxin Zhuang** (BeiGene Inc.)

Polya Tree Based Nearest Neighborhood Regression

Régression par méthode des plus proches voisins, basée sur l'arborescence de Pólya

Parametric regression, such as linear regression, plays an important role in statistics. The use of parametric regression models typically involves the specification of a regression function of the covariates, the distribution of response and the link between the response and covariates, which are commonly at the risk of misspecification. In this talk, we introduce a fully non-parametric regression model, a Polya tree (PT) based nearest neighborhood regression. To approximate the true conditional probability measure of the response given the covariate value, we construct a PT-distributed probability measure of the response in the nearest neighborhood of the covariate value of interest. Our proposed method gives consistent and robust estimators, and has a faster convergence rate than the kernel density estimation. We conduct extensive simulation studies and analyze a Combined Cycle Power Plant dataset to compare the performance of our method relative to kernel density estimation, PT density estimation, and linear dependent tail-free process (LDTFP). The studies suggest that the proposed method exhibits the superiority to the kernel and PT density estimation methods in terms of the estimation accuracy and convergence rate and to LDTFP in terms of robustness.

La régression paramétrique, comme la régression linéaire, joue un rôle important en statistique. L'utilisation de modèles de régression paramétrique implique généralement la spécification d'une fonction de régression des covariables, la distribution de réponse ainsi que le lien entre la réponse et les covariables, qui risquent couramment l'erreur de spécification. Notre exposé présente un modèle de régression entièrement non paramétrique, soit une régression par méthode des plus proches voisins, basée sur l'arborescence de Pólya (PT). À des fins d'approximation de la mesure de probabilité conditionnelle réelle de la réponse, et compte tenu de la valeur des covariables, nous construisons une mesure de probabilité par distribution PT de la réponse par plus proches voisins de la valeur d'intérêt des covariables. La méthode que nous proposons donnent des estimateurs convergents et robustes et son taux de convergence est plus rapide que l'estimation par noyau de la densité. Nous menons des études de simulation exhaustives et analysons un ensemble de données d'une centrale électrique à cycle combiné pour comparer la performance de notre méthode à l'estimation par noyau de la densité, à l'estimation par PT de la densité et au processus sans queue dépendant linéaire (LDTFP). Les études suggèrent que la méthode proposée se montre supérieure aux méthodes d'estimation de la densité par noyau et par PT sur le plan de l'exactitude de l'estimation et du taux de convergence et au LDTFP sur le plan de la robustesse.

Modern Statistical Learning with Complex Large-scale Data
Apprentissage statistique moderne avec données complexes à grande échelle

Chair/Président: Chen Xu

Organizer/Responsable: Chen Xu

Room/Salle: AT 102

Date: Monday May 29 / lundi 29 mai

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-16:00]

Shili Lin (The Ohio State University)

Sparse Estimation in Finite Mixture of Varying Coefficient Regression Models

Estimation éparse dans des mélanges finis de modèles de régression à coefficients variables

A generalization of the commonly used finite mixture of regressions models, in which the effects of some covariates may vary with respect to an index variable such as time, has been introduced to better accommodate complex features in real data applications, including genetic data. Such a model could be used for investigating heterogeneous effects in unknown subpopulations, varying effects over time, and many covariates. However, existing methods are not amenable to accommodate all these features simultaneously. In this talk, we will describe a new methodology based on a regularized likelihood approach for simultaneous parameter estimation and variable selection. We will present large-sample properties of the proposed method, as well as finite sample performance through simulation. We will apply the proposed methodology to a genetic dataset to demonstrate its utility.

Une généralisation des actuels mélanges finis des modèles de régressions, dont les effets de certaines covariables peuvent varier selon une variable d'indice comme le temps, a été présentée pour davantage accommoder les caractéristiques complexes des applications avec des données réelles, y compris les données génétiques. Un tel modèle pourrait servir à étudier les effets hétérogènes dans des sous-populations inconnues, des effets variables dans le temps et plusieurs covariables. Par contre, les méthodes actuelles ne sont pas suffisamment souples pour tenir compte de tous ces aspects simultanément. Lors de cet exposé, nous décrirons donc une nouvelle méthodologie basée sur une approche de vraisemblance régularisée pour l'estimation de paramètres simultanée et la sélection de variables. Nous présenterons des propriétés de grands échantillons de la méthode proposée, ainsi que la performance de l'échantillon fini à l'aide de simulations. Nous appliquerons la méthodologie proposée à un ensemble de données génétique pour illustrer son utilité.

[16:00-16:30]

Shaista Jaffer (University of Ottawa)

Event-driven Trading Strategies with Explainable Machine Learning

Stratégies de négociation événementielle à l'aide de l'apprentissage automatique explicable

Using machine learning (ML) techniques for stock price forecasting solely based on historical data faces several challenges. Stock prices are influenced by multiple factors that lead to inconsistent and non-stationary patterns, making these algorithms less effective. Additionally, they can only discover correlations and not causal relationships, resulting in a loss of effectiveness during market shifts. Furthermore, finance datasets may be too small for ML to be effective. Our project aims

La prévision des cours boursiers au moyen de techniques d'apprentissage automatique uniquement basées sur des données historiques présente plusieurs difficultés. Les cours des actions sont influencés par de multiples facteurs qui produisent des modèles incohérents et non stationnaires, ce qui rend ces algorithmes moins efficaces. En outre, ils ne peuvent découvrir que des corrélations et non des relations de cause à effet, ce qui entraîne une perte d'efficacité lors des fluctuations du marché. De plus, les ensembles de données financières sont parfois trop restreints pour que l'ap-

to address these issues by focusing on predicting well-researched corporate events such as mergers and acquisitions, bankruptcy, IPOs, and integrating economic modeling into ML. We experimented with various approaches, such as dynamic investment model with reinforcement learning, and deep learning with knowledge graphs, and compared their results. All these approaches have an average prediction accuracy of 80%, significantly higher than the 30% accuracy achieved through stock price forecasting.

prentissage automatique soit efficace. Dans le cadre de notre projet, nous cherchons à résoudre ces problèmes en nous concentrant sur la prévision d'événements d'entreprise bien documentés (fusions et acquisitions, faillites et introductions en bourse) et en intégrant la modélisation économique dans l'apprentissage machine. Nous avons expérimenté différentes approches, telles que le modèle d'investissement dynamique avec apprentissage par renforcement et l'apprentissage profond avec les graphes de connaissances, et nous avons comparé les résultats. Toutes ces approches ont une précision de prévision moyenne de 80 %, ce qui est nettement supérieur à la précision de 30 % obtenue par la prévision du cours de l'action.

[16:30-17:00]

Kaili Jing (Xi'an JiaoTong University)

Low Gradient-based Subsampling for Corrupted Massive Data

Sous-échantillonnage basé sur un faible gradient pour mégadonnées corrompues

Subsampling is one of basic methodologies used for coupling with massive data. In the past studies, the large gradient subsampling is normally suggested, which assigns the subsampling probability directly proportional to the gradient of loss. Such a strategy is effective for uncontaminated massive data but not for corrupted massive data. Based on an observation that the use of large gradient brings the danger of assigning higher sampling probabilities to the heavily corrupted observations, we propose a low gradient subsampling procedure for corrupted massive data. We show that the low gradient of loss corresponds to the uncontaminated data in high probability and therefore may express the impact of the use of heavily corrupted observations. We also show that the procedure is statistically consistent with achieving a nearly optimal estimation precision. The promising performance of the procedure is supported on a variety of numerical simulations and real data applications.

Le sous-échantillonnage est l'une des méthodologies de base employée pour le couplage de mégadonnées. Dans des études antérieures, on suggère généralement le sous-échantillonnage de fort gradient, qui donne la probabilité de sous-échantillonnage directement proportionnelle au gradient de perte. Une telle stratégie est efficace pour les mégadonnées non contaminées, mais pas pour celles qui sont corrompues. En nous fondant sur l'observation voulant que l'utilisation d'un fort gradient comporte le danger d'attribuer de forte probabilités d'échantillonnage à des observations lourdement corrompues, nous proposons une procédure de sous-échantillonnage à faible gradient pour les mégadonnées corrompues. Nous démontrons que le faible gradient de perte correspond en forte probabilité aux données non contaminées, et donc peut exprimer l'impact de l'utilisation d'observations grandement corrompues. De plus, la procédure est statistiquement cohérente avec la réalisation d'une précision d'estimation presque optimale. La performance prometteuse de cette procédure est soutenue par de nombreuses simulations numériques et d'applications à des données réelles.

In celebration of 50 years of research in Stochastics at Carleton University: Session in Honour of Donald A. Dawson
Célébrations de 50 ans de recherche en stochastique à l'Université Carleton : Session en l'honneur de Donald A. Dawson

Chair/Président: Bouchra Nasri

Organizer/Responsable: Bouchra Nasri

Room/Salle: AT 101

Date: Monday May 29 / lundi 29 mai

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-16:00]

Bruno N. Rémillard (HEC Montréal) **Jean Vaillancourt** (HEC Montreal)

Central limit theorem for martingales

Un théorème central limite pour martingales

When the limiting compensator of a sequence of martingales is continuous, we obtain a weak convergence theorem for the martingales; the limiting process can be written as a Brownian motion evaluated at the compensator and we find sufficient conditions for both processes to be independent. We will give examples of applications to occupation times processes and for financial volatility measures. This is joint work with Jean Vaillancourt.

Lorsque la limite d'une suite de compensateurs de martingales est continue, on obtient un théorème de convergence faible pour les martingales ; le processus limite peut s'écrire comme un mouvement brownien évalué au compensateur et nous trouvons des conditions suffisantes pour que les deux processus soient indépendants. Nous donnerons des exemples d'applications pour des processus de temps d'occupation, ainsi que pour des mesures de volatilité. Il s'agit de travaux conjoints avec Jean Vaillancourt.

[16:00-16:30]

Shui Feng (McMaster University)

Relative Entropy and Hierarchical Dirichlet Process

Entropie relative et processus de Dirichlet hiérarchique

Hierarchical Dirichlet process is a discrete random measure serving as an important prior in the Bayesian nonparametrics. It is motivated with the study of groups of clustered data. This talk focuses on its connection to relative entropy through large deviations. The explicit form reveals the impact of different hierarchies. Some related results will be discussed as time permit including annealed large deviations for Bayesian bootstrap and the Dirichlet posteriors.

Le processus de Dirichlet hiérarchique est une mesure aléatoire discrète servant de distribution a priori importante dans les approches bayésiennes non paramétriques. Il est motivé par l'étude de groupes de données groupées. Cette présentation se concentre sur son lien avec l'entropie relative à travers de grandes déviations. La forme explicite révèle l'impact des différentes hiérarchies. Certains résultats connexes seront discutés si le temps le permet, y compris les grands écarts recuits pour le bootstrap bayésien et les distributions a posteriori de Dirichlet.

[16:30-17:00]

Jean Vaillancourt (HEC Montréal) **Bruno N. Rémillard** (HEC Montréal)

Another central limit theorem plus a few applications

Un autre théorème limite central et quelques applications

A new convergence theorem is presented, for sequences of martingales with càdlàg trajectories (right continuous with left limits everywhere) in the sense of the weak

Un nouveau théorème de convergence est présenté pour les suites de martingales à trajectoires càdlàg (continues par la droite et ayant des limites par la gauche) en terme de la topologie faible

dual topology on Hilbert space, under very weak conditions. Examples are provided to show that these conditions are easy to check and yield useful asymptotic results, especially when the limit is a mixture of stochastic processes with discontinuities, notably for: stock price approximation, occupation time for planar Brownian motion and functional principal component analysis. This is joint work with Bruno Rémillard.

duale sur un espace d'Hilbert, sous des conditions très faibles. On fournit des exemples illustrant que ces conditions sont faciles à vérifier tout en livrant des résultats asymptotiques utiles, surtout lorsque la limite de la suite est un mélange de processus aléatoires ayant des discontinuités, notamment pour : l'approximation des prix d'actions, le temps d'occupation du movement brownien planaire et l'analyse en composantes principales fonctionnelles. Il s'agit de travaux conjoints avec Bruno Rémillard.

New Tools in Teaching Statistics
Nouveaux outils pour l'enseignement de la statistique

Chair/Président: Wanhua Su

Organizer/Responsable: Wanhua Su

Room/Salle: TB 208

Date: Monday May 29 / lundi 29 mai

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-16:00]

Becky Wei Lin (Simon Fraser University)

The Design and Use of Shiny Apps for Teaching Introductory Statistics Courses

Conception et utilisation des applications Shiny dans l'enseignement de cours d'introduction aux statistiques

In most introductory statistics courses, a wide spectrum of topics to deliver leaves the instructor limited room for introducing statistical language like R at the same time. In addition, most students in these courses have no prior computer coding experience. Consequently, learning both statistics and mastering R coding skills can be overwhelming for most students, which can also be a big challenge for instructors as well. Furthermore, most students in these courses either have little math background or are not very mathematically inclined. In many social science and life science majors, students report a high level of statistics anxiety. To address all the challenges, we design and create a collection of Fun Apps of Statistical Tools (FAST) with R Shiny Apps. A quick review of these fun apps will be introduced in this talk, and examples of how to use them to create online quizzes and homework assignments will be also illustrated.

Dans la plupart des cours d'introduction aux statistiques, la panoplie de sujets à aborder laisse peu de temps aux enseignants pour présenter le langage statistique comme R. De plus, la plupart des étudiants dans ces cours n'ont aucune expérience antérieure en programmation informatique. Conséquemment, il peut être trop exigeant d'apprendre à la fois les statistiques et acquérir des aptitudes de programmation en R, sans compter que cela peut représenter un grand défi pour les enseignants aussi. D'ailleurs, la plupart des étudiants inscrits à ces cours ont très peu d'antécédents en mathématiques ou y sont peu doués. En sciences humaines et en science de la nature, de nombreux étudiants disent se sentir très anxieux à l'idée d'aborder les statistiques. Afin de résoudre ces défis, nous avons conçu et créé une collection d'outils statistiques amusants en application (FAST) avec les applications Shiny R. Lors de cet exposé, nous survolerons certaines de ces applications amusantes et nous illustrerons à l'aide d'exemple comment créer des examens pratiques et des devoirs en ligne.

[16:00-16:30]

Jonathan Herman (University of Toronto at Mississauga)

MathMatize: A New Tool for Mathematical and Statistical Education

MathMatize : un nouvel outil pour l'enseignement des mathématiques et des statistiques

We will give an overview of MathMatize, a math education platform for secondary and post-secondary math. MathMatize offers assessment and polling tools for instructors, with a content editor built for math instruction. There are several different question types available, and a growing community bank of exercises available. The system has a computational engine, tolerance settings, and randomization capabilities to address a variety of use cases. Examples on how to use MathMatize to create online quizzes and live polls will be illustrated.

Nous présenterons MathMatize, une plateforme d'enseignement des mathématiques pour le secondaire et le post-secondaire. MathMatize propose des outils d'évaluation et de sondage pour les enseignants, ainsi qu'un éditeur de contenu conçu pour l'enseignement des mathématiques. Cette plateforme contient plusieurs types de questions et une banque d'exercices en pleine expansion. Elle dispose d'un moteur de calcul, de paramètres de tolérance et de capacités de randomisation pour répondre à diverses utilisations. Nous illustrerons des exemples d'utilisation de MathMatize pour créer des questionnaires en ligne et des sondages en direct.

New Developments at the Interface of Finance and Pension
Nouveaux développements à l'interface des finances et des pensions

Chair/Président: Jean-François Bégin

Organizer/Responsable: Jean-François Bégin

Room/Salle: RB 1200

Date: Monday May 29 / lundi 29 mai

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-16:00]

Thomas Salisbury (York University)

Regulatory Constraints and the Riccati Tontine

Les contraintes réglementaire et la tontine de Riccati

Canada now hosts two retail tontine products. The one offered by Guardian financial (GuardPath) has an interesting design, prompted by regulatory constraints. Survivors split the proceeds at the end of 20 years, but those who die or leave early (say at time t) receive a fraction k(t) of their account value Z(t) back. The regulatory constraint is that k(t)E[Z(t)] must be at least their initial deposit. This type of regulatory constraint has not received much attention in the literature. We will look at the choice of k(t) that optimizes the expected return to survivors, in a representative agent model, and will show that it satisfies a Riccati equation independently of the size of the tontine pool. We will also examine the effect of a stochastic hazard rate on the tontine, particularly if it invests in assets correlated with mortality. This is joint work with Moshe Milevsky.

Le Canada a maintenant deux tontines pour les consommateurs. Celle de Guardian Financial (GuardPath) présente un plan intéressant, à cause de la réglementation financière. Les survivants se partagent le produit après 20 ans, mais ceux qui meurent ou qui quittent tôt (par exemple, au moment t) reçoivent une fraction k(t) de la valeur de leur compte Z(t). La contrainte réglementaire est que k(t)E[Z(t)] ne doit pas être inférieure au dépôt initial. Ce type de contrainte réglementaire n'a pas reçu beaucoup d'attention dans la littérature. Nous nous intéresserons au choix de k(t) qui optimise le retour attendu aux survivants, dans un modèle d'agent représentatif, et montrerons qu'il satisfait une équation de Riccati indépendent du nombre de participants dans la tontine. Nous examinerons également l'effet d'un taux de risque aléatoire, par exemple si la tontine s'investit dans des actions corrélés à la mortalité. Ce travail a été réalisé conjointement avec Moshe Milevsky.

[16:00-16:30]

Barbara Sanders (Simon Fraser University) **Jean-François Bégin** (Simon Fraser University) **Wenyuan Zhou** (Simon Fraser University)

To Merge or not to Merge? On the Impacts of Pension Plan Consolidation

Fusionner ou ne pas fusionner ? Des impacts de la consolidation des régimes de retraite

We investigate the benefits and drawbacks of pension plan consolidation by quantifying the impact of mergers. To do so, we propose an exhaustive framework that combines all the relevant components of pension plan operation. Specifically, we include a realistic economic scenario generator, a stochastic mortality model that captures differences among sub-populations, a cost model with economies of scale, and an asset allocation methodology based on utility maximization. Three

Nous étudions les avantages et les inconvénients de la consolidation des régimes de retraite en quantifiant l'impact des fusions. Pour ce faire, nous proposons un cadre exhaustif qui combine toutes les composantes pertinentes du fonctionnement des régimes de retraite. Plus précisément, nous incluons un générateur de scénarios économiques réalistes, un modèle de mortalité stochastique qui tient compte des différences entre sous-populations, un modèle de coûts avec des économies d'échelle et une méthodologie d'allocation d'actifs basée sur la maximisation

groups of measures are constructed from this framework to quantify mergers' solvency and welfare impacts: plan-related risk measures assessing profits from an economic capital perspective, consumption-based metrics to understand members' benefits, and contribution risk measures capturing the risk from the sponsor's viewpoint. We apply the framework to two hypothetical mergers and find that the merger is favourable for small plans in both cases.

de l'utilité. Nous construisons trois groupes de mesures à partir de ce cadre pour quantifier les impacts des fusions sur la solvabilité et le bien-être : des mesures de risque liées au régime évaluant les bénéfices du point de vue du capital économique, des mesures basées sur la consommation pour comprendre les avantages pour les membres, et des mesures de risque de contribution capturant le risque du point de vue du promoteur. Nous appliquons ce cadre à deux fusions hypothétiques et constatons que la fusion est favorable aux petits régimes dans les deux cas.

[16:30-17:00]

David Saunders (University of Waterloo)

Risk Sharing in Pension Plans: Optimality and Fairness

Partage du risque dans les régimes de pension : optimalité et équité

We will discuss mathematical problems that arise in the analysis of group pension plans, related to the fairness and optimality of the plan design. In particular, we focus on designs aiming to provide plan members with safe and stable lifetime consumption, while maintaining fairness across generations.

Nous aborderons les problèmes mathématiques qui surviennent dans l'analyse d'un régime de pension collectif relié à l'équité et l'optimalité de la conception du régime. Tout particulièrement, nous nous concentrons sur les conceptions cherchant à fournir aux participants au régime une consommation de durée de vie stable et sécuritaire, tout en maintenant une équité à travers les générations.

Analytic Approaches for Novel Data Sources
Approches analytiques pour nouvelles sources de données

Chair/Président: Hugh Chipman

Room/Salle: TB 236

Date: Monday May 29 / lundi 29 mai

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-15:45]

Cédric Beaulac (Université du Québec à Montréal)

Developing neural network architectures for functional data analysis

Développer de nouvelles architectures de réseaux de neurones pour l'analyse de données fonctionnelles

I will introduce the recent findings of my research group about the integration of neural networks in the analysis of functional data. First, we propose a new layer architecture, a functional output layer, allowing us to use neural networks to output functional responses. This can be used to solve the function-on-scalar regression problem in a non-linear way. Second, we propose a concept for functional weights which can project functional data to a scalar representation. By combining these weights with the functional output layer previously established, we create a functional autoencoder. This model is both able to return a finite representation of functional data and a smooth representation of that data. Finally, inspired by the successes of convolutional neural networks, we propose to address the connections in the functional output layer in a meaningful way to both reduce the number of parameters drastically and to capture the temporal correlation of functional data.

Dans cette présentation, j'introduirai les grandes lignes du programme de recherche que je développe avec mes étudiants qui porte sur l'intégration de méthodes d'apprentissage dans l'analyse de données fonctionnelles. Tout d'abord, nous proposons une modification des réseaux de neurones afin de retourner une courbe. Cela nous permet donc de proposer une solution non linéaire pour la régression de données fonctionnelles. Ensuite, nous proposons un paramètre fonctionnel nous permettant de projeter nos données sur un ensemble de scalaires. En combinant ces résultats, nous établissons un autoencodeur fonctionnel. Ce dernier peut à la fois nous retourner une représentation finie des données fonctionnelles en plus de lisser les données entrées. Finalement, inspiré des réseaux de neurones convolutif, nous proposons de réduire le nombre de connexions dans les couches fonctionnelles afin de réduire drastiquement le nombre de paramètres tout capturant la corrélation temporelle des données.

[15:45-16:00]

Mahsa Panahi (Department of Statistics and Actuarial Science, University of Waterloo) **Stefan Steiner** (University of Waterloo) **Jeroen de Mast** (University of Amsterdam)

Identifying Dominant Causes using Leveraged Study Designs

Identification des causes dominantes à l'aide de plans d'étude à effet de levier

Excessive variation in critical-to-quality process outputs is a common challenge in manufacturing industries. To mitigate it, determining the dominant cause(s) of variation is an effective and recommended first step. Shainin's group comparison procedure is a way to identify these dominant cause(s) through a structured investigation plan that exploits the idea of leveraging. Our study evaluates the soundness of this procedure, and we find Shainin's analysis procedure inefficient and,

La variation excessive des résultats des processus essentiels à la qualité est un problème courant dans les industries manufacturières. Pour l'atténuer, il est utile et recommandé dans un premier temps de déterminer la ou les causes dominantes de la variation. La procédure de comparaison de groupes de Shainin est un moyen d'identifier ces causes dominantes grâce à un plan d'investigation structuré qui exploite l'idée de l'effet de levier. Notre étude évalue le bien-fondé de cette procédure, et nous constatons que la procédure d'analyse de Shainin est inefficace

more importantly, unreliable in identifying the dominant cause. As an alternative, we propose a new, efficient, and reliable analysis procedure based on maximum likelihood. A critical evaluation demonstrates the superiority of our proposed analysis procedure. We also provide a tangible example and compare the outcomes of both methods.

et, plus important encore, qu'elle n'est pas fiable pour identifier la cause dominante. Comme alternative, nous proposons une nouvelle procédure d'analyse efficace et fiable basée sur le maximum de vraisemblance. Une évaluation critique démontre la supériorité de la procédure d'analyse proposée. Nous fournissons également un exemple concret et comparons les résultats des deux méthodes.

[16:00-16:15]

Matthew Baxter (University of Guelph)

Comparing Ensemble Methods for a Two-Level Individual Level Model Parameter Estimation

Comparaison de méthodes d'ensemble pour l'estimation des paramètres d'un modèle à deux niveaux individuels

Disease transmission models play a large role in determining appropriate containment strategies for many infectious disease outbreaks, including the recent COVID-19 epidemic. However, virus transmission often differs between cities. To account for this, a two-level spatial individual-level model (ILM) is used to capture the disease spread between and within cities. We then investigate the use of classification and regression ensemble methods using a random forest algorithm to estimate the two-level ILM parameters, rather than traditional yet computationally expensive procedures such as Bayesian MCMC. We compare the ability of these methods to estimate the parameters for the two-level ILM, and whether the resulting parameter estimates can successfully model the disease transmission.

Les modèles de transmission des maladies jouent un rôle important dans la détermination des stratégies d'endiguement appropriées pour de nombreuses épidémies, y compris la récente épidémie de COVID-19. Cependant, la transmission du virus diffère souvent d'une ville à l'autre. Pour en tenir compte, nous utilisons un modèle spatial à deux niveaux individuels (ILM) pour capturer la propagation de la maladie entre les villes et à l'intérieur de celles-ci. Nous étudions ensuite l'utilisation de méthodes d'ensemble de classification et de régression utilisant un algorithme de forêt aléatoire pour estimer les paramètres du modèle, plutôt que des procédures traditionnelles mais coûteuses en termes de calcul, telles que le MCMC bayésien. Nous comparons la capacité de ces méthodes à estimer les paramètres de l'ILM à deux niveaux, et nous vérifions si les estimations des paramètres qui en résultent peuvent modéliser avec succès la transmission de la maladie.

[16:15-16:30]

Jinxin Yuan (University of Manitoba)

Bayesian Analysis of Dolly Varden Mark-recapture Data in Canadian Arctic

Analyse bayésienne de données de marquage-recapture du Dolly Varden dans l'archipel arctique canadien

In mark-recapture studies, survival parameters are often subject to individual heterogeneity and affected by environmental effects and observation errors. Besides that, recent interest has included additional complexities such as individual covariance and random effects within the statistical framework. To meet the requirements of analysing complex mark-recapture or mark-recapture-recovery data, novel Bayesian fitting state-space models provide a practical tool by coupling a model of mechanistic movement properties (known as process model) with a model of the observation methods (known as observation model). In this study, the Bayesian analysis was conducted on Dolly Varden Mark-recapture Data in Canadian Arctic and unbiased biological information about Dolly Varden can be summarized from mark-recapture data which can be helpful to determine if the

Dans les études de marquage-recapture, les paramètres de survie sont souvent soumis à l'hétérogénéité individuelle et affectés par le effets environnementaux et les erreurs d'observation. Par ailleurs, l'intérêt s'est récemment porté, entre autres choses, sur des complexités additionnelles comme la covariance individuelle et les effets aléatoires à l'intérieur du cadre statistique. Pour répondre aux critères d'analyse des données complexes de marquage-recapture et marquage-recapture-récupération, de nouveaux modèles d'ajustement bayésiens espace-état fournissent un outil pratique en appariant les propriétés d'un modèle mécaniste (appelé modèle de processus) à un modèle de méthodes d'observation (appelé modèle d'observation). Pour cette étude, on a procédé à l'analyse bayésienne de données de marquage-recapture du Dolly Varden dans l'archipel arctique canadien. De l'information d'ordre biologique sans biais sur le Dolly Varden peut être résumée à partir des données de marquage-recapture, ce qui peut

population was sustainably harvested and build an environment friendly and sustainably fishing community.

être utile pour déterminer si la population de poissons a été exploitée de façon durable et pour édifier une communauté de pêche écologique et durable.

[16:30-16:45]

Trang Bui (University of Waterloo) **Stefan Steiner** (University of Waterloo) **Nathaniel T. Stevens** (University of Waterloo)

General Additive Network Effect Models

Modèles généraux d'effet additif de réseau

Social network companies often carry out experiments to test product changes and new ideas. In such experiments, the outcome of one user may be influenced by the treatment assignment of other users, which necessitates specialized design and analysis methods. We introduce the general additive network effect (GANE) model, which unifies many existing outcome models under an encompassing model-based framework. The model is interpretable and flexible in modeling both the treatment effect and the network influence. We show that maximum likelihood estimators are consistent and asymptotically normal for a family of model specifications. Causal quantities of interest such as the global treatment effect are defined and expressed as functions of the GANE model parameters, and hence inference can be carried out accordingly. We further propose the "power-degree" (POW-DEG) specification of the GANE model, whose performance, together with that of other specifications, is investigated via simulations.

Les entreprises de réseaux sociaux mènent souvent des expériences pour tester des changements de produit et de nouvelles idées. Dans de telles expériences, le résultat d'un utilisateur peut être influencé par l'attribution de traitement d'autres utilisateurs, ce qui demande un plan et des méthodes d'analyses adaptées. Nous présentons le modèle général d'effet additif de réseau (GANE), qui unifie plusieurs modèles de résultats actuels dans un cadre global basé sur un modèle. Ce modèle est interprétable et flexible pour la modélisation de l'effet de traitement et de l'influence du réseau. Nous démontrons que les estimateurs du maximum de vraisemblance sont convergents et asymptotiquement normaux pour une famille de spécifications du modèle. Des quantités causales d'intérêt, telles que l'effet global du traitement, sont définies et exprimées en tant que fonctions des paramètres du modèle GANE, et l'inférence peut donc être effectuée en conséquence. De plus, nous proposons la spécification ≪puissance-degré≫ (POW-DEG) du modèle GANE et étudions sa performance, conjointement avec les autres spécifications, au moyen de simulations.

[16:45-17:00]

Dena L. Schanzer (Public Health Agency of Canada (retired)) **Douglas G. Woolford** (University of Western Ontario)

Perspectives on Data Translation from Fields Outside of the Statistical Science

Perspectives sur la traduction de données provenant de domaines externes aux sciences statistiques

Considerable advances have been made in developing tools and guidelines for assessing evidence for policy recommendations, including evidence-informed decisions. Within Statistics these tend to focus on the quality of the analytical methods. Guidelines from other areas focus more on the overall quality of the estimates, while including an opinion on the methodology from a data expert. For example, the GRADE system guides how to rate the quality of evidence used in health-related decision making based on risk of bias, publication bias, imprecision (random error), inconsistency and indirectness. We highlight the GRADE system while also touching on some of the recommendations of commonly used tools and guidelines from other disciplines discussed in the edited volume "Applied Data Science: Data Translators Across the Disciplines" (Woolford et

Le développement d'outils et de lignes directrices a connu des avancées considérables pour évaluer les faits amassés aux fins de recommandations quant à des politiques, sans oublier les décisions axées sur les faits. En statistiques, on se concentre généralement sur la qualité des méthodes analytiques. Dans d'autres domaines, les lignes directrices se concentrent plutôt sur la qualité globale des estimations, tout en incluant une opinion d'un expert en données à propos de la méthodologie. Par exemple, le système ≪GRADE≫ enseigne comment noter la qualité de des faits amassés et utilisés dans une prise de décision relative à la santé basée sur le risque de biais, de biais de publication, d'imprécision (erreur aléatoire), d'incohérence et d'incertitude. Nous soulignons le système GRADE tout en abordant certaines recommandations d'outils et de lignes directrices couramment employés dans d'autres disciplines mentionnées dans le volume édité ≪Applied Data Science : Data Translators Across the Disci-

al., forthcoming)", as reviewers often find that the information required to assess someone else's analysis is often missing or difficult to obtain.

plines» (Woolford et al., à venir), puisque les évaluateurs trouvent souvent que l'information nécessaire pour évaluer l'analyse d'autrui est insuffisante ou difficile à obtenir.

On Causal Inference Methods and Applied Statistics
Des méthodes d'inférence causale et de la statistique appliquée

Chair/Président: Celia M.T. Greenwood

Room/Salle: TB 238

Date: Monday May 29 / lundi 29 mai

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-15:45]

Nirodha M. Epasinghege Dona (Simon Fraser University)

A Causal Investigation of Pace of Play in Soccer

Étude causale du rythme de jeu au soccer

This study provides a comprehensive investigation of playing with pace in soccer. The investigation begins by introducing two quantitative definitions of pace whose calculations are facilitated through the availability of player tracking data. In the study, the primary scientific question concerns whether playing with pace is an advantageous strategy in terms of playing style. This is a question that has not been adequately resolved in either soccer or hockey. Here, we use methods of causal inference to investigate the relationship between pace in soccer and shots. It is determined that playing with more pace than the opponent throughout a match confers an advantage of approximately two additional shots per game. As a byproduct of our analysis, other soccer insights related to pace are obtained.

Cette étude se penche de plus près sur le rythme du jeu au soccer. Nous introduisons tout d'abord deux définitions quantitatives du rythme dont le calcul est facilité par la disponibilité des données de suivi sur les joueurs. La question principale de cette étude est de savoir jouer avec rythme est une stratégie avantageuse sur le plan du style de jeu. C'est une question à laquelle aucune solution adéquate n'a été trouvée ni au soccer, ni au hockey. Ici, les méthodes d'inférence causale sont utilisées pour étudier la relation entre le rythme au soccer et les tirs. Nous avons déterminé qu'un jeu mieux rythmé que celui de l'adversaire pendant le match confère un avantage d'environ deux tirs de plus par match. Une compréhension d'autres aspects du soccer relatifs au rythme a découlé de notre analyse.

[15:45-16:00]

Jinko Graham (Simon Fraser University) **Payman Nickchi** (Simon Fraser University) **Charith Karunarathna** (Simon Fraser University)

An Exploration of Linkage Fine-Mapping on Sequences from Case-Control Studies

Une exploration de la cartographie fine de liaison sur les séquences tirées d'études cas-témoins

Linkage analysis maps genes for a heritable trait by identifying genomic regions with excess relatedness among individuals with similar trait values. Analysis may be conducted in families, or in unrelated individuals from a population. Here, we focus on linkage analysis in a population sample, but use sequences rather than individuals as our unit of observation. We propose two ways to associate similarity in relatedness of sequences with similarity in their trait values and compare the resulting linkage methods to two genotypic-association methods. We also introduce a procedure to label case sequences as potential carriers of causal variants, based

L'analyse de liaison cartographie les gènes à la recherche d'un trait héréditaire en identifiant les régions ayant un excès d'apparenté entre des individus ayant des valeurs de traits semblables. L'analyse peut être menée dans des familles, ou avec des individus sans lien de parenté dans une population. Ici, nous nous concentrons sur l'analyse de liaison dans un échantillon de population, mais utilisons des séquences plutôt que des individus en guise d'unité d'observation. Nous proposons deux façons d'associer la similarité aux séquences d'apparentés ayant des similitudes dans leurs valeurs de trait et comparer les méthodes de liaison obtenues à deux méthodes d'association génotypique. Nous présentons aussi une procédure pour étiqueter les séquences cas

on inferred relatedness to other case sequences. Our simulation results indicate that linkage methods based on sequence-relatedness improve gene localization and perform at least as well as genotypic-association methods for detecting its rare causal variants.

en tant que porteurs potentiels de variants causaux, basés sur l'apparentée inférée relative aux autres séquences cas. Nos résultats de simulation indiquent que les méthodes de liaison basées sur l'apparenté de séquence facilitent la localisation de gène et sont au moins aussi efficaces que les méthodes d'association génotypique dans la détection de variants causaux rares.

[16:00-16:15]

Yuliang Shi (University of Waterloo) **Yeying Zhu** (University of Waterloo) **Joel A. Dubin** (University of Waterloo)

Variable Selection for Causal Modeling in Missing Exposure Problems

Sélection de variables pour la modélisation causale dans des problèmes d'exposition manquante

In causal inference, properly selecting the propensity score (PS) model has been widely investigated in observational studies. However, when the data are missing at random (MAR) on the exposure, there are very few studies investigating how to simultaneously select both imputation and PS models, and providing a criterion for evaluating the overall performance of both models in a non-simulation setting. In this paper, we investigate the variable selection issue for both the imputation and PS models based on the different roles of the covariates represented in a causal diagram. Then, we propose a new criterion to evaluate the selected models, which combines the performance of the imputation and outcome models. The simulation studies show that the traditional criterion does not perform well when the exposure variable is MAR. In comparison, the newly proposed criterion builds up a strong correlation with RMSE, which can guide researchers to select the proper imputation and PS models.

De nombreuses études d'observation ont étudié la sélection adéquate de modèles de score de propension (PS) en inférence causale. Cependant, lorsqu'il y a des données manquantes aléatoirement (MAR) dans l'exposition, très peu d'études ont cherché à savoir comment choisir simultanément l'imputation et les modèles de PS, et comment obtenir des critères pour évaluer la performance globale des deux modèles dans un cadre hors simulation. Dans cet article, nous étudions le problème de sélection de variables pour l'imputation et les modèles de PS basés sur les différents rôles des covariables représentés dans le diagramme causal. Nous proposons ensuite un nouveau critère pour évaluer les modèles sélectionnés, ce qui combine la performance de l'imputation et des modèles de résultat. Les études en simulations démontrent que le critère traditionnel réagit mal lorsque la variable d'exposition est MAR. En comparaison, le nouveau critère proposé conserve une corrélation robuste avec RMSE, ce qui permet d'orienter les chercheurs pour choisir l'imputation et les modèles de PS adéquats.

[16:15-16:30]

Mengjie Bian (McMaster University) **Angelo J. Canty** (McMaster University)

Investigating the Winner's Curse in Mendelian Randomization

Enquête sur la malédiction du vainqueur dans la randomisation mendélienne

Mendelian Randomization (MR) uses genetic markers to estimate the causal effect of an exposure on an outcome. The markers are often selected based on their significance in a genome-wide association study (GWAS). The selection causes bias in the association estimates known as the "winner's curse" which may also result in the MR estimate of the causal effect being biased. We performed simulation studies to assess the potential bias from winner's curse in MR and applied existing methods to correct the genetic marker-exposure association which is then used in causal estimation. We also adapted an existing bootstrap method to work when only summary-level data is available and show our adapted

La randomisation mendélienne (MR) se sert de marqueurs génétiques pour estimer l'effet causal d'une exposition à un résultat. Les marqueurs sont souvent sélectionnés en fonction de leur pertinence dans une étude d'association pangénomique (GWAS). La sélection cause un biais dans les estimations de l'association, que l'on connaît sous le nom de ≪malédiction du vainqueur≫, qui peut aussi biaiser le résultat dans l'estimation de la MR de l'effet causal. Nous avons réalisé des études en simulations afin d'évaluer le biais potentiel de la malédiction du vainqueur dans la MR et avons appliqué les méthodes courantes pour corriger l'association marqueur-exposition génétique dont on se sert ensuite pour l'estimation causale. Nous avons aussi adapté une méthode existante par bootstrap pour qu'elle fonctionne lorsque les données sont

method works comparably to the original method. We show that winner's curse can bias causal estimates from MR. Some correction methods can reduce the bias at the expense of increasing the variability resulting in wider confidence intervals and relatively unchanged power to detect causal effects.

seulement disponibles au niveau agrégé et démontrons que cette méthode adaptée fonctionne de façon comparable à l'originale. Nous illustrons comment la malédiction du vainqueur peut biaiser les estimations causales tirées de la MR. Certaines méthodes de correction peuvent amoindrir, mais au prix d'une hausse de variabilité causant l'élargissement des intervalles de confiance et une puissance relativement inchangée pour repérer les effets causaux.

Chair/Président: Mai Ghannam

Room/Salle: TB 202

Date: Monday May 29 / lundi 29 mai

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-15:45]

Balage Don Harshani Hiranthika De Silva (University of Manitoba) **Saman Muthukumarana** (University of Manitoba) **Tony Szturm** (University of Manitoba)

Bayesian Analysis for Comparing Computerized Manual Dexterity Performance Metrics Among Children with Cerebral Palsy

Comparaison de mesures de dextérité manuelle informatisées chez les enfants atteints d'infirmité motrice cérébrale par analyse bayésienne

Cerebral Palsy (CP) is one of the most common motor neurodevelopmental disorders which affects three of every thousand live births in North America. In this study, we use a Bayesian approach to advance the science of manual dexterity outcome measurement for young children. Fifty children with a diagnosis of CP and fifty healthy, typically developing children between the ages of 4 to 12 years of age were enrolled in the study. Firstly, the frequentist Analysis of Variance (ANOVA) method was used and then for comparison, a Bayesian Analysis of Variance (BANOVA) method was applied to analyze the various manual dexterity metrics obtained from a novel computer game-based assessment tool. The Bayesian methods but not the typical ANOVA analysis revealed a significant age effect on all performance metrics. A significant difference in performance was also observed between the different objects tested. Finally, we develop a Bayesian Seemingly Unrelated Regression model with multiple responses.

L'infirmité motrice cérébrale (CP) est l'un des troubles neurodéveloppementaux moteurs les plus courants, touchant trois enfants sur mille naissances vivantes en Amérique du Nord. Dans cette étude, nous utilisons une approche bayésienne pour faire progresser la science de la mesure de la dextérité manuelle chez les jeunes enfants. Cinquante enfants ayant reçu un diagnostic de CP et cinquante enfants en bonne santé, au développement typique, âgés de 4 à 12 ans, ont été recrutés pour l'étude. Dans un premier temps la méthode fréquentiste d'analyse de la variance (ANOVA) a été utilisée, puis, à titre de comparaison, une méthode bayésienne d'analyse de la variance (BANOVA) a été appliquée pour analyser les différentes mesures de la dextérité manuelle obtenues à partir d'un outil d'évaluation basé sur un jeu informatique. Contrairement à l'analyse ANOVA classique, les méthodes bayésiennes ont révélé un effet significatif de l'âge sur toutes les mesures de performance. Une différence significative de performance a également été observée entre les différents objets testés. Enfin, nous développons un modèle bayésien de régression apparemment sans rapport, avec des réponses multiples.

[15:45-16:00]

Arjun Banik (University of Victoria) **Laura Cowen** (University of Victoria) **Saman Muthukumarana** (University of Manitoba) **Colin Gallagher** (Fisheries and Ocean, Canada)

Analyzing demographic parameters of anadromous Dolly Varden (Salvelinus malma malma) using Bayesian multi-state capture-recapture modelling

Analyse de données démographiques du Dolly Varden anadrome (Salvelinus malma malma) à l'aide d'une modélisation multi-états de capture-recapture bayésienne

For mark-recapture data, the Cormack-Jolly-Seber (CJS) model is extremely useful for analyzing the demographic characteristics of various species. Although the lack of sufficient observable data limits the precision of

Pour des données de marquage-recapture, le modèle Cormack-Jolly-Seber (CJS) est très utile pour l'analyse des caractéristiques démographiques de diverses espèces. Même si le manque de données observables suffisantes limite la précision des estima-

traditional estimations, Bayesian inference can be useful for handling the sparsity and acknowledges estimation uncertainty. We developed a novel Bayesian multi-state mark-recapture model, tested its performance on simulated data sets, and then applied it to the anadromous northern form of Dolly Varden collected from Firth River for estimating their breeding status (spawner or non-spawner), survival probability, and their state transition probabilities. Our model predicts, the likelihood of spawning in subsequent years is 55% and spawning to silver (non-spawning) has a higher probability than silver to spawning. Also, the estimates differ depending on sex. Despite Dolly Varden's complex biological characteristics, our model estimates appear to be biologically reliable.

tions traditionnelles, l'inférence bayésienne peut être utile pour traiter la parcimonie et reconnaître l'incertitude de l'estimation. Nous avons développé un modèle multi-états marquage-recapture bayésien, en avons évalué la performance avec des ensembles de données simulées et l'avons ensuite appliqué à une forme nordique anadrome du Dolly Varden dans la rivière Firth afin d'estimer l'état de frai (état reproducteur ou non-reproducteur) de ce poisson, la probabilité de survie et les probabilités de transition de l'état. Les prédictions de notre modèle : le taux de vraisemblance de la reproduction dans les années subséquentes est de 55 % et la probabilité de l'état de frai vers l'état de non-frai est plus élevée que celle de l'état de non-frai vers l'état de frai. De plus, les estimations diffèrent en fonction du sexe. En dépit des caractéristiques biologiques complexes du Dolly Varden, les estimations de notre modèle semblent fiables sur le plan biologique.

[16:00-16:15]

Masud Rana (University of Saskatchewan) **Shahedul Khan** (University of Saskatchewan) **Punam Pahwa** (University of Saskatchewan)

Covariate measurement error in spatial analysis of count data

Erreur de mesure de covariables dans l'analyse spatiale de données de dénombrement

Measurement error (ME) arises whenever we fail to observe a variable precisely. Covariate ME violates the assumption that covariates are fixed and accurately measured, resulting in biased parameter estimates and inflated variance estimates in a regression model. The objective is to investigate the impact of misspecified covariate ME in spatially correlated data. We formulated Bayesian hierarchical Poisson regression models with the error-prone covariates described by classical error model, Berkson error model, and no consideration of ME. We also considered three types of correlation structure in each model: no random effect, unstructured random effect, and the BYM2 model (an extension of Besag-York-Mollie [BYM] model). We presented comparative studies, illustrated via real and simulated data. The simulation results revealed that ignoring ME and conducting naive analysis for spatially correlated count data using the BYM2 model attenuates the regression coefficient towards the null effect.

Une erreur de mesure (EM) survient lorsqu'on n'observe pas précisément une variable. L'EM de covariables ne respecte pas l'hypothèse selon laquelle les covariables sont fixes et précisément mesurées, ce qui cause des estimations de paramètre biaisées et des estimations de variance exagérées dans un modèle de régression. L'objectif est d'étudier l'impact d'une EM de covariable mal spécifiée dans des données spatialement corrélées. Nous avons formulé des modèles de régression de Poisson hiérarchiques bayésiens avec des covariables sujettes à erreur et décrites par un modèle d'erreur classique, un modèle d'erreur de Berkson, et sans tenir compte de l'EM. Nous avons aussi examiné trois types de structure de corrélation dans chaque modèle : une sans effet aléatoire, une avec effet aléatoire non structuré, et une étant le modèle BYM2 (une extension du modèle Besag-York-Mollie [BYM]). Nous avons présenté des études comparatives, illustrées à partir de données simulées et réelles. Les résultats de simulation révèlent qu'ignorer l'EM et réaliser une analyse naïve pour des données de dénombrement corrélées spatialement au moyen de modèle BYM2 atténue le coefficient de régression vers l'effet nul.

[16:15-16:30]

Mingwei Xu (University of Waterloo)

Bayesian Integration Method for Parameter Estimation in Ordinary Differential Equation

Méthode d'intégration bayésienne pour l'estimation de paramètre dans une équation différentielle ordinaire

Parameter inference and uncertainty quantification of ordinary differential equations (ODEs) from noisy observations play a significant role in many scientific

L'inférence des paramètres et la quantification de l'incertitude des équations différentielles ordinaires (ODE) tirées d'observations bruyantes jouent un rôle important dans de nombreux domaines

fields. Currently, most parameter estimation methods that bypass numerical integration tend to rely on basis functions or Gaussian processes to approximate the ODE solution and its derivatives. Due to the sensitivity of the ODE solution to its derivatives, these methods can be hindered by estimation error, especially when only sparse time-course observations are available. We present a Bayesian collocation framework that operates on the integrated form of the ODEs and avoids the expensive use of numerical solvers. Our methodology has the capability to handle general nonlinear ODE systems. We demonstrate the accuracy of our method for estimating the parameters and recovering the system trajectories in a simulation study. A real data example is also provided.

scientifiques. Présentement, la plupart des méthodes d'estimation des paramètres qui contournent l'intégration numérique ont tendance à s'appuyer sur des fonctions de base ou des processus gaussiens pour approximer la solution de l'ODE et de ses dérivés. En raison de la sensibilité de la solution d'ODE à ses dérivés, ces méthodes peuvent être entravées par des erreurs d'estimation, tout particulièrement lorsque seules des observations temporelles éparses sont disponibles. Nous présentons un cadre de collocation bayésien qui opère sur la forme intégrée des ODE et évite l'utilisation coûteuse de résolveurs numériques. Notre méthodologie a la capacité de gérer des systèmes d'ODE non linéaires généraux. Nous démontrons la précision de notre méthode pour estimer les paramètres et récupérer les trajectoires du système dans une étude de simulation. Nous fournissons aussi un exemple avec des données réelles.

[16:30-16:45]

Ajmery Jaman (McGill University)

Penalized G-estimation for Effect Modifier Selection in the Structural Nested Mean Model for Repeated Outcomes

G-estimation pénalisée pour la sélection de modificateurs d'effet dans le modèle structurel de moyenne imbriquée pour des résultats répétés

Structural nested mean models (SNMMs) are useful causal models for estimating the potentially modified effect of a time-varying treatment on the mean of an outcome in the presence of time-varying confounding. Among the measured confounders, only some may cause heterogeneous treatment effects and a data-driven selection of these effect modifiers (EMs) maybe interesting, especially when the true EMs are unknown. We propose a penalized efficient score function using a non-convex smoothly clipped absolute deviation (SCAD) penalty for estimation of the causal effect of a treatment with simultaneous selection of EMs in the SNMM for repeated outcomes. We theoretically demonstrate the asymptotic properties of our doubly robust estimator. We also conducted simulation study that verifies the oracle property of the proposed penalized G-estimator. The method is applied for selecting the EMs for hemodiafiltration (HDF), a dialysis technique for treating patients with end-stage renal disease (ESRD).

Les modèles structurels de moyenne imbriquée (SNMM) sont des modèles causaux pratiques pour estimer l'effet potentiellement modifié du traitement variable avec le temps sur la moyenne d'un résultat en présence de facteurs de confusion variables avec le temps. Parmi les facteurs de confusion mesurés, seulement certains pourraient causer des effets de traitement hétérogènes. Il pourrait alors être avantageux de faire une sélection de ces modificateurs d'effet (EM) basée sur des données, particulièrement lorsque les EM réels sont inconnus. Nous proposons une fonction de score pénalisée efficace au moyen d'une pénalité non convexe à écart absolu (SCAD) pour estimer l'effet causal d'un traitement avec sélection simultanée d'EM dans le SNMN pour des résultats répétés. Nous démontrons théoriquement les propriétés asymptotiques de notre estimateur doublement robuste. Nous avons aussi mené des études en simulations qui vérifient la propriété de l'oracle du G-estimateur pénalisé proposé. La méthode est appliquée afin de choisir les EM pour l'hémodiafiltration, une technique de dialyse servant à traiter les patients souffrant d'insuffisance rénale terminale (IRSU).

[16:45-17:00]

Samuel Morrissette (University of Manitoba) **Saman Muthukumarana** (University of Manitoba) **Maxime Turgeon** (University of Manitoba)

Parsimonious Bayesian Model-Based Clustering with Dissimilarities

Regroupement avec dissimilarité basé sur un modèle bayésien parcimonieux

Clustering techniques are used to group observations and discover interesting patterns within data. Model-

Les techniques de regroupement servent à grouper les observations et découvrir des schémas intéressants dans les données.

based clustering is one such method that is an attractive choice due to the specification of a generative model for the data and the ability to calculate model-selection criteria. However, when only distances between observations are available, model-based clustering can no longer be used and heuristic algorithms without the aforementioned advantages are often used instead. As a solution, Oh and Raftery (2007) suggest a Bayesian model-based clustering method, named BMCD, that only requires a distance matrix as input. We extend the BMCD method by proposing additional models, alternative model selection criteria, strategies for reducing computing time, and providing a software implementation in the R programming language. These extensions ensure that the method is effective at capturing clusters of various shapes and sizes—even in high-dimensional space.

Le regroupement basé sur un modèle est l'une de ces méthodes et représente un choix avantageux en raison de la spécification d'un modèle génératif pour les données et la capacité de calculer les critères de sélection de modèles. Cependant, lorsqu'on a seulement les distances entre les observations, le regroupement à base de modèle devient inutilisable. Souvent, on adopte alors les algorithmes heuristiques, mais ils n'offrent pas les avantages décrits précédemment. En guise de solution, Oh et Raftery (2007) suggèrent une méthode de regroupement basée sur un modèle bayésien (BMCD) qui ne demande qu'une matrice de distance en guise d'information. Nous élargissons la méthode BMCD en proposant des modèles supplémentaires, des critères de sélection de modèles de rechange, des stratégies de réduction de temps de calcul, et procurons une intégration logicielle dans la langue de programmation R. Ces extensions assurent l'efficacité de la méthode pour saisir les regroupements de plusieurs formes et tailles, même dans un espace de haute dimension.

Data science methods and applications II
Méthodes et applications de la science des données II

Chair/Président: Pingzhao Hu

Room/Salle: TB 240

Date: Monday May 29 / lundi 29 mai

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-15:45]

Sebastian F. Calcetero (University of Toronto) **Andrei L. Badescu** (University of Toronto) **X. Sheldon Lin** (University of Toronto)

A Credibility Index Approach for Effective a Posteriori Ratemaking with Large Insurance Portfolios

Une approche d'indice de crédibilité pour une tarification a posteriori efficace avec de grandes portefeuilles d'assurance

A posteriori ratemaking in insurance aims to find premiums that account for policyholders' attributes and their claim history using a Bayesian model. Most data-driven Bayesian models are mathematically intractable, and thus, premiums must be obtained via numerical methods, e.g MCMC. However, these methods can be computationally expensive and prohibitive for large portfolios. Additionally, these methods are "black-box" procedures as there is no expression showing how the claim history is used to upgrade the premiums. In this talk, a methodology is proposed to derive a closed-form expression to compute credibility premiums for any given Bayesian model by introducing a credibility index that serves as an efficient summary statistic of a policyholder's claim history. By using this closed-form solution, the computational burden of a posteriori ratemaking for large portfolios can be reduced through the use of surrogate modeling, and it also provides a transparent way of computing premiums.

La tarification a posteriori en assurance vise à trouver des primes qui tiennent compte des attributs des assurés et de leur historique de réclamations en utilisant un modèle bayésien. La plupart des modèles bayésiens basés sur les données sont mathématiquement intractables et les primes doivent donc être obtenues par des méthodes numériques, comme MCMC. Cependant, ces méthodes peuvent être coûteuses et prohibitives pour de grands portefeuilles. Dans cette présentation, une méthodologie est proposée pour dériver une expression fermée pour calculer les primes de crédibilité pour tout modèle bayésien donné en introduisant un indice de crédibilité qui sert de statistique de synthèse efficace de l'historique des réclamations d'un assuré. En utilisant cette solution fermée, la charge de calcul de la tarification a posteriori pour les grands portefeuilles peut être réduite grâce à l'utilisation de modèles de substitution.

[15:45-16:00]

Shimon Nauenberg (University of Toronto) **Nicholas Mitsakakis** (CHEO Research Institute and University of Toronto)

Weak Supervision for Detecting Mental Health Issues from Social Media Data

Supervision faible pour la détection des problèmes de santé mentale à partir des données des médias sociaux

Abundant social media data can be used to train classification models for detecting mental health problems. However, when using supervised learning, domain experts must hand label posts. Here, we use weak supervision and estimate true depression labels of Reddit posts—a forum-based network– by using known from the literature or suspected correlates. We construct label functions - deterministic noisy labels of each post. By

Les données abondantes des médias sociaux peuvent être utilisées pour former des modèles de classification permettant de détecter les problèmes de santé mentale. Cependant, lors de l'utilisation de l'apprentissage supervisé, les experts du domaine doivent étiqueter manuellement les messages. Ici, nous utilisons une supervision faible et estimons les véritables étiquettes de dépression des messages de Reddit, un réseau basé sur des forums, en utilisant des corrélats connus de la littérature ou suspectés. Nous construisons

assuming a graphical model correlating the true label Y with the label functions, we estimate the posterior probability of Y, given the label function values. We convert these probabilities to binary labels (depressed/not depressed) and train a Support Vector Machine model, using text sentiment, polarity, and n-grams as features. We test our model against manually labeled data. Our study shows the usefulness of weakly supervised learning, for detecting mental health problems in social media posts.

des fonctions d'étiquetage - des étiquettes déterministes avec du bruit pour chaque message. En supposant un modèle graphique corrélant la véritable étiquette Y avec les fonctions d'étiquetage, nous estimons la probabilité postérieure de Y, étant donné les valeurs des fonctions d'étiquetage. Nous convertissons ces probabilités en étiquettes binaires (déprimé/non déprimé) et formons un modèle de machine à vecteur de support, en utilisant le sentiment du texte, la polarité et les n-grammes comme caractéristiques. Nous testons notre modèle sur des données étiquetées manuellement. Notre étude montre l'utilité de l'apprentissage faiblement supervisé pour détecter les problèmes de santé mentale dans les messages des médias sociaux.

[16:00-16:15]

Zhaoran Hou (University of Waterloo)

Estimating Boltzmann Averages for Protein Structural Quantities Using Sequential Monte Carlo

Estimer les moyennes de Boltzmann pour les quantités structurelles protéiques au moyen de la séquentielle de Monte Carlo

Sequential Monte Carlo (SMC) methods are widely used to draw samples from intractable target distributions. Particle degeneracy can hinder the use of SMC when the target distribution is highly constrained or multimodal. As a motivating application, we consider the problem of sampling protein structures from the Boltzmann distribution. We propose a general SMC method that propagates multiple descendants for each particle, followed by resampling to maintain the desired number of particles. Simulation studies demonstrate the efficacy of the method for tackling the protein sampling problem. As a real data example, we use our method to estimate the number of atomic contacts for a key segment of the SARS-CoV-2 viral spike protein.

Les méthodes séquentielles de Monte Carlo (SMC) sont largement employées pour tirer des échantillons à partir de distributions de cibles complexes. La dégénérescence de particule peut nuire à l'utilisation de la SMC lorsque la distribution de cible est hautement contrainte ou multimodale. En guise d'application illustrative, nous considérons le problème d'échantillonnage de structures protéiques tirées de la distribution de Boltzmann. Nous proposons une méthode SMC générale qui propage de multiples descendants pour chaque particule, suivi d'un rééchantillonnage pour conserver le nombre désiré de particules. Des études en simulations démontrent l'efficacité de notre méthode pour s'attaquer au problème d'échantillonnage protéique. En guise d'exemple avec des données réelles, nous utilisons notre méthode pour estimer le nombre de contrats atomiques pour un segment clé de la protéine de spicule virale SARS-CoV-2.

[16:15-16:30]

Jie Jian (University of Waterloo)

Dynamic Community Detection of the International Apple Trading Network

Détection dynamique de communautés dans le réseau international de commerce de pommes

Networks of agricultural trading amounts between countries over a period of time are of great interest. Partitioning the countries (nodes) into distinct communities, where countries within a community are more similar in connection patterns, is important for analyzing agricultural market trends and identifying potential opportunities, as well as the supply chain risks and vulnerabilities. Using the apple trading network as an example, we study the community detection in such a weighted time-varying network where each edge has a value associated with it and the value changes over

Les réseaux d'échanges agricoles entre pays sur une période donnée sont d'un grand intérêt. La répartition des pays (nœuds) en communautés distinctes, où les pays au sein d'une communauté sont plus similaires en termes de connexions, est importante pour analyser les tendances du marché agricole et identifier les opportunités potentielles, ainsi que les risques et les vulnérabilités de la chaîne d'approvisionnement. En utilisant le réseau commercial des pommes comme exemple, nous étudions la détection de communautés dans un tel réseau pondéré variable dans le temps, où chaque arête est associée à une valeur qui évolue dans le temps. Nous modélisons les données du réseau à partir d'un modèle sto-

time. We model the network data from a generalized stochastic block model where the edge weights follow a compound distribution, whose parameters are affected by both the community memberships of the nodes and some covariates dynamically. To make the model more robust and interpretable, we require smooth effects of the covariates over time. This is joint work with Dr. Peijun Sang and Dr. Mu Zhu.

chastique généralisé par blocs dans lequel les poids des arêtes suivent une distribution composée, dont les paramètres sont affectés à la fois par l'appartenance des nœuds à une communauté et par certaines covariables de manière dynamique. Pour rendre le modèle plus robuste et plus facile à interpréter, nous exigeons que les effets des covariables soient lissés dans le temps. Ce travail est réalisé en collaboration avec Peijun Sang et Mu Zhu.

[16:30-16:45]

McNealis Vanessa (McGill University) **Erica E. M. Moodie** (McGill University) **Nema Dean** (University of Glasgow)

Joint Mixed-Effects Models for Causal Inference in Network-Based Observational Studies

Modèles à effets mixtes conjoints pour l'estimation d'effets causaux en présence d'interférence réseau

Causal inference on populations embedded in social networks poses technical challenges, since the typical no interference assumption may no longer hold. Existing methods developed in the context of network interference rely upon the assumption of no unmeasured confounding. However, when faced with multilevel network data, there may be a latent factor influencing both the exposure and the outcome at the cluster level. We propose an approach that combines a parametric joint mixed-effects model for the outcome and the exposure with direct standardization to identify and estimate causal effects in the presence of network interference and unmeasured cluster confounding. In simulations, we compare our proposed method with linear mixed– and fixed effects models and show that unbiased estimation is achieved using the joint model. Having derived valid tools for estimation, we examine the effect of maternal college education on adolescent school performance using the Add Health data.

L'estimation d'effets causaux à partir de données de réseaux sociaux pose des défis de taille, puisque les méthodes conventionnelles d'inférence causale présupposent l'absence d'interférence entre les individus. De plus, les méthodes existantes fonctionnent en l'absence de facteur de confusion non mesuré. Cependant, en présence de données de réseaux multi niveaux, il est plausible qu'un facteur latent influence à la fois l'exposition et l'issue au niveau du groupement. Nous proposons un modèle paramétrique à effets mixtes conjoint pour l'exposition et l'issue pour estimer des effets causaux en présence d'interférence réseau et de facteurs de confusion non mesurés. Des simulations comparant le modèle conjoint à des modèles à effets fixes et mixtes ordinaires montrent que l'estimateur proposé parvient à atténuer le biais de confusion. À l'aide du modèle conjoint, nous investiguons l'effet de l'éducation maternelle sur la performance scolaire dans l'étude Add Health.

[16:45-17:00]

Karina Kwan (McGill University) **Celia M. T. Greenwood** (McGill University) **David Soave** (Wilfrid Laurier University)

A Weighted Casebase Framework for Analyzing Case-Cohort Study Data

Cadre de travail basé sur les cas pondérés pour l'analyse de données d'études cas-cohortes

Case-cohort studies are attractive for studying rare diseases where obtaining additional data from a subset of participants, such as genomic sequencing, is infeasible for the entire study cohort. In analyzing such studies, individuals must be appropriately weighted to account for the biased case/control sampling. The casebase framework uses logistic regression to estimate fully parametric hazard functions with the benefit of smooth-in-time absolute risk functions. We propose a weighted casebase framework to obtain unbiased estimates of relative and absolute risks using weighted logistic regres-

Les études cas-cohorte sont intéressantes pour l'étude des maladies rares, par exemple pour le séquençage génomique, lorsqu'il est impossible d'obtenir des données supplémentaires auprès de l'ensemble de la cohorte étudiée. Lors de l'analyse de ces études, les individus doivent être pondérés de manière appropriée pour tenir compte de l'échantillonnage biaisé cas/témoins. Le cadre de travail basé sur les cas utilise la régression logistique pour estimer des fonctions de risque entièrement paramétriques avec l'avantage que présentent les fonctions de risque absolu lissées dans le temps. Nous proposons un cadre de travail basé sur les cas pondérés pour obtenir des estimations non biaisées des risques relatifs et absolus

sion with the appropriate model offset. Standard errors for the risk estimates can be obtained through bootstrapping. We study the numerical properties of the approach in comparison with weighted Cox models through simulation studies. The proposed framework is applied to model cancer risk in a case-cohort data set from the European Prospective Investigation into Cancer and Nutrition study.

en utilisant la régression logistique pondérée avec le décalage de modèle approprié. Nous obtenons les erreurs standard pour les estimations de risque par bootstrapping. Nous étudions les propriétés numériques de l'approche et les comparons aux modèles de Cox pondérés par le biais d'études de simulation. Nous appliquons ensuite le cadre proposé à la modélisation du risque de cancer dans un ensemble de données de cohorte de cas provenant de l'Étude européenne prospective sur le cancer et la nutrition.

SSC 2022 Gold Medal Address
Allocution du récipiendaire de la Médaille d'or de la SSC 2022

Chair/Président: Wendy Lou

Organizer/Responsable: Wendy Lou

Room/Salle: RB 2200

Date: Tuesday May 30 / mardi 30 mai

Time/Heure: 08:30-09:50

Abstract/Résumé

[08:30-09:50]

David A. Stephens (McGill University)

What Makes a 'Bayesian' Analysis Bayesian ?

Qu'est-ce qui fait qu'une analyse bayésienne est « bayésienne » ?

The principles of Bayesian statistical inference, established from different formulations including the mathematical result known as de Finetti's representation, have been accepted as settled statistical theory for (at least) decades, and although the subjectivist origins are controversial to some, what constitutes a 'Bayesian' analysis would largely be agreed upon by all observers. Recently, however, there has been a perceptible shift away from fully Bayesian methods to more hybrid approaches that rely on combinations of ideas from different statistical paradigms, with an increasing focus on large sample and purely frequentist logic for justification. This focus often brings useful advances, but often also lacks an acknowledgement that such non-Bayesian compromises are being made. In this talk I will review some of the recent developments in this direction, and discuss when the 'Bayesian' label is justifiable and when it is more questionable.

Le principe d'inférence statistique bayésienne, établi à partir de différentes formulations comme le résultat mathématique connu sous le nom dereprésentation de Finetti, a été accepté en tant que théorie statistique depuis au moins des décennies. Son origine subjective peut être controversée pour certains, mais il en reste qu'une majorité d'observateurs s'accordent sur ce qui constitue une analyse bayésienne. Par contre, récemment, on remarque une déviation des méthodes purement bayésiennes vers des approches hybrides qui se fondent sur des combinaisons d'idées tirées de différents paradigmes statistiques et se concentrent de plus en plus sur de grands échantillons et une logique purement fréquentiste en guise de justification. Ces approches apportent souvent des avancées pratiques, mais on reconnaît peu que des compromis non bayésiens sont faits. Lors de cet exposé, je survolerai certains développements récents dans cette direction, et aborderai la question à savoir si le titre «bayésien» est justifiable ou plutôt discutable.

Practical challenges and novel methods for estimating causal effect from observational data
Défis pratiques et nouvelles méthodes d'estimation de l'effet causal à partir de données d'observation

Chair/Président: Mohammad Ehsanul Karim

Organizer/Responsable: Mohammad Ehsanul Karim

Room/Salle: AT 102

Date: Tuesday May 30 / mardi 30 mai

Time/Heure: 10:20-11:50

Abstract/Résumé

[10:20-10:50]

Lan Wen (University of Waterloo)

Robust Estimation of the Effect of Intervening Variable in Settings with Unmeasured Confounding

Estimation robuste de l'effet de la variable d'intervention dans les situations comportant des confondantes non mesurées

We present new results on average causal effects in settings with unmeasured exposure-outcome confounding. Our proposed causal estimands are motivated by practical questions corresponding to an average causal effect of an intervening variable, a manipulable descendant of the exposure that precedes the outcome. We argue that effects of intervening variables are replicable and have clear policy implications, and furthermore are non-parametrically identified by the classical frontdoor formula without assuming cross-world conditions or ill-defined exposure interventions. Semiparametric estimators that guarantee sample bounds are constructed for the proposed estimands. New theoretical results are illustrated using data from the National Health and Nutrition Examination Survey.

Nous présentons de nouveaux résultats concernant les effets causaux moyens dans des cadres comportant des confondantes non mesurées d'exposition et de résultat. Les estimateurs causaux que l'on propose sont motivés par des questions pratiques correspondant à l'effet causal moyen d'une variable intervenante, c'est-à-dire un descendant manipulable de l'exposition qui précède le résultat. Nous soutenons que les effets des variables intervenantes sont reproductibles et ont manifestement des répercussions sur les politiques. De plus, ils sont identifiés de façon non paramétrique par la formule classique de type "porte d'entrée" sans supposer des conditions qui prévalent dans tous les scénarios ou des interventions d'exposition mal définies. Nous construisons des estimateurs semi-paramétriques qui garantissent des limites d'échantillonnage pour l'estimateur proposé. Nous illustrons de nouveaux résultats théoriques au moyen de données tirées de la National Health and Nutrition Examination Survey.

[10:50-11:20]

Mireille Schnitzer (Université de Montréal) **Ashkan Ertefaie** (University of Rochester) **Denis Talbot** (Université Laval) **Guanbo Wang** (Harvard University) **David Berger** (Université de Montréal) **Jennifer O'Loughlin** (Université de Montréal) **Marie-Pierre Sylvestre** (Université de Montréal)

Longitudinal outcome-adaptive and marginal fused LASSO for confounder selection and model pooling with time-varying treatments

LASSO fusionné adaptatif longitudinal pour la sélection de facteurs de confusion et regroupement de modèles à temps de traitement variables

Data sparsity is a common problem when conducting causal inference with time-varying binary treatments, especially when treatment can change over many time-points. Many methods involve weighting by the inverse of the probability of treatment, which requires modeling the probability of treatment at each time point. Under

La rareté des données est un problème important pour l'inférence causale avec des traitements binaires variant dans le temps, en particulier lorsque le traitement peut changer sur de nombreux points dans le temps. Plusieurs méthodes utilisent la pondération par l'inverse de la probabilité de traitement, ce qui nécessite de modéliser la probabilité de traitement à chaque temps. Il est pos-

sparsity, it is possible to pool these models over time, but when correlations between covariates and treatment vary over time, this can lead to bias. Furthermore, with a large covariate space assumed to be a non-minimal sufficient adjustment set, reducing the adjustment set can greatly improve the variance of the estimator. We consider a novel approach to longitudinal confounder selection using a longitudinal outcome adaptive fused LASSO that will data-adaptively select covariates and collapse the treatment model parameters over time-points with the goal of improving the efficiency of the estimator while minimizing confounding bias.

sible de regrouper ces modèles dans le temps, mais lorsque les corrélations entre les covariables et le traitement varient dans le temps, cela peut entraîner des biais. De plus, avec un grand espace de covariables supposé être un ensemble d'ajustement suffisant non minimal, la réduction de l'ensemble d'ajustement peut grandement améliorer la variance de l'estimateur. Nous considérons une nouvelle approche de la sélection longitudinale des facteurs de confusion à l'aide d'un LASSO fusionné adaptatif longitudinal qui sélectionnera les covariables de manière adaptative aux données et réduira les paramètres du modèle de traitement au fil du temps dans le but d'améliorer l'efficacité de l'estimateur tout en minimisant le biais de confusion.

[11:20-11:50]

Denis Talbot (Université Laval)

A generalized double-robust Bayesian model averaging approach to causal effect estimation

Une approche généralisée et doublement robuste de moyenne de modèles bayésienne pour l'estimation d'un effet causal

We propose a Bayesian model averaging approach to confounder selection for estimating the effect of a binary or continuous exposure on a binary or continuous outcome. The approach is double robust, leading to less stringent modeling assumptions than either outcome-modeling or exposure-modeling based approaches. The Bayesian framework is used to yield inferences that account for the model selection in a principled way. The proposed implementation is sufficiently efficient from a computational perspective to be applied to analyze large databases, such as health administrative data. In simulation studies based on such data, our proposed approach was observed to be able to capitalize on proxies to reduce unmeasured confounders bias.

Nous proposons une approche bayésienne de moyenne de modèles pour la sélection des facteurs confondants afin d'estimer l'effet d'une exposition binaire ou continue sur une issue binaire ou continue. L'approche est doublement robuste et repose ainsi sur des hypothèses de modélisation moins contraignantes que les approches basées sur la modélisation de l'issue ou sur la modélisation de l'exposition. Le cadre bayésien est utilisé pour produire des inférences qui tiennent adéquatement compte de la sélection du modèle. L'implantation proposée est suffisamment efficace d'un point de vue computationnel pour permettre l'analyse de grandes bases de données, telles que les données médico-administratives. Dans des études de simulation basées sur de telles données, il a été observé que l'approche proposée était capable de capitaliser sur des proxies pour réduire les biais de confusion attribuables à des facteurs non mesurés.

The Collaborative Statistician: Learning the Tricks of the Trade
Le statisticien collaboratif : Apprendre les ficelles du métier

Chair/Président: Bouchra Nasri

Organizer/Responsable: Tolu Sajobi

Room/Salle: TB 210

Date: Tuesday May 30 / mardi 30 mai

Time/Heure: 10:20-11:50

Abstract/Résumé

[10:20-10:35]

Lehana Thabane (McMaster University)

Statistical Leadership in the Time of Crisis: Lessons from COVID Trials

Leadership statistique en temps de crise : Leçons des essais cliniques COVID

The unprecedented and global challenge of COVID-19 have brought together international expert researchers, clinicians, and methodologists to better understand the epidemiology of the disease and find appropriate cures. Collaborative statisticians have played important roles in the vaccines and drugs for COVID-19. Using several examples from Canada-led trials and other international projects, I will describe the type of leadership and innovations that collaborative statisticians bring to these collaborative teams in the development of life-saving interventions for COVID-19 disease.

Le défi mondial sans précédent qu'a représenté la COVID-19 a rassemblé chercheurs, cliniciens et méthodologistes internationaux afin de mieux comprendre l'épidémiologie de la maladie et de trouver des remèdes appropriés. Les statisticiens collaboratifs ont joué un rôle important dans la mise au point des vaccins et des médicaments contre la COVID-19. À l'aide d'exemples d'essais menés au Canada et d'autres projets internationaux, je décrirai le type de leadership et d'innovations que les statisticiens collaboratifs apportent à ces équipes collaboratives dans le développement d'interventions vitales pour la COVID-19.

[10:35-10:50]

Lisa M. Lix (University of Manitoba)

Lessons on Collaborations in Health Services & Policy Research: Integrating Knowledge Translation with Statistical Research

Leçons sur les collaborations dans le domaine de la recherche sur les services et les politiques de santé : Intégrer l'application des connaissances à la recherche statistique

Health funding agencies are increasingly requesting that proposals include researcher and decision-maker partnerships in collaborative and multidisciplinary teams, with representatives from communities, industry, and government agencies. In this talk I will provide examples of national and international collaborations in health services and health policy research that integrate multiple disciplines, including biostatistics. These examples will focus on national and international prescription drug safety and effectiveness studies, international health outcome investigations, and studies to develop and implement new patient-reported outcome measures in clinical settings. I will describe the roles that biostatisticians can play in study design, data quality eval-

Les organismes de financement de la santé demandent de plus en plus souvent que les propositions incluent des partenariats entre chercheurs et décideurs au sein d'équipes collaboratives et multidisciplinaires, avec des représentants des communautés, de l'industrie et des agences gouvernementales. Dans cet exposé, je donnerai des exemples de collaborations nationales et internationales dans le domaine de la recherche sur les services et les politiques de santé qui intègrent plusieurs disciplines, y compris la biostatistique. Ces exemples porteront sur des études nationales et internationales sur la sécurité et l'efficacité des médicaments sur ordonnance, des enquêtes internationales sur les résultats en matière de santé et des études visant à développer et à mettre en œuvre de nouvelles mesures de résultats rapportés par les patients en environnement clinique. Je décrirai les rôles que les biostatisticiens

uation, and methods development in policy-relevant research. I will also describe opportunities for biostatisticians to participate in integrated knowledge translation, and lessons that can be learned from these activities.

peuvent jouer dans la conception des études, l'évaluation de la qualité des données et le développement de méthodes dans la recherche pertinente pour les politiques. Je décrirai également les possibilités pour les biostatisticiens de participer à l'application intégrée des connaissances et les leçons qui peuvent être tirées de ces activités.

[10:50-11:05]

Joanna Elizabeth Mills Flemming (Dalhousie University)

Lessons Learned while Developing Statistical Methods to Better Understand the Ocean and its Inhabitants

Leçons apprises lors du développement de méthodes statistiques cherchant à comprendre davantage l'océan et ses habitants

I begin with a heterogenous mix of examples from my own research that demonstrate both the utility of statistical methods for improving our understanding of the ocean and its inhabitants, and the need for further developments. In doing so I provide descriptions of two very common sources of marine data: fisheries catches and surveys, and marine animal tracking systems. I conclude by highlighting statistical problems on the immediate ocean research horizon, and the technical challenges they pose.

Je débute par un mélange hétérogène d'exemples tirés de ma propre recherche qui illustre à la fois l'utilité des méthodes statistiques pour avancer nos connaissances sur l'océan et ses habitants, mais aussi le besoin accru de leurs développements. Ainsi, j'offre des descriptions de deux sources très courantes de données marines : des enquêtes et des prises de pêche, et les systèmes de repérage d'animaux marins. Je conclus en soulignant les problèmes statistiques dans l'horizon immédiat de la recherche océanique, ainsi que les défis techniques qu'ils génèrent.

[11:05-11:20]

Tolu Sajobi (University of Calgary)

Time is Brain: Lessons on Statistical Collaborations in a Fast-Paced Multidisciplinary World of Acute Stroke

Le temps, c'est du cerveau : Leçons sur les collaborations statistiques dans le monde multidisciplinaire et rapide de l'accident vasculaire cérébral aigu

The practice of biostatistics in multidisciplinary environments is exciting, professionally gratifying, and filled with several opportunities for a rewarding career. However, biostatisticians who work in these settings often operate in two different worlds (statistics vs applied research) while managing differences in expectations, cultures, responsibilities, and opportunities. They typically navigate several challenges, such as communication, managing workloads, negotiating power imbalances, reconciling competing work expectations, fostering relationships, developing contextual research knowledge, developing leadership roles. In this talk, I will share my experience working as the lead statistician for a world-class stroke program to support several ground-breaking studies and trials for acute stroke interventions. Lessons on important personal and professional skills needed to thrive in such a fast-paced multidisciplinary setting will be discussed.

La pratique de la biostatistique dans des environnements multidisciplinaires est passionnante, gratifiante sur le plan professionnel et offre de nombreuses possibilités de carrière enrichissante. Cependant, les biostatisticiens qui travaillent dans ces environnements évoluent souvent dans deux mondes différents (statistique et recherche appliquée) tout en gérant les différences d'attentes, de cultures, de responsabilités et d'opportunités. Ils doivent généralement relever plusieurs défis, tels que la communication, la gestion de la charge de travail, la négociation des déséquilibres de pouvoir, la conciliation d'attentes professionnelles concurrentes, le développement de relations, l'acquisition de connaissances contextuelles en matière de recherche et le développement d'un rôle de leader. Dans cet exposé, je partagerai mon expérience en tant que statisticien en chef d'un programme d'AVC de classe mondiale pour soutenir plusieurs études et essais révolutionnaires sur les interventions en cas d'AVC aigu. Je discuterai des leçons à tirer sur les compétences personnelles et professionnelles importantes nécessaires pour s'épanouir dans un environnement multidisciplinaire au rythme aussi rapide.

[11:20-11:50]

The Collaborative Statistician: Learning the Tricks of the Trade
Le statisticien collaboratif : Apprendre les ficelles du métier

Bouchra Nasri (Université de Montréal)

Panel Discussion and Q&A

Débat et questions-réponses

Statistical collaboration and/or consulting are the bridge through which statisticians contribute to meaningful multidisciplinary research. There is a growing demand for statisticians who are knowledgeable and adept at providing statistical leadership on multidisciplinary teams to advance scientific inquiries and innovations. The proposed session brings together leading methodologists and applied statisticians with a track record of success in collaborative research in the life sciences. Speakers will share their experience working in collaborative research environments and important skill set needed for effective collaboration. Attendees for this conference session will learn about the importance of statistical leadership in collaborative environments, the essential personal and professional skills needed to thrive in a multidisciplinary environment, and strategies to navigate the challenges that arise in these settings This session will be particularly useful for trainees, new investigators, established investigators, and those working in industry, government, and business.

La collaboration et/ou la consultation en statistique sont les passerelles par lesquelles les statisticiens contribuent à une recherche multidisciplinaire significative. Il y a une demande croissante pour des statisticiens compétents et capables de fournir un leadership statistique à des équipes multidisciplinaires pour faire avancer les enquêtes et les innovations scientifiques. La session proposée réunit méthodologistes et statisticiens appliqués de premier plan qui ont fait leurs preuves dans la recherche collaborative en sciences de la vie. Les intervenants partageront leur expérience dans des environnements de recherche collaborative et les compétences importantes nécessaires à une collaboration efficace. Les participants apprendront l'importance du leadership statistique dans les environnements collaboratifs, les compétences personnelles et professionnelles essentielles nécessaires pour prospérer dans un environnement multidisciplinaire, et les stratégies permettant de relever les défis qui se posent dans ces contextes. Cette session sera particulièrement utile pour les stagiaires, les nouveaux chercheurs, les chercheurs établis, et ceux qui travaillent dans l'industrie, le gouvernement, et les entreprises.

Statistics in Schools
La statistique dans les écoles

Chair/Président: Samantha-Jo Caetano

Organizer/Responsable: Samantha-Jo Caetano

Room/Salle: AT 101

Date: Tuesday May 30 / mardi 30 mai

Time/Heure: 10:20-11:50

Abstract/Résumé

[10:20-10:50]

Chrystal E. Smith (York University)

Using R-programming in the K-12 Classroom for Mathematical modelling and statistical literacy development.

Utiliser le logiciel R dans la salle de classe K-12 pour la modélisation mathématique et l'apprentissage statistique

This talk discusses intersections between the Ontario Curriculum for Mathematics for grades 7 and 8 (Ontario Ministry of Education 2020) and instances of technological and statistical learning that are embedded within the curriculum. It highlights opportunities for the use of R programming in the mathematics classroom to develop the technological and programming skills needed for further study in mathematics and statistics as well as the self-efficacy and problem-solving skills needed in the workplace. Some approaches to instructional design will be presented, with a focus on mathematical modelling so that students can use R in real life situations with creativity, simplicity, and confidence.

Je présente dans cette séance des intersections entre le curriculum de mathématiques de l'Ontario pour les 7e et 8e années (ministère de l'Éducation 2020) et des instances d'apprentissage statistiques et technologiques qui s'y trouvent. Il s'agit d'opportunités pour l'utilisation du logiciel R dans la salle de classe de mathématiques afin de développer les compétences techniques et de programmation nécessaires pour des études avancées en mathématiques et statistiques, ainsi que pour développer des compétences d'auto-efficacité et de résolution de problèmes en milieu de travail. Quelques approches pédagogiques seront présentées en mettant l'accent sur la modélisation mathématique de sorte que les élèves puissent utiliser le logiciel R dans des situations de la vraie vie avec confiance, créativité et en toute simplicité.

[10:50-11:20]

Diana Katherine Skrzydlo (University of Waterloo)

High School Outreach – Machine Learning, Finance, and Data Bias

Rayonnement au niveau secondaire – Apprentissage automatique, finance et biais dans les données

Over the last few years I've had the opportunity to give presentations to several high school groups relating to statistics. There are many benefits of academics doing this kind of outreach work: students (and their teachers!) enjoy hearing about the practical applications of what they are seeing in class, and are curious about the use of statistics in their future careers; and you get to share your knowledge and get students excited about statistics. In this presentation, I will share details about the topics I spoke about, including machine learning, finance, and data bias, as well as some advice if you find yourself with a similar opportunity.

Au cours des dernières années, j'ai eu l'occasion de donner des présentations à plusieurs groupes d'écoles secondaires portant sur les statistiques. Les retombées pour les universitaires qui font ce type de travail de rayonnement sont nombreuses : les étudiants (et leurs enseignants !) aiment entendre parler des applications pratiques de ce qu'ils voient en classe et sont curieux de l'utilisation des statistiques dans leur carrière ; et vous avez l'occasion de partager vos connaissances et de soulever l'enthousiasme des étudiants pour les statistiques. Dans cette présentation, je partagerai des détails sur les sujets dont j'ai parlé, y compris l'apprentissage automatique, la finance et le biais de données, ainsi que quelques conseils si vous vous retrouvez avec une opportunité similaire.

[11:20-11:50]

Nathalie Sjarova (Statistics Canada) **Georgia Fakiolas** (Statistics Canada / Statistique Canada) **Dmitri Bakker** (Statistics Canada / Statistique Canada) **Miranda Cox** (Statistics Canada / Statistique Canada) **Warren McDougald** (Statistics Canada / Statistique Canada)

Statistics Canada Census Education Toolkits

Trousses éducatives du Recensement de Statistique Canada

As more data sources are competing for students' attention, the need for credible, trustworthy information is more vital than ever. Statistics Canada has released a series of educational toolkits to help educators bring 2021 Census data into classrooms, libraries and lecture halls across the country. As Canada's national statistical office, Statistics Canada is committed to fostering statistical literacy and building the next generation of census data users. These user-friendly toolkits are designed for use in elementary and secondary-school classrooms, and by CEGEP, college and university students looking to use census data in their research. This presentation will provide an overview on the 2021 Census Toolkits and how they can be used by post-secondary students and professors.

Alors que de plus en plus de sources de données attirent l'attention des étudiants, le besoin en information crédible et fiable est plus important que jamais. Statistique Canada a lancé une série de trousses éducatives afin d'aider les éducateurs à intégrer les données du Recensement de 2021 dans les classes, les bibliothèques, et les amphithéâtres à travers le pays. En tant qu'organisme national de statistique du Canada, Statistique Canada est déterminé à appuyer la littératie statistique et l'éducation de la prochaine génération d'utilisateurs des données du recensement. Ces trousses conviviales sont conçues pour être utilisées dans le cadre de l'enseignement primaire et secondaire. Elles s'adressent aussi aux étudiants des cégeps, des collèges et des universités qui souhaitent utiliser les données du recensement dans le cadre de leurs recherches. Cette présentation fournira un aperçu des Trousses du Recensement de 2021 et des manières dont elles peuvent être utilisées par les étudiants et professeurs du niveau postsecondaire.

Recent Advances in mixture-model based classification
Avancées récentes de la classification basée sur des modèles de mélange

Chair/Président: Ryan P Browne

Organizer/Responsable: Ryan P Browne

Room/Salle: AT 301

Date: Tuesday May 30 / mardi 30 mai

Time/Heure: 10:20-11:50

Abstract/Résumé

[10:20-10:50]

Jeffrey L. Andrews (University of British Columbia, Okanagan)

Variable Importance in Clustering with Finite Mixture Models

Importance des variables dans le regroupement avec des modèles de mélange finis

We introduce a feature ranking approach specific to clustering with finite mixture models and consider its potential usage in a variable selection context. Our approach focuses on treating the log-likelihood as the optimization target, and measures changes therein upon marginal permutation of the variables. We apply this approach to several real and simulated data sets, and include discussion of usage with stability selection and other broadly applicable variable selection techniques.

Nous introduisons une approche de classement des caractéristiques spécifique au regroupement avec des modèles de mélanges finis et examinons son utilisation potentielle dans un contexte de sélection de variables. Notre approche se concentre sur le traitement de la log-vraisemblance comme objectif d'optimisation, et mesure les changements qui y sont liés lors de la permutation marginale des variables. Nous appliquons cette approche à plusieurs ensembles de données réelles et simulées, et nous incluons une discussion sur l'utilisation avec la sélection de stabilité et d'autres techniques de sélection de variables largement applicables.

[10:50-11:20]

Sanjeena Dang (Carleton University) **Anjali Silva** (University of Toronto) **Xiaoke Qin** (Carleton University) **Steven Rothstein** (University of Guelph) **Paul David McNicholas** (McMaster University)

Clustering matrix-variate discrete data

Regroupement de données discrètes à variables matricielles

Three-way data structures or matrix-variate data are frequent in biological studies. In RNA sequencing, three-way data structures are obtained when high-throughput transcriptome sequencing data are collected for n genes across p conditions at r occasions. Matrix variate distributions offer a natural way to model three-way data. In this work, a mixture of matrix variate Poisson-log normal distributions is proposed for clustering read counts from RNA sequencing. By considering the matrix variate structure, the number of covariance parameters to be estimated is reduced and the components of resulting covariance matrices provide a meaningful interpretation. We propose three different frameworks for parameter estimation - a Markov chain Monte Carlo based approach, a variational Gaussian approximation-based

Les structures de données à trois voies ou les données à variables matricielles sont fréquentes dans les études biologiques. Dans le domaine du séquençage de l'ARN, des structures de données à trois voies sont obtenues lorsque des données de séquençage du transcriptome à haut débit sont collectées pour n gènes dans p conditions à r occasions. Les distributions de variables matricielles offrent un moyen naturel de modéliser les données à trois voies. Dans le cadre de ces travaux, un mélange de distributions de variables matricielles Poisson-log normal est proposé pour regrouper les nombres de lectures provenant du séquençage de l'ARN. La prise en compte de la structure de variables matricielles permet de réduire le nombre de paramètres de covariance à estimer, et les composants des matrices de covariance obtenues fournissent une interprétation significative. Nous proposons trois cadres différents pour l'estimation des paramètres : une approche reposant sur la

approach, and a hybrid approach. The models are applied to both real and simulated data, and we demonstrate that the proposed approaches can recover the underlying cluster structure.

méthode de Monte-Carlo par chaînes de Markov, une approche reposant sur l'approximation gaussienne variationnelle et une approche hybride. Les modèles sont appliqués à des données réelles et simulées, et nous démontrons que les approches proposées peuvent récupérer la structure sous-jacente des grappes.

[11:20-11:50]

Alexander Sharp (University of Waterloo) **Ryan P. Browne** (University of Waterloo)

Maximum Contribution to the Likelihood: Precise Estimation with the Stochastic Expectation-Maximization Algorithm

Contribution des maxima à la vraisemblance : Estimation précise avec l'algorithme espérance-maximisation stochastique

The stochastic EM algorithm replaces the E-step with a Monte Carlo approximation, trading monotonicity for the potential to escape local maxima. Estimation techniques include averaging the tail of the chain, and choosing the value in the chain associated with the largest likelihood value. We demonstrate that the latter estimator diverges from the maximum likelihood estimate with high probability as the dimensionality of the parameter increases, but that it is also more precise in terms of chain length when the parameter is a scalar. Based on these findings, we propose a new estimator which we prove achieves this same level of precision for inference of multidimensional parameters. Simulation studies demonstrate the benefits of the proposed estimator when compared to topical approaches. Some discussion is given on limitations of the proposed estimator and potential avenues for increasing applicability are considered.

L'algorithme EM stochastique remplace l'étape E par une approximation de Monte Carlo, troquant la monotonicité contre la possibilité d'échapper aux maxima locaux. Les techniques d'estimation comprennent le calcul de la moyenne de la queue de la chaîne et le choix de la valeur de la chaîne associée à la plus grande valeur de vraisemblance. Nous démontrons que ce dernier estimateur diverge de l'estimation du maximum de vraisemblance avec une probabilité élevée lorsque la dimensionnalité du paramètre augmente, mais qu'il est également plus précis en termes de longueur de chaîne lorsque le paramètre est un scalaire. Sur la base de ces résultats, nous proposons un nouvel estimateur dont nous prouvons qu'il atteint ce même niveau de précision pour l'inférence de paramètres multidimensionnels. Des études de simulation démontrent les avantages de l'estimateur proposé par rapport aux approches topiques. Nous discutons des limites de l'estimateur proposé et envisageons des pistes potentielles pour en augmenter l'applicabilité.

Data Science Ethics and Fairness
Éthique et équité en science des données

Chair/Président: Joel Ostblom

Organizer/Responsable: Joel Ostblom

Room/Salle: TB 208

Date: Tuesday May 30 / mardi 30 mai

Time/Heure: 10:20-11:50

Abstract/Résumé

[10:20-11:05]

Sina Fazelpour (Northeastern University, Canada)

Disciplining deliberation: On interpreting machine learning trade-offs in sociotechnical systems

Discipliner la délibération : de l'interprétation des compromis de l'apprentissage automatique dans les systèmes sociotechniques

Formal analyses of decision scenarios routinely uncover trade-offs between different desiderata of value to decision-makers. This paper focuses on two highly publicized formal trade-offs in prediction-based decision-making—between predictive accuracy and fairness and between predictive accuracy and interpretability. In many cases, these trade-offs are taken at face value by researchers, practitioners, and policy-makers alike. That is, these trade-offs between different formal properties of predictive algorithms are seen as statements about practical value tensions in the deployment behavior of the sociotechnical systems in which those algorithms are embedded. Thus interpreted, the trade-offs have formed a core focus of normative engagement with the relevant technologies, offering both a particular division of labor between technical and ethical efforts, and a specific framing of the ethical questions. This paper argues against this attitude. We offer four sets of reasons for exercising caution in interpreting these trade-offs, pertaining to their (i) conceptual validity, (ii) empirical, (iii) compositional, and (iv) dynamic significance. We show how the resulting misinterpretations induce a pernicious myopia on our normative engagements with and aspirations for predictive technologies. We end by drawing out the epistemic and ethical implications of these considerations for the evaluation, design, and governance of ML-based decision-making.

Les analyses formelles des scénarios de décision révèlent régulièrement des compromis entre différents critères de valeur pour les décideurs. Cet article se concentre sur deux compromis formels très médiatisés dans la prise de décision basée sur la prédiction - entre la précision de la prédiction et l'équité et entre la précision de la prédiction et l'interprétabilité. Dans de nombreux cas, ces compromis sont pris au pied de la lettre par les chercheurs, les praticiens et les décideurs. En d'autres termes, ces compromis entre différentes propriétés formelles des algorithmes prédictifs sont considérés comme des déclarations sur les tensions de valeur pratique dans le comportement de déploiement des systèmes sociotechniques dans lesquels ces algorithmes sont intégrés. Ainsi interprétés, les compromis ont constitué un point central de l'engagement normatif avec les technologies concernées, offrant à la fois une division particulière du travail entre les efforts techniques et éthiques, et une formulation spécifique des questions éthiques. Le présent document s'inscrit en faux contre cette attitude. Nous proposons quatre séries de raisons de faire preuve de prudence dans l'interprétation de ces compromis, concernant leur (i) validité conceptuelle et leur signification (ii) empirique, (iii) compositionnelle et (iv) dynamique. Nous montrons comment les interprétations erronées qui en résultent induisent une myopie pernicieuse sur nos engagements normatifs et nos aspirations en matière de technologies prédictives. Nous terminons en soulignant les implications épistémiques et éthiques de ces considérations pour l'évaluation, la conception et la gouvernance de la prise de décision basée sur l'apprentissage automatique.

[11:05-11:50]

Heather Krause (We All Count)

Data Equity: How not to use data like a racist
Équité des données : comment ne pas utiliser les données comme un raciste

Collecting, analyzing and communicating data are not neutral activities. The process of creating statistical evidence is very much dependent on the worldview and cultural values of the people involved - from those designing the data collection, those doing the analysis, and those being studied. Even such an apparently simple mathematical process as taking an average is highly influenced by your point of view. In this talk, we will explore tools and processes we can use to help surface and correct unintentional social bias, sexism, racism, colonialism, and other types of latent discrimination in the way we collect data, analyze data and communicate about data. The talk will demonstrate a framework that formalizes the statistical evidence-generation processes within an equity context and provides structure to statistical research projects. The authors develop the Data Equity Framework and describe the seven steps required to fully operationalize this tool. The framework is then applied to real-world case studies.

La collecte, l'analyse et la communication de données ne sont pas des activités neutres. Le processus de création de preuves statistiques dépend fortement de la vision du monde et des valeurs culturelles des personnes impliquées, qu'il s'agisse de celles qui conçoivent la collecte des données, de celles qui effectuent l'analyse ou de celles qui sont étudiées. Même un processus mathématique apparemment aussi simple que l'établissement d'une moyenne est fortement influencé par votre point de vue. Dans cet exposé, nous explorerons les outils et les processus que nous pouvons utiliser pour aider à mettre en évidence et à corriger les préjugés sociaux involontaires, le sexisme, le racisme, le colonialisme et d'autres types de discrimination latente dans la manière dont nous collectons et analysons les données et communiquons à propos d'elles. L'exposé présentera un cadre qui formalise les processus de production de données statistiques dans un contexte d'équité et qui structure les projets de recherche statistique. Les auteurs développent un Cadre d'équité des données et décrivent les sept étapes nécessaires pour rendre cet outil pleinement opérationnel. Le cadre est ensuite appliqué à des études de cas réels.

Spotlight on CANSSI postdocs
Vitrine des postdocs de l'INCASS

Chair/Président: Andrea Benedetti

Organizer/Responsable: Andrea Benedetti

Room/Salle: RB 2200

Date: Tuesday May 30 / mardi 30 mai

Time/Heure: 10:20-11:50

Abstract/Résumé

[10:20-10:38]

Gracia Dong (University of Toronto - St. Michael's Hospital)

Analyzing Indian Mortality with Temperature Attribution using Case-Crossover Models

Analyse de la mortalité indienne avec attribution de température au moyen de modèles de cas croisés

The majority of studies on the health effects of temperature either focus on high-income countries or on the effects of extreme temperature events such as heat waves in low- and middle- income countries. Following the work in Fu et al. (PLoS Med (2018)), we assess the effects of apparent temperature on medical causes of death for adults, children and neonates in India. Using hourly gridded temperature data at a 0.25 * 0.25 degree resolution and daily mortality data from the Million Deaths Study from 2004 to 2013, we quantify the effects of both hot and cold temperature on mortality using an overdispersed case-crossover model.

La majorité des études sur les effets de la température sur la santé se concentrent soit sur les pays à revenu élevé, soit sur les effets de températures extrêmes comme des vagues de chaleur dans les pays à faible ou moyen revenu. En poursuivant le travail de Fu et al. (PLoS Med (2018)), nous évaluons les effets de la température apparente sur les causes médicales de décès pour les adultes, les enfants et les nouveau-nés en Inde. Au moyen de données rectangulaires sur la température à une résolution de 0,25 par 0,25 degré et des données quotidiennes sur la mortalité tirées de la Million Deaths Study de 2004 à 2013, nous quantifions les effets des températures chaudes et froides sur la mortalité à l'aide d'un modèle de cas croisé surdispersé.

[10:38-10:56]

François-Michel Boire (HEC Montréal) **Hatem Ben-Ameur** (HEC Montréal) **Mark Reesor** (Wilfrid Laurier University) **Lars Stentoft** (Western University)

Capital Structure Modelling with Dynamic Programming Under Alternative Processes

Modélisation de la structure du capital par programmation dynamique en présence des processus alternatifs

We develop a stochastic dynamic program (SDP) for valuing American-style options when the underlying follows the Constant Elasticity of Variance (CEV) model of Cox (1975) and the Normal Inverse Gaussian (NIG) Lévy process of Barndorff-Nielsen (1995). We discuss how the extension of the option pricing framework to alternative stochastic processes can be applied to a capital structure model with a contingent claim approach to equity valuation as seen in Merton (1973), Geske (1977, 1979), and Leland (1994). Since firm asset values are not observable on a regular basis, we propose an extension to the quasi-maximum likelihood (QML) approach of Ben-Abdellatif et al. (2021) where

Cet article développe un programme dynamique stochastique (SDP) pour l'évaluation d'options de style américain lorsque le modèle sous-jacent suit le modèle CEV (Constant Elasticity of Variance) de Cox (1975) et le processus de Lévy NIG (Normal Inverse Gaussian) de Barndorff-Nielsen (1995). L'extension du cadre d'évaluation des options à ces processus stochastiques alternatifs est appliquée à un modèle de structure du capital dans lequel l'évaluation des actions est basée sur les actifs de la firme, tel que dans les modèles de Merton (1973), Geske (1977, 1979) et Leland (1994). Puisque la valeur des actifs de l'entreprise n'est pas observable sur une base régulière, nous proposons une extension de l'approche de quasi-maximum de vraisemblance (QML) de Ben-Abdellatif et al. (2021) où les données sur les actions et

both equity and options data can be used jointly to calibrate model parameters. Next, we empirically compare the importance of model features like variance elasticity and non-Gaussian innovations for computing yield spreads and default probabilities.

les options peuvent être utilisées conjointement pour calibrer les paramètres du modèle. Ensuite, nous comparons empiriquement l'importance des caractéristiques du modèle telles que l'élasticité de la variance et les innovations non gaussiennes pour l'analyse de la structure à terme des écarts de rendement et des probabilités de défaut.

[10:56-11:14]

Ismaila Ba (York University)

Linear response variational Bayes (LRVB) approximation for a zero-inflated probabilistic PCA (ZIPPCA) model

Approche bayésienne variationnelle de réponse linéaire (LRVB) pour un modèle d'ACP probabiliste avec excès de zéros (ZIPPCA)

Dimension reduction techniques are among the most essential tools in the analysis of high-dimensional data. Generalized principal component analysis (PCA) is an extension to standard PCA that has been widely used to identify low-dimensional features in high-dimensional discrete data. For microbiome data in particular, the multinomial PCA is a natural counterpart of the standard PCA but fails to account for the excessive number of zero values. To allow for sparsity, zero-inflated multivariate distributions can be used. We propose a ZIPPCA model for extracting microbial compositions and a classification variational approximation algorithm to fit the model. Furthermore, we use the mean-field type variational family to approximate the posterior distribution and a LRVB approximation to provide improved covariance estimates. Simulation study and an application to a microbiome dataset are presented to compare the performance of the proposed model with that of the existing methods.

Les techniques de réduction des dimensions sont parmi les outils les plus essentiels dans l'analyse des données en haute dimension. L'analyse en composantes principales (ACP) généralisée est une extension de l'ACP standard qui a été largement utilisée pour identifier les caractéristiques de faible dimension dans les données discrètes en haute dimension. Pour les données du microbiome en particulier, l'ACP multinomiale est une contrepartie naturelle de l'ACP standard, mais elle ne tient pas compte du nombre excessif de valeurs nulles. Pour tenir compte de l'éparsité, il est possible d'utiliser des distributions multivariées avec excès de zéros. Nous proposons un modèle ZIPPCA pour extraire les compositions microbiennes et un algorithme d'approximation variationnelle de classification pour ajuster le modèle. En outre, nous utilisons la famille variationnelle de type champ moyen pour approximer la distribution des probabilités a posteriori et une approximation LRVB pour fournir des estimations de covariance améliorées. Nous présentons une étude de simulation et une application à un ensemble de données de microbiome pour comparer la performance du modèle proposé avec celle des méthodes existantes.

[11:14-11:32]

Dylan Z. Spicker (McGill University) **Erica E. M. Moodie** (McGill University)

Preserving Patient Privacy in Dynamic Treatment Regimes

Préservation de la vie privée des patients dans les régimes de traitement dynamiques

Precision medicine is a branch of evidenced-based medicine where individual level characteristics are used to tailor treatment recommendations. Dynamic treatment regimes (DTRs) provide one framework to formalize precision medicine. DTRs rely on sensitive patient information and as such there is a strong need to guarantee privacy for individuals' data in this setting. There is a growing literature which demonstrates that the publication of summary statistics or model parameters has the capacity to leak private information about the individuals in the dataset, even when the data are themselves protected. Differential privacy has emerged as a popular

La médecine de précision est une branche de la médecine factuelle dans laquelle les caractéristiques individuelles sont utilisées pour adapter les recommandations de traitement. Les régimes de traitement dynamiques (RTD) constituent un cadre permettant de formaliser la médecine de précision. Ces régimes s'appuient sur des informations sensibles concernant les patients et il est donc indispensable de garantir la confidentialité des données individuelles dans ce contexte. De plus en plus d'ouvrages démontrent que la publication de statistiques sommaires ou de paramètres de modèles peut entraîner la fuite d'informations privées sur les personnes figurant dans l'ensemble de données, même lorsque les données sont elles-mêmes protégées. La confi-

standard for providing rigorous guarantees on the level of privacy afforded to individuals. We demonstrate the need to consider rigorous privacy standards for DTRs, present a differentially private DTR estimation technique, and illustrate the trade-offs between privacy and utility which must be balanced in this setting.

dentialité différentielle est devenue une norme populaire pour fournir des garanties rigoureuses sur le niveau de protection de la vie privée des individus. Nous démontrons la nécessité d'envisager des normes de confidentialité rigoureuses pour les RTD, présentons une technique d'estimation de RTD différentiellement confidentielle et illustrons les compromis entre la confidentialité et l'utilité qui doivent être faits dans ce contexte.

[11:32-11:50]

Antonio Herrera Martin (University of Toronto) **Gwendolyn Eadie** (University of Toronto) **Radu Craiu** (University of Toronto) **David Stenning** (Simon Fraser University) **Derek Bingham** (Simon Fraser University) **Bryan Gaensler** (University of Toronto)

Classification of FRBs

Classification des sursauts radio rapides (FRB)

Astronomy is a data-rich scientific discipline with multiple opportunities for collaboration with statisticians. Our research focuses on data for fast radio bursts (FRBs), which are radio signals produced by unknown phenomena outside our galaxy. It was initially assumed they were caused by a process that would destroy the source of the signal, causing them not to repeat. However, it was found that some of the signals do indeed repeat without exhibiting a specific period. This study's first scientific objective is to classify repeaters, i.e., sources that emit multiple FRBs over time, and non-repeaters based on measured characteristics. The project is the result of a collaboration between astronomers and statisticians, developed within the Canadian Hydrogen Intensity Mapping Experiment (CHIME). We will discuss previous work in this area along with some criticisms. New regression-based methodology will be introduced and numerically illustrated using real and simulated data.

L'astronomie, une discipline scientifique riche en données, offre de multiples possibilités de collaboration avec les statisticiens. Notre recherche est axée sur les données relatives aux sursauts radio rapides (FRB), qui sont des signaux radio produits par des phénomènes inconnus à l'extérieur de notre galaxie. On a supposé au départ que les FRB étaient causés par un processus qui détruisait la source du signal, empêchant par conséquent leur répétition. On a cependant découvert qu'en réalité certains signaux se répétaient sans manifester de période précise. Le premier objectif de cette étude est de classifier les répéteurs, c.-à-d. les sources qui émettent de multiples FRB au fil du temps et les non-répéteurs en fonction des caractéristiques mesurées. Ce projet qui découle d'une collaboration entre astronomes et statisticiens a été développé dans le cadre de l'expérience canadienne de cartographie de l'intensité de l'hydrogène (Canadian Hydrogen Intensity Mapping Experiment, CHIME). Nous discuterons de travaux antérieurs dans ce domaine et de certaines critiques. Une nouvelle méthodologie basée sur la régression sera présentée et une illustration numérique en sera faite à l'aide de données réelles et simulées.

Recent Advances in Small Area Estimation
Progrès récents en estimation des petites surfaces

Chair/Président: Zeinab Mashreghi

Organizer/Responsable: Mahmoud Torabi

Room/Salle: TB 236

Date: Tuesday May 30 / mardi 30 mai

Time/Heure: 10:20-11:50

Abstract/Résumé

[10:20-10:50]

Jonnagadda N. Rao (Carleton University) **Victor Esteavo** (Statistics Canada) **Jean-François Beaumont** (Statistics Canada) **Keven Bosa** (Statistics Canada)

Comparison of area level and unit level estimators when unit level data are available

Comparaison des estimateurs au niveau du domaine et au niveau d'unité quand des données au niveau de l'unité sont disponibles

Fay (2018) proposed the use of survey regression estimators, based on unit level data, as direct estimators in the area level model. He demonstrated that the resulting EBLUP estimators are comparable in terms of efficiency to the EBLUP estimator under the unit level model. We report the results of a simulation study on efficiency comparisons under normal and log-normal unit errors. We also study the estimation of mean squared prediction error of the estimators in terms of relative bias and coefficient of variation. We then study the effect of outliers in the unit level data on the relative performance of the area level and unit level estimators in terms of efficiency. Extension to informative sampling will also be studied.

R. Fay (2018) a proposé d'utiliser les estimateurs par régression d'enquête, basés sur les données au niveau d'unité, comme estimateurs directs dans le modèle au niveau du domaine. Il a montré que sur le plan de l'efficacité, les estimateurs EBLUP (meilleur prédicteur linéaire sans biais empirique) obtenus sont comparables à l'estimateur EBLUP dans le modèle au niveau de l'unité. Nous faisons état des résultats d'une étude de simulation portant sur les comparaisons de l'efficacité sous erreurs de l'unité normales et log-normales. Nous étudions également l'estimation de l'erreur de prédiction par moindres carrés des estimateurs quant à leurs biais relatif et coefficient de variation. Nous examinons aussi l'effet des valeurs aberrantes des données au niveau de l'unité sur la performance relative sur le plan de l'efficacité des estimateurs au niveau du domaine et au niveau de l'unité. Une extension à un échantillonnage informatif est aussi à l'étude.

[10:50-11:20]

Samuel Sombo (Statistics Canada) **Jean-François Beaumont** (Statistics Canada) **Keven Bosa** (Statistics Canada) **Cynthia Bocci** (Statistics Canada)

Using Random Forest for Small Area Estimation

L'utilisation de forêts aléatoires pour l'estimation sur petits domaines

When domain sample sizes are small, design-consistent direct estimators can be unstable. To improve the precision of direct estimators, the Fay-Herriot (FH) area level model is often used. In its original form, the FH model assumes a linear linking model with constant error variance. It also requires knowing the variance of direct estimators. Direct variance estimators can be considered but they are generally unstable for small sample

Avec une petite taille d'échantillon dans les domaines, les estimateurs directs convergents par rapport au plan de sondage peuvent être instables. Pour améliorer la précision de ces estimateurs, on utilise souvent le modèle de Fay-Herriot (FH) au niveau des domaines. À la base, le modèle de FH suppose un modèle de liaison linéaire avec variances des erreurs constante et des estimateurs directs connue. Les estimateurs directs de la variance sont généralement instables pour de petites tailles d'échantillon, mais

sizes. To solve this problem, the variance estimates are usually smoothed. The assumptions underlying the FH and smoothing models are not always satisfied in practice, and it is not always easy to correct the models adequately. In this context, it may be desirable to have access to non-parametric methods as random forests because they depend less strongly on the validity of model assumptions. In this presentation, we will expose the proposed approach and show its properties through an empirical study.

lisser les estimations de variance aide à régler ce problème. Or, les hypothèses des modèles de FH et de lissage ne sont pas toujours satisfaites et il n'est pas évident de corriger adéquatement les modèles. Dans ce cas, utiliser une méthode non paramétrique comme les forêts aléatoires peut aider, car elle dépend moins de la validité des hypothèses des modèles. Dans cette présentation, nous exposerons l'approche proposée ainsi que ses propriétés au moyen d'une étude empirique.

[11:20-11:50]

Gauri Datta (U.S. Census Bureau) **Juhyung Lee** (University of Florida) **Jiacheng Li** (University of Georgia)

Pseudo-Bayesian Small Area Estimation

Estimation pseudo-bayésienne pour petits domaines

In small area estimation (SAE), sample sizes from various subpopulations or areas are often too small to accurately estimate its mean, and one borrows strength from similar areas through appropriate models based on useful covariates. The empirical best linear unbiased prediction method has been the dominant frequentist approach in SAE. It relies on estimation of model parameters based on the marginal distribution of the data. The observed best prediction (OBP) is an alternative method that estimates the parameters by minimizing an objective function derived from the total mean squared prediction error. We use this objective function to construct a pseudo-posterior distribution for the model parameters. The proposed Bayesian method is computational, easily implementable, provides nonnegative estimates of uncertainty of point estimates and reasonable interval estimates of the small area means. Applications to real data and simulations studies show that the proposed method is competitive.

Dans l'estimation de petits domaines (SAE), les tailles d'échantillons tirés de plusieurs sous-populations ou domaines sont souvent trop petites pour estimer avec précision sa moyenne, et l'on tire une robustesse à partir de domaines similaires grâce à des modèles adéquats basés sur des covariables utiles. La meilleure méthode empirique de prédiction non biaisée linéaire est l'approche fréquentiste dominante pour la SAE. Elle dépend de l'estimation des paramètres de modèles basés sur la distribution marginale des données. La meilleure prédiction observée (OBP) est une autre option qui estime les paramètres en minimisant une fonction objective dérivée du total de l'erreur de prédiction quadratique moyenne. Nous utilisons cette fonction objective pour construire une pseudo-distribution a posteriori pour les paramètres de modèle. La méthode bayésienne proposée est calculatoire, aisément intégrable et procure des estimations non négatives de l'incertitude des estimations ponctuelles ainsi que des estimations d'intervalle raisonnables des moyennes de petits domaines. Des applications dans des études en simulations et à partir de données réelles illustrent la compétitivité de la méthode proposée.

Chair/Président: Yifan Li

Room/Salle: RB 1200

Date: Tuesday May 30 / mardi 30 mai

Time/Heure: 10:20-11:50

Abstract/Résumé

[10:20-10:35]

Deli Li (Lakehead University) **Yu Miao** (Henan Normal University, China) **Yongcheng Qi** (University of Minnesota Duluth, USA)

A new type of results on probabilities of moderate deviations for i.i.d. random variables

Nouveau type de résultats sur les probabilités de déviations modérées pour les variables aléatoires iid

Let $\{X, X_n; n \geq 1\}$ be a sequence of i.i.d. non-degenerate real-valued random variables with $\mathbb{E}X^2 < \infty$. Let $S_n = \sum_{i=1}^n X_i,\ n \geq 1$. Let $g(\cdot): [0,\infty) \to [0,\infty)$ be a nondecreasing regularly varying function with index $\rho \geq 0$ and $\lim_{t\to\infty} g(t) = \infty$. Let μ and σ^2 be the mean and variance of X. In this work, on the scale $g(\log n)$, we obtain precise asymptotic estimates for probabilities of moderate deviations of the form $\log \mathbb{P}\left(S_n - n\mu > x\sqrt{ng(\log n)}\right)$, $\log \mathbb{P}\left(S_n - n\mu < -x\sqrt{ng(\log n)}\right)$, and $\log \mathbb{P}\left(|S_n - n\mu| > x\sqrt{ng(\log n)}\right)$ for all $x > 0$. Unlike those known results in the literature, the results established depend on both the variance and the asymptotic behavior of the tail distribution of X.

On suppose que $\{X, X_n; n \geq 1\}$ est une séquence de variables aléatoires iid non dégénérées à valeurs réelles avec $\mathbb{E}X^2 < \infty$. On suppose que $S_n = \sum_{i=1}^n X_i$, $n \geq 1$. On suppose que $g(\cdot): [0,\infty) \to [0,\infty)$ est une fonction régulièrement variable non décroissante d'indice $\rho \geq 0$ et $\lim_{t\to\infty} g(t) = \infty$. On suppose que μ et σ^2 sont la moyenne et la variance de X. Dans le cadre de ces travaux, à l'échelle $g(\log n)$, nous obtenons des estimations asymptotiques précises pour les probabilités d'écarts modérés de la forme $\log \mathbb{P}\left(S_n - n\mu > x\sqrt{ng(\log n)}\right)$, $\log \mathbb{P}\left(S_n - n\mu < -x\sqrt{ng(\log n)}\right)$, et $\log \mathbb{P}\left(|S_n - n\mu| > x\sqrt{ng(\log n)}\right)$ pour tout $x > 0$. Contrairement aux résultats connus dans la littérature, les résultats établis dépendent à la fois de la variance et du comportement asymptotique de la queue de distribution de X.

[10:35-10:50]

Ziang Zhang (University of Toronto) **Alex Stringer** (University of Waterloo) **Patrick Brown** (University of Toronto) **James Stafford** (University of Toronto)

Bayesian Smoothing and Inference of Derivatives through Integrated Wiener Processes

Lissage bayésien et inférence de dérivées par l'entremise de processus intégrés de Wiener

In many applications that involve the inference of an unknown smooth function, the inference of its derivatives will often be just as important as that of the function itself. To make Bayesian inferences of both the function and its derivatives, the p-th order Integrated Wiener's Process (IWP) is considered as the prior. Methods for constructing a finite element (FEM) approximation of an IWP exist but do not allow appropriate inference for derivatives. In this work, we propose a new FEM approximation based on overlapping splines (O-spline), which provides consistent inference for all derivatives.

Dans plusieurs applications comportant l'inférence d'une fonction lisse inconnue, l'inférence à propos de ses dérivées est souvent aussi importante que celle de la fonction elle-même. Afin de réaliser des inférences bayésiennes pour la fonction ainsi que pour ses dérivées, nous considérons le processus intégré de Wiener (IWP) d'ordre p comme l'a priori. Il existe des méthodes de construction d'une approximation d'élément fini (FEM) d'un IWP, mais elles ne permettent pas une inférence appropriée des dérivées. Dans le cadre de ce travail, nous proposons une nouvelle approximation de FEM basée sur les splines superposées (O-spline), qui procurent une inférence cohérente pour tous les

It is shown both analytically and empirically that the proposed approximation converges to the true IWP as the number of knots increases. We further provide a unified and interpretable way to define priors for the smoothing parameter. Finally, we apply the proposed method for the analysis of COVID death rates where the inference is carried on both the function and its derivatives.

dérivées. On démontre analytiquement et empiriquement que l'approximation proposée converge vers le vrai IWP au fur et à mesure que le nombre de nœuds augmente. Nous procurons de plus une manière unifiée et interprétable de définir les a priori du paramètre de lissage. Enfin, nous appliquons la méthode proposée à l'analyse du taux de mortalité de la COVID dans laquelle l'inférence est réalisée à partir de la fonction et de ses dérivées.

[10:50-11:05]

Kilani Ghoudi (United Arab Emirates University) **Mohamed Chaouch** (Qatar University) **Naamane Laib** (CY University)

Joint parametric specification checking of conditional mean and volatility in time series models with martingale difference innovations

Vérification de spécification paramétrique conjointe de la moyenne conditionnelle et volatilité dans des modèles de séries temporelles avec innovations de la différence de martingale

Using cumulative residual processes, we introduce powerful joint specification tests for conditional mean and variance functions in the context of nonlinear time series with martingale difference innovations. The main challenge comes from the fact that, the cumulative residual process does not admit a distribution-free limit. A solution consists in transforming the process to achieve a distribution-free limit or in approximating the limit distribution using numerical or re-sampling techniques. The three approaches are considered and compared. The tests have nontrivial power against a class of root-n local alternatives and are suitable when the conditioning set is infinite-dimensional. Numerical results show that the powers of tests based on re-sampling or numerical approximation are in general slightly better than those based on martingale transformation.

Au moyen de processus résiduels cumulatifs, nous présentons de puissants tests de spécification conjoints pour la moyenne conditionnelle et les fonctions de variances dans un contexte de séries chronologiques non linéaires avec innovations de la différence de martingale. Le défi principal provient du fait que le processus résiduel cumulatif n'a pas une limite sans distribution. En guise de solution, on peut transformer le processus pour obtenir une limite sans distribution ou approximer la distribution limite à l'aide de techniques numériques ou de rééchantillonnage. Nous évaluons et comparons les trois approches. Les tests ont une puissance non négligeable contre une classe d'alternatives locales de racine n et sont appropriés lorsque l'ensemble de conditionnement est de dimension infinie. Les résultats numériques démontrent que les puissances de tests basés sur le rééchantillonnage ou l'approximation numérique sont en général légèrement supérieures à celles des tests basés sur la transformation de martingale.

[11:05-11:20]

Haodi Liang (University of British Columbia) **Haodi Liang** (University of British Columbia) **Jiahua Chen** (University of British Columbia)

Global Consistency of Empirical Likelihood

Cohérence globale et vraisemblance empirique

The overwhelmingly favoured maximum likelihood estimator (MLE) is renowned for its strong consistency and optimality. When the likelihood is smooth, we often solve the score equation and regard the solution obtained as the MLE. If multiple solutions exist, it can be difficult to judge whether the solution at hand is the MLE. Empirical likelihood is a popular platform for non-parametric inference. The resulting profile EL function possesses many nice properties of a parametric likelihood and can be used to make inference for unknown parameters. However, similar issue arises

L'estimateur de choix du maximum de la vraisemblance (MLE) est reconnu pour sa forte cohérence et optimalité. Lorsque la vraisemblance est lisse, on résout souvent l'équation de score et percevons la solution obtenue comme MLE. Par contre, il peut être difficile de juger si la solution obtenue est MLE lorsque plusieurs solutions existent. La vraisemblance empirique est une plateforme populaire pour l'inférence non paramétrique. La fonction de profil EL générée possède une variété de propriétés avantageuses d'une vraisemblance paramétrique et peut servir à effectuer une inférence pour des paramètres inconnus. Cependant, des problèmes similaires surviennent lorsque la fonction de profil a

when the profile EL function has multiple local maxima. Hence it is important to find conditions under which the global maximum of the EL function is strongly consistent. We can enlarge the set of estimating functions under EL to ensure the global consistency in many cases. Furthermore, the global consistency result motives an approach to judging if a local maximum is in fact global.

plusieurs maximums locaux. C'est pourquoi il est important de trouver des conditions selon lesquelles le maximum global de la fonction EL est fortement cohérent. Dans plusieurs cas, nous pouvons élargir l'ensemble de fonctions d'estimation selon l'EL pour assurer la cohérence globale. De plus, le résultat de cohérence globale justifie une approche pour juger si un maximum local est effectivement global.

[11:20-11:35]

Pankaj Uttam Bhagwat (University of Sherbrooke) **William Strawderman** (Rutgers University, USA)

On the construction of Bayes minimax multiple shrinkage estimators

Construction d'estimateurs minimax de rétrécissement multiples bayésiens

For the canonical problem of estimating a multivariate normal mean under squared-error loss, we study the construction of minimax Bayes multiple shrinkage estimators. Multiple shrinkage estimators are beneficial in practice as they resolve the dilemma of choosing the minimax shrinkage estimator when conflicting prior knowledge suggests that more than one estimator can be effective. We show that, when particular spherical priors are used which lead to scalable marginal densities, the Stein's minimaxity condition of superharmonicity of the square root of the marginal density gives rise to a feasible method for constructing minimax pseudo-Bayes multiple shrinkage estimators. The general framework developed also leads to construction of proper priors resulting in admissible minimax multiple shrinkage estimators. Most notably, we show that adjusted Strawderman-type priors yield proper Bayes minimax (and admissible) multiple shrinkage estimators. These include frequently used t-priors.

En réponse au problème canonique de l'estimation de la moyenne normale multivariée sous perte d'erreur quadratique, nous étudions la construction d'estimateurs minimax de rétrécissement multiples bayésiens. En pratique, ceux-ci sont avantageux puisqu'ils résolvent le dilemme du choix de l'estimateur minimax de rétrécissement lorsque des connaissances a priori contradictoires laissent supposer que plus d'un estimateur peut être performant. Nous montrons que dans le cas où sont utilisés des a priori sphériques particuliers qui entraînent des densités marginales évolutives, la condition minimax de Stein de superharmonicité de la racine carrée de la densité marginale engendre une méthode de construction d'estimateurs de rétrécissement minimax multiples pseudo-bayésiens. Le cadre général développé entraîne aussi la construction d'a priori adéquats qui donnent pour résultat des estimateurs de rétrécissement minimax multiples admissibles. Nous montrons aussi que les a priori ajustés de type Strawderman produisent des estimateurs adéquats (et admissibles) de rétrécissement minimax multiples bayésiens. Ils comprennent des a priori (t) souvent utilisés.

[11:35-11:50]

Yunhong Lyu (University of Windsor) **Sévérien Nkurunziza** (University of Windsor)

On generalized square-root diffusion processes and Inference problems

Sur les processus de diffusion de type racines carrées et problèmes d'inférence

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 Cox-Ingersoll-Ross process. We also study an inference problem concerning the drift parameter of the proposed process. We focus on the scenario where the parameter might satisfy some restrictions. Regardless the difficulties due to the absence of an explicit solution and to the non-stationarity of the proposed process, we derive the Unrestricted Estimator and the Restricted Estimator as well as their joint asymptotic normality. We also establish a statistic test for testing

Dans cet exposé, nous proposons un process stochastique adapté à des données financières positives assujetties à une tendance de retour à la moyenne cyclique. Le modèle proposé est une généralisation du process de Cox-Ingersoll-Ross. De plus, nous étudions un problème d'inférence statistique concernant le paramètre de dérive dudit processus. À ce sujet, nous considérons le scénario où le paramètre de dérive pourrait satisfaire à certaines restrictions. Malgré les difficultés dues à l'absence de la solution explicite et à la non-stationnarité du processus, nous établissons l'estimateur sans restriction (EC) et l'estimateur restreint (RE) ainsi que leur normalité asymptotique conjointe. Nous élaborons également les estimateurs à rétrécissement ainsi qu'un test statis-

the hypothesized restriction and propose some shrinkage estimators. Further, we derive the asymptotic power of the proposed test as well as the asymptotic relative efficiency of the proposed estimators. The obtained simulation results corroborate our theoretical results.

tique pour tester la restriction. De plus, nous établissons la puissance asymptotique du test élaboré ainsi que l'efficacité asymptotique des estimateurs proposés. Les résultats de simulation corroborent nos résultats théoriques.

Chair/Président: Yan Yuan

Room/Salle: TB 238

Date: Tuesday May 30 / mardi 30 mai

Time/Heure: 10:20-11:50

Abstract/Résumé

[10:20-10:35]

Mehdi Dagdoug (University of Ottawa) **Camelia Goga** (Université de Franche-Comté) **David Haziza** (University of Ottawa)

High-dimensional model-assisted estimation for survey data

Estimation assistée par modèle dans un contexte de haute dimension pour des données d'enquête

In surveys, covariates are often used at the estimation stage to improve the efficiency of point estimators. The customary asymptotic framework in surveys assumes that both the population and the sample sizes increase to infinity, whereas the number of auxiliary variables is assumed to be fixed. This framework is appropriate when the number of auxiliary variables is negligible with respect to the sample size. In practice, however, one may encounter situations where this assumption is not satisfied. We investigate the impact of the number of auxiliary variables on the efficiency of point estimators. Sufficient conditions are obtained for the high-dimensional consistency of several model-assisted estimators. We also investigate the properties of several variance estimation procedures in a high-dimensional setting. We present the results from a simulation study that assesses the behavior of commonly used variance estimation procedures in high dimensions.

Dans les enquêtes, les covariables sont souvent utilisées à l'étape de l'estimation pour améliorer l'efficacité des estimateurs ponctuels. Le cadre asymptotique habituel des enquêtes suppose que la taille de la population et la taille de l'échantillon s'accroissent à l'infini et que le nombre de variables auxiliaires est fixe. Ce cadre est approprié lorsque le nombre de variables auxiliaires est négligeable quant à la taille de l'échantillon. En pratique, on retrouve cependant des situations où cette supposition n'est pas satisfaite. Nous examinons l'impact du nombre de variables auxiliaires sur l'efficacité des estimateurs ponctuels. On obtient des conditions suffisantes pour la cohérence en haute dimension de plusieurs estimateurs assistés par modèle. Nous examinons aussi les propriétés de plusieurs procédures d'estimation de la variance dans un contexte de haute dimension. Nous présentons enfin les résultats d'une étude par simulation qui évalue le comportement de procédures d'estimation de la variance couramment utilisées dans des contextes de haute dimension.

[10:35-10:50]

Boyi Hu (Simon Fraser University)

Simultaneous Functional Quantile Regression

Régression fonctionnelle simultanée des quantiles

We propose a novel framework that multiple quantiles can be considered simultaneously. This can overcome the two major limitations of the existing methods - namely the existing methods cannot guarantee the monotonicity of conditional quantiles or control the smoothness of the slope function estimator as a bivariate function in the estimation procedure. With the help of a bivariate spline basis, we propose a constrained objective function that includes the information of multiple

Nous proposons un nouveau cadre dans lequel plusieurs quantiles peuvent être pris en compte simultanément. Cela permet de surmonter les deux principales limitations des méthodes existantes, à savoir qu'elles ne peuvent pas garantir la monotonicité des quantiles conditionnels ou contrôler la régularité de l'estimateur de la fonction de pente en tant que fonction bivariée dans la procédure d'estimation. À l'aide d'une base spline bivariée, nous proposons une fonction objective contrainte qui inclut l'information des quantiles multiples. La méthode proposée permet de surmonter les

quantiles. The proposed method can overcome the two limitations mentioned above.

deux limitations mentionnées ci-dessus.

[10:50-11:05]

Andrea Payne (Carleton University) **Anjali Silva** (University of Toronto) **Steven Rothstein** (University of Guelph) **Paul David McNicholas** (McMaster University) **Sanjeena Dang** (Carleton University)

Utilizing a Family of Mixtures of Multivariate Poisson Log-Normal Distributions for Clustering High Dimensional Multivariate Count Data

Utilisation d'une famille de mélanges de distributions log-normales de Poisson multivariées pour le regroupement de données de comptage multivariées en haute dimension

Mixtures of multivariate Poisson lognormal (MPLN) models have used efficiently to analyze multivariate count data seen in bioinformatics, which is high dimensional and frequently exhibits over-dispersion In the MPLN model, the latent variable comes from a multivariate Gaussian distribution and the counts, which are conditional on this latent variable, are modeled using a Poisson distribution. As well, the MPLN model can account for over-dispersion and allows for correlation between the variables due its hierarchical structure. Here, we extend the mixture of multivariate Poisson-log normal distributions for high dimensional data by incorporating a factor analyzer structure in the latent space. A parsimonious family of mixtures of Poisson log-normal distributions are proposed by decomposing the covariance matrix and imposing constraints on these decompositions. We demonstrate the performance of the model using simulated and real datasets.

Les mélanges de modèles multivariés Poisson log-normal (MPLN) ont été utilisés efficacement pour analyser les données de comptage multivariées observées en bioinformatique, qui sont hautement dimensionnelles et présentent souvent une surdispersion. Dans le modèle MPLN, la variable latente provient d'une distribution gaussienne multivariée et les comptages, qui sont conditionnels à cette variable latente, sont modélisés à l'aide d'une distribution de Poisson. En outre, le modèle MPLN peut tenir compte de la surdispersion et permet une corrélation entre variables grâce à sa structure hiérarchique. Ici, nous étendons le mélange de distributions de Poisson log-normal multivariées pour les données en haute dimension en incorporant une structure d'analyseur de facteurs dans l'espace latent. Nous proposons une famille parcimonieuse de mélanges de distributions de Poisson log-normal en décomposant la matrice de covariance et en imposant des contraintes sur ces décompositions. Nous démontrons la performance du modèle en utilisant des ensembles de données simulées et réelles.

[11:05-11:20]

Yuan Zhong (York University) **Wei Xu** (University Health Network, University of Toronto) **Xin Gao** (York University)

High-dimensional Data Integration with Multiple Heterogeneous Tasks

Intégration de données en haute dimension avec de multiples tâches hétérogènes

Data integration is the process of extracting information from multiple sources and analyzing different related data sets simultaneously. In this paper, we propose to use the mixed L21-regularized composite quasi-likelihood function to perform multi-task feature learning with different types of responses, including continuous and discrete responses. For high dimensional settings, our result establishes the sign recovery consistency and estimation error bounds of the penalized estimates under regularity conditions. Simulation studies and real data analysis examples are provided to illustrate the utility of the proposed method to combine correlated platforms with heterogeneous tasks and perform the joint sparse estimation.

L'intégration de données est un processus d'extraction d'informations à partir de sources multiples et d'analyse simultanée de différents ensembles de données connexes. Dans cet article, nous proposons d'utiliser la fonction de quasi-vraisemblance composite mixte L21-régularisée pour effectuer un apprentissage multitâche des caractéristiques avec différents types de réponses, notamment continues et discrètes. Pour les environnements en haute dimension, notre résultat établit la cohérence de la récupération du signe et les limites de l'erreur d'estimation des estimations pénalisées dans des conditions de régularité. Nous fournissons des études de simulation et des exemples d'analyse de données réelles pour illustrer l'utilité de la méthode proposée pour combiner des plates-formes corrélées avec des tâches hétérogènes et effectuer l'estimation éparse conjointe.

[11:20-11:35]

Victoire Michal (McGill University) **Jon Wakefield** (University of Washington) **Alexandra Schmidt** (McGill University)

Comparison of Small Area Prediction methods with high dimensional auxiliary information

Comparaison de méthodes de prédiction pour des petits domaines en présence d'information auxiliaire de haute dimension

We propose a comparison of different variable selection methods for model-based small area estimation when the number of areas with sampled data is a small fraction of the total areas for which estimates are required. However, there are abundant auxiliary data for both the sampled and the non-sampled areas. Three areal-level procedures are studied: a frequentist approach with forward variable selection based on the AIC; a Bayesian shrinkage method with a horseshoe prior; a random forest prediction approach. Further, to measure the uncertainty of random forest estimates, we propose a modification of the split conformal procedure that relaxes the assumption of identically distributed data points. The proposed procedure to compute prediction intervals is compared to the original approach in a simulation study. The three variable selection methods are compared in a simulation study considering different relationships between the outcome and the covariates and different sampling designs.

Nous nous proposons de comparer différentes méthodes de sélection de variables pour l'estimation de petits domaines lorsque le nombre de domaines avec des données échantillonnées est une petite fraction du nombre total de domaines pour lesquels des estimations sont requises. De plus, de nombreuses données auxiliaires sont disponibles pour les domaines échantillonnés et non échantillonnés. Trois méthodes appliquées au niveau des domaines sont étudiées : une approche fréquentiste avec sélection ascendante de variables basée sur l'AIC ; une méthode bayésienne utilisant la distribution a priori Horseshoe ; une approche de prédiction par forêts aléatoires. De plus, pour mesurer l'incertitude des estimations par forêts aléatoires, nous proposons une modification de la méthode ≪ split conformal ≫ qui assouplit l'hypothèse de données identiquement distribuées. La procédure proposée pour calculer les intervalles de prédiction est comparée à l'originale dans une étude de simulation. Les trois méthodes de sélection de variables sont comparées dans une étude de simulation prenant en compte différentes relations entre la variable réponse et les covariables ainsi que différents plans d'échantillonnage.

Chair/Président: Sayantee Jana

Room/Salle: TB 240

Date: Tuesday May 30 / mardi 30 mai

Time/Heure: 10:20-11:50

Abstract/Résumé

[10:20-10:35]

Chel Hee Lee (Alberta Health Services & University of Calgary) **Jason Waechter** (University of Calgary) **Christopher Doig** (University of Calgary) **Mikelis Bickis** (University of Saskatchewan)

Imprecise Inference for Clinical Reasoning under Near or Complete Ignorance

Inférence imprécise pour le raisonnement clinique selon une ignorance complète ou presque complète

Physicians' clinical diagnostic decision-making involves cognitive reasoning skills and disease-specific knowledge to inform the collection and analysis of patient data which is not always sufficient and readily available. When information is little or unavailable, modeling a near- or complete-prior ignorance is not free from the dogma of precision in Bayesian inference. We adopted Walley (1991) 's imprecise probabilities and quantified the uncertainty associated with a diagnostic clinical reasoning course that measured the students' evaluation of initial observation, clinical history, physical examination, and investigation results. Two low information probability models illustrate the representation invariance principle, translation behavior of prior, degree of imprecision, and prior-data conflict using our developed R-package, iDDX. The discussion ends with an elicitation strategy for this inference and a comparison to the priors commonly used in Bayesian inference.

La prise de décision relative au diagnostique clinique par les médecins requiert des aptitudes de raisonnement cognitif et une connaissance spécifique aux maladies afin d'informer la collecte et l'analyse de données de patient qui ne seront pas toujours suffisantes ou à portée de main. Quand il y a peu ou pas d'information, modéliser une ignorance complète ou presque complète respecte tout de même le dogme de précision des inférences bayésiennes. Nous avons adopté les probabilités imprécises de Walley (1991) et quantifié l'incertitude associée à un cours de raisonnement clinique diagnostique mesurant l'évaluation des étudiants relative à l'observation initiale, l'historique clinique, l'examen physique et les résultats d'enquête. Deux modèles de probabilité à faible information illustrent le principe d'invariance de représentation, le comportement de traduction de l'a priori, le degré d'imprécision et le conflit des données a priori au moyen des notre bibliothèque R spécialement conçue : iDDX. L'exposé se termine par une stratégie d'élicitation pour cette inférence et une comparaison par rapport à l'a priori employé dans l'inférence bayésienne.

[10:35-10:50]

Peizhi Li (Queen's University) **Yingwei Peng** (Queen's University) **Jianing Zheng** (Dongbei University of Finance and Economics)

An Empirical Comparison between Gradient Boosting Methods and Cox's Proportional Hazards Model for Right-censored Survival Data

Une comparaison empirique entre les méthodes de boosting de gradient et le modèle de risques proportionnels de Cox pour les données de survie censurées à droite

Gradient boosting methods become popular in recent years to analyze right-censored survival data where Cox's PH is the widely used statistical model. However, there are very limited studies on the differences between the two approaches. We compare two boosting methods with Cox's PH model: one is the GBDT and

Les méthodes de boosting de gradient sont devenues populaires récemment pour l'analyse de données de survie censurées à droite alors que le modèle de risques proportionnels (PH) de Cox est l'approche le plus souvent employé. Cela dit, très peu d'études examinent les différences entre ces deux approches. Nous comparons donc deux méthodes de boosting (GBDT et GBMB) avec

the other is GBMB. A simulation study is conducted to investigate the performance of the methods in practice where only the main effects of covariates are included. The results show that the boosting methods outperform Cox's PH model in both the relative and absolute risk estimation in the PH model except when Cox's PH model is fully specified with nonlinear and interaction covariates effects. It indicates that the boosting methods, particularly the GBDT, is a very competitive method for right-censored survival data if complicated covariate effects may exist but are unknown to the investigator. We illustrate an application of the boosting methods with a real data analysis.

le modèle PH de Cox. Nous menons une étude en simulations afin d'examiner la performance des méthodes en pratique lorsqu'on inclut seulement les effets principaux des covariables. Les résultats démontrent que les méthodes de boosting surpassent le modèle PH de Cox dans l'estimation du risque aussi bien relatif qu'absolu, sauf si le modèle PH de Cox est entièrement spécifié avec des effets linéaires et d'interaction des covariables. Cela indique que les méthodes de boosting, tout particulièrement GBDT, sont très performantes pour les données de survie censurée à droite lorsqu'il existe des effets de covariables complexes, mais que le chercheur ne les connaît pas. Nous illustrons une application des méthodes de boosting avec une analyse de données réelles.

[10:50-11:05]

Kelly Ramsay (York University) **Shojaeddin Chenouri** (University of Waterloo)

FKWC Tests for Differences in the Covariance Structure of Functional Data

Des tests FKWC pour reconnaître les différences dans la structure de covariance des données fonctionnelles

We present a new class of robust, nonparametric multisample hypothesis tests for differences in the covariance operator of functional data, called functional Kruskal-Wallis for covariance (FKWC) tests. FKWC tests use a pooled center-outward ordering of the sampled functions to measure differences in covariance structure between samples. The ordering is based on functional data depth, which we show is connected to the covariance operator. We will present some theoretical aspects and simulation study results of the FKWC tests. We also discuss how FKWC tests handle various challenges associated with functional data, such as computation, high-dimensionality, and outliers.

Nous présentons une nouvelle classe de tests d'hypothèse robustes, non paramétriques et pour échantillons multiples dans le but de déceler les différences dans l'opérateur de covariance des données fonctionnelles. Nous les appelons les tests fonctionnels Kruskal-Wallis pour la covariance (FKWC). Les tests FKWC se servent d'un classement regroupé centre-extérieur des fonctions échantillonnées pour mesurer les différences des structures de covariance entre les échantillons. Ce classement est basé sur la profondeur de données fonctionnelles, que l'on démontre comme étant relié à l'opérateur de covariance. Nous présenterons certains aspects théoriques et des résultats d'étude en simulations relative aux tests FKWC. Nous discuterons aussi de quelle manière les tests FKWC peuvent gérer de nombreux défis associés aux données fonctionnelles, comme le calcul, la haute dimension et les valeurs aberrantes.

[11:05-11:20]

Alex Stringer (University of Waterloo) **Blair Bilodeau** (University of Toronto) **Yanbo Tang** (Imperial College London)

Stochastic Convergence Rates and Applications of Adaptive Quadrature in Bayesian Inference

Taux de convergence stochastique et applications de quadrature adaptative dans l'inférence bayésienne

Adaptive numerical quadrature is used to normalize posterior distributions in many Bayesian models. We provide the first stochastic convergence rates for the error incurred when normalizing a posterior distribution using adaptive quadrature under standard regularity conditions. We give novel approximations to moments, marginal densities, and quantiles, and provide convergence rates for several of these summaries. Low- and high-dimensional applications are presented, the latter using adaptive quadrature as one component

La quadrature numérique adaptative est adoptée pour normaliser les distributions a posteriori dans de nombreux modèles bayésiens. Nous procurons les premiers taux de convergence stochastique pour l'erreur qui survient lors de la normalisation d'une distribution a posteriori à l'aide de quadrature adaptative selon des conditions de régularité standard. Nous offrons de nouvelles approximations pour les moments, les densités marginales et les quantiles, puis procurons des taux de convergence pour plusieurs de ces sommaires. Nous présentons des applications de grande et faible dimension. Pour celles de grande dimension, on adopte une qua-

of a more sophisticated approximation framework, for which limited theory is given. Extension of the theory and computational framework to mixed models for cluster-correlated data, and implications (theoretical and computational) for the ubiquitous Laplace approximation and Laplace approximate marginal likelihood are discussed.

drature adaptative en guise de composante d'un cadre d'approximation sophistiqué, à partir duquel on obtient une théorie limitée. Nous aborderons l'extension de la théorie et le cadre informatique à des modèles mixtes pour des données corrélées par grappe, et les répercussions (théoriques et informatiques) sur l'omniprésente approximation de Laplace et la vraisemblance marginale approximative de Laplace.

[11:20-11:35]

James H. McVittie (University of Regina)

Nonparametric Survival Function Estimation using Current Lifetimes: An alternative to the Grenander Estimator

Estimation de fonction de survie non paramétrique à l'aide de durées de vie actuelles : une solution de remplacement à l'estimateur Grenander

A classical problem in survival analysis is to estimate the failure time distribution from the current lifetimes measured from the date of the initiating event to the date when the cohort was cross-sectionally sampled in a prevalent cohort study design. When the underlying initiating event dates are generated from a stationary Poisson point process, it is possible to estimate the survival function nonparametrically using the Grenander estimator of the current lifetime decreasing density function. As the survival function estimator depends on the Grenander estimator evaluated at time 0, the resulting survival function estimator is inconsistent. We propose an alternative nonparametric maximum likelihood estimation procedure to directly estimate the survival function and show the resulting estimator is both uniformly consistent and weakly convergent. We provide some simulation results and estimate the distribution of the length of NBA player careers.

Un problème classique en analyse de survie est d'estimer la distribution de temps d'échec en fonction des durées de vie actuelles mesurées à partir de la date de l'événement déclencheur lorsque la cohorte a été échantillonnée de façon transversale dans le cadre d'étude de cohorte prévalente. Lorsque les dates d'événement déclencheur sous-jacent sont générées à partir d'un processus ponctuel stationnaire de Poisson, il est possible d'estimer la fonction de survie de façon non paramétrique au moyen de l'estimateur de Grenander de la fonction de densité de durée de vie en baisse actuelle. Puisque l'estimateur de fonction de survie dépend de l'estimateur de Grenander évalué au temps 0, l'estimateur de fonction de survie produit devient incohérent. C'est pourquoi nous proposons une procédure d'estimation de vraisemblance maximum non paramétrique dans le but d'estimer directement la fonction de survie et démontrons que l'estimateur produit est doté d'une cohérence uniforme et d'une convergence faible. Nous procurons quelques résultats de simulation et estimons la distribution de la durée de carrière de joueurs de la NBA.

[11:35-11:50]

Gurbakhshash Singh (Central Connecticut State University) **Gordon Fick** (University of Calgary (Emeritus))

Considerations for the use of a logit or a log link in generalized linear models

Réflexions sur l'utilisation d'une fonction logit ou log dans les modèles linéaires généralisés

For binary outcomes, Logistic regression is the most widely used model based method. The binary outcome is modeled to a set of explanatory variables using the logit link. This model permits the estimation of log odds and functions of log odds. The log-binomial model, using the log link, permits the estimation of log probabilities and functions of log probabilities. We will provide scenarios where models based on the logit link and models based on the log link provide similar inferences and interpretations making the choice of the link rather arbitrary. We then discuss scenarios where the choice of link is crucial to the resulting inferences and interpre-

Pour des résultats binaires, la régression logistique est la méthode basée sur un modèle la plus largement utilisée. Les résultats binaires sont modélisés à un ensemble de variables explicatives utilisant le logit. Ce modèle permet l'estimation des cotes log et des fonctions des cotes log. Le modèle log-binomial utilisant la fonction log permet l'estimation des log-probabilités et des fonctions des log-probabilités log. Nous présentons des scénarios dans lesquels les modèles basés sur le logit et les modèles basés sur le log fournissent des inférences et interprétations similaires, ce qui rend le choix du lien plutôt arbitraire. Nous discutons ensuite des scénarios dans lesquels le choix d'un lien est essentiel aux inférences et interprétations qui en résultent. Nous fournis-

tations. We then provide a similar study with ordinal outcomes, where we will be comparing the proportional odds model with the proportional probability model.

sons aussi une étude semblable avec des résultats ordinaux dans laquelle nous comparons le modèle de cotes proportionnelles au modèle de probabilité proportionnelle.

Chair/Président: Md. Erfanul Hoque

Room/Salle: TB 202

Date: Tuesday May 30 / mardi 30 mai

Time/Heure: 10:20-11:50

Abstract/Résumé

[10:20-10:35]

Yixiu Liu (University of Manitoba) **Depeng Jiang** (University of Manitoba)

Joint Modeling of Longitudinal and Survival Data: An Approach via Poisson Models for the Survival Part under Bayesian Framework

Modélisation conjointe de données de survie et longitudinales : une approche par modèles de Poisson pour la survie dans un cadre bayésien

Data of repeated measurements (longitudinal) and time-to-events (survival) are commonly recorded in a longitudinal study. The joint model (JM) for longitudinal and survival data has been extensively discussed recently to analyze this type of data. Poisson model is an alternative approach to Cox proportional hazard model for survival data, however, it has not been discussed if the Poisson approach can be used in JMs. In this study, we propose to use Poisson model as the survival part in a JM under Bayesian framework. We conducted comprehensive simulation studies to assess the performance of the proposed method, and evaluate the effect of time intervals being split using the Poisson approach, censoring rate, and sample size. The performance of the proposed method was also compared to a published R package for JMs – JMbayes. In addition, data from the Manitoba Follow-Up Study were used to illustrate the merits of our proposed methods.

Dans une étude longitudinale, on enregistre couramment des données de mesures répétées (longitudinales) et de durée de vie (survie). Le modèle conjoint (JM) pour les données longitudinales et de survie a été largement abordé récemment pour l'analyse de ce type de données. Le modèle de Poisson est une option qui diverge du modèle de risque proportionnel de Cox pour les données de survie. Cependant, on ne sait toujours pas si l'approche de Poisson peut servir dans des JM. Dans le cadre de cette étude, nous proposons de nous servir du modèle de Poisson pour la partie survie d'un JM selon un cadre bayésien. Nous menons des études approfondies en simulations afin de juger la performance de la méthode proposée et évaluer l'effet de la division des intervalles temporels au moyen de l'approche de Poisson, du taux de censure et de la taille d'échantillon. La performance de la méthode proposée est aussi comparée à une bibliothèque R publiée pour les JM (JMbayes). De plus, nous avons utilisé des données provenant de l'étude Manitoba Follow-Up Study pour illustrer les qualités de notre méthode.

[10:35-10:50]

Aida Eslami (Université Laval) **Hervé Abdi** (School of Behavioral and Brain Sciences, The University of Texas at Dallas, USA)

Cluster analysis of qualitative and mixed variables

Regroupement de variables qualitatives et mixtes

Cluster analysis of variables is of paramount interest in multivariate data analysis because it can be used to uncover the underlying structure of the data, detect redundancies, and reduce the number of variables. To cluster variables a standard strategy uses (di)similarity coefficients computed between the variables. A more appealing strategy combines cluster analysis of variables with

Le regroupement de variables est d'un intérêt majeur dans l'analyse des données multivariées car elle peut être utilisée pour révéler la structure sous-jacente des données, détecter les redondances et réduire le nombre de variables. La stratégie standard afin de regrouper des variables utilise des coefficients de (dis)similarités calculées entre les variables. Une stratégie plus attrayante consiste à combiner le regroupement de variables et la

data reduction. The aim of our research is to develop a new clustering approach that could be particularly relevant for the cluster analysis of qualitative and quantitative variables based on a multi-block approach—an under studied domain. Our approach bears some similarities to CLUSTATIS (a method of clustering a collection of datasets measured on the same individuals) originally developed to cluster blocks of variables.

réduction de la dimension des données. L'objectif de notre recherche est de développer une nouvelle approche, basée sur une approche multi-blocs, qui serait particulièrement utile pour le regroupement de variables qualitatives et quantitatives–un domaine peu étudié. Notre approche présente certaines similitudes avec CLUSTATIS (une méthode de regroupement appliquée à une collection d'ensembles de données mesurées sur les mêmes individus) développée pour regrouper des blocs de variables.

[10:50-11:05]

Xiaoke Qin (Carleton University)

Clustering Human Gut Microbiome Data Using Finite Mixtures of Generalized Dirichlet-Multinomial Distributions

Regroupement de données sur le microbiome intestinal humain à l'aide de mélanges finis de distributions multinomiales de Dirichlet généralisées

Varying composition of human gut microbiome has been associated with health conditions and pathogenesis of diseases. Because of this, the clustering of microbiome data has been of increasing interest. In this presentation, we propose to use the finite mixture of generalized Dirichlet-multinomial models (GDM), which allows for a flexible correlation matrix. Furthermore, we discuss the non-permutation invariance of GDM. A generalized expectation-maximization algorithm is used for parameter estimation. Using a series of simulated datasets, the performance of the model is evaluated in terms of underlying cluster structure recovery and parameter estimation. We also applied the model to two real-life human microbiome datasets.

Une variation de la composition du microbiome intestinal humain a été associée à divers troubles de santé et à la pathogenèse de maladies. Le regroupement des données du microbiome suscite donc un intérêt croissant. Dans cette présentation, nous proposons d'utiliser le mélange fini de modèles multinomiaux de Dirichlet généralisés (GDM), qui permet une matrice de corrélation flexible. En outre, nous discutons de la non-invariance à la permutation des GDM. Nous employons un algorithme d'espérance-maximisation généralisé pour l'estimation des paramètres. À l'aide d'une série de jeux de données simulées, nous évaluons les performances du modèle en termes de récupération de la structure sous-jacente des grappes et d'estimation des paramètres. Nous appliquons également le modèle à deux ensembles de données de microbiomes humains réels.

[11:05-11:20]

François A Marshall (Queen's University)

Efficient Reconstruction of Low-frequency Oscillations in Long Cyclostationary Processes: Novel Coherence Diagnostics

Reconstruction efficace des oscillations à basse fréquence dans des processus quasi-périodique : nouveaux diagnostics de cohérence

One often assumes reconstructing the low-frequency oscillation of a J-variate stochastic process: the j'th vector entry belongs to an almost-periodic process; its zero-mean noise term is stationary; and its j'th and k'th processes are independent. Send the J processes through a common lowpass filter of high stopband attenuation. Additionally permitting nonstationarity, a lowpass filter suitable for a J=1 process is also suitable for a J=2 process – all univariate processes involved having been drawn from a common finite-dimensional distribution. Two novel diagnostic metrics are here introduced akin to the cross spectrum for the J=2 process when its second process is replaced by a reconstruction of its first process (the first metric a regression coefficient, and the

On s'intéresse souvent à reconstruire l'oscillation basse fréquence d'un processus stochastique à J variables : la j ème entrée du vecteur appartient à un processus quasi-périodique ; son terme de bruit de moyenne nulle est stationnaire ; et ses j ème et k ème processus sont indépendants. Envoyer les J processus à travers un filtre passe-bas commun d'atténuation élevée de la bande atténuée. En laissant tomber l'hypothèse de stationnarité, un filtre passe-bas adapté à un processus à J = 1 convient également à un processus à J = 2 si tous les processus univariés impliqués ont été tirés d'une distribution commune de dimension finie. Deux nouveaux indicateurs sont ici introduits, apparentées au spectre croisé pour le processus à J=2 lorsque son deuxième processus est remplacé par une reconstruction de son premier processus (le premier indicateur est un coefficient de régression et le second est invariant aux

second invariant to noise linear effects). For the stratum times series of a complex clinical human-seizure survey, one finds that the lowpass filter should more strongly reduce ripple height in the nonstationary case.

effets linéaires du bruit). Pour la série temporelle de strates d'une enquête complexe clinique sur les crises d'épilepsie, on constate que le filtre passe-bas devrait réduire plus fortement la hauteur d'ondulation dans le cas non stationnaire.

Recent Advances in Compositional Data Analysis and Applications
Avancées récentes dans l'analyse et les applications des données compositionnelles

Chair/Président: Connie Stewart

Organizer/Responsable: Connie Stewart, Michail Tsagris

Room/Salle: TB 208

Date: Tuesday May 30 / mardi 30 mai

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-14:00]

Michail Tsagris (University of Crete) **Abdulaziz Alenazi** (Northern Border University) **Connie Stewart** (University of New Brunswick)

A Jensen-Shannon divergence based k-NN algorithm for missing value imputation in compositional data

Algorithme k-NN basé sur la divergence de Jensen-Shannon pour l'imputation des valeurs manquantes dans les données de composition

A novel non-parametric method to impute missing values in compositional data is suggested. The method is based on the k-NN algorithm and utilises the Jensen-Shannon divergence and is devised solely for missing values and not rounded zeros. An advantage of the proposed method is its non-parametric nature which unlike restrictive parametric methods, makes no assumption about the structure of the data. Further, this method is applicable even when compositional data contain zero values. Simulation studies using artificial and real data exhibit the superior performance of the proposed algorithm over two competing algorithms at various levels, not only in terms of accuracy but also in terms of computational efficiency.

Nous proposons une nouvelle méthode non paramétrique pour imputer les valeurs manquantes dans les données de composition. La méthode est basée sur l'algorithme k-NN et utilise la divergence de Jensen-Shannon. Elle est conçue uniquement pour les valeurs manquantes et non pour les zéros arrondis. L'un des avantages de la méthode proposée est sa nature non paramétrique qui, contrairement aux méthodes paramétriques restrictives, ne fait aucune hypothèse sur la structure des données. En outre, cette méthode est applicable même lorsque les données de composition contiennent des valeurs nulles. Des études de simulation utilisant des données artificielles et réelles montrent la supériorité de l'algorithme proposé par rapport à deux algorithmes concurrents à différents niveaux, non seulement en termes de précision mais aussi en termes d'efficacité de calcul.

[14:00-14:30]

Janice Lea Scealy (Australian National University)

Score matching for microbiome compositional data

Appariement de score pour les données de composition du microbiome

Compositional data are challenging to analyse due to the non-negativity and sum-to-one constraints on the sample space. It is often the case with microbiome compositional data that many of the components are highly right-skewed, with large numbers of zeros. A major limitation of currently available estimators for compositional models is that they either cannot handle many zeros in the data or are not computationally feasible in moderate to high dimensions. We derive a new set of novel score matching estimators applicable to distributions on a Rie-

Les données de composition sont difficiles à analyser en raison des contraintes de non-négativité et de somme de un sur l'espace d'échantillonnage. Dans le cas des données de composition du microbiome, il arrive souvent que de nombreuses composantes soient fortement asymétriques à droite, avec un grand nombre de zéros. L'une des principales limites des estimateurs actuels pour les modèles de compositions est soit qu'ils ne peuvent pas traiter plusieurs zéros dans les données ou que le calcul informatique devient infaisable dans des dimensions grandes ou modérées. Nous dérivons donc un nouvel ensemble d'estimateurs d'appariement de

mannian manifold with boundary, of which the standard simplex is a special case. The score matching method is applied to estimate the parameters in a new flexible model for compositional data and we show that the estimators are scalable and available in closed form. We apply the new model and estimators to real microbiome compositional data and show that the model provides a good fit to the data.

score applicables aux distributions dans une variété riemannienne avec limites, et dont le simplex standard est un cas spécial. La méthode d'appariement de score est appliquée afin d'estimer les paramètres dans un nouveau modèle flexible pour des données de composition et nous démontrons que les estimateurs sont extensibles et peuvent être écrits sous forme analytique. Nous appliquons le nouveau modèle et les estimateurs à des données de composition réelles sur le microbiome et démontrons que le modèle s'ajuste bien aux données.

[14:30-15:00]

Hongzhe Lee (University of Pennsylvania)

A Tensor Decomposition Model for Longitudinal Microbiome Studies

Modèle de décomposition tensorielle pour études longitudinales du microbiome

Longitudinal microbiome studies can help delineate true biological signals from the high inter-individual variability that is common in microbiome data. However, there are few methods available for unsupervised dimension reduction of time course microbial abundance observations. Existing methods do not fully observe the distribution characteristics of such data types, namely, zero-inflation, compositionality, and overdispersion. We present a tensor decomposition model and a semi-parametric quasi-likelihood estimation method for the decomposition of longitudinal microbiome data, by generalizing existing approaches in tensor decomposition of Gaussian data. Optimization is performed through projected gradient descent additionally allowing interpretability constraints. We show through simulation studies our method is able to recover low rank structures from microbiome time course data, better than existing approaches. Lastly, we apply our method to two existing longitudinal microbiome studies, to detect global microbial changes associated with dietary and pharmaceutical effects, as well as infant birth modes.

Les études longitudinales du microbiome peuvent aider à délimiter les véritables signaux biologiques de la grande variabilité interindividuelle qui est courante dans les données microbiomiques. Cependant, il existe peu de méthodes disponibles pour la réduction non supervisée de la dimension des observations de l'abondance microbienne au cours du temps. Les méthodes existantes ne tiennent pas pleinement compte des caractéristiques de distribution de ces types de données, à savoir l'excès de zéros, la compositionnalité et la surdispersion. Nous présentons un modèle de décomposition tensorielle et une méthode semi-paramétrique d'estimation de la quasi-vraisemblance pour la décomposition des données microbiomiques longitudinales, en généralisant les approches existantes de la décomposition tensorielle des données gaussiennes. L'optimisation est réalisée par la descente du gradient projeté en tenant compte des contraintes d'interprétabilité. Nous montrons à travers des études de simulation que notre méthode est capable de récupérer des structures de faible rang à partir de données microbiomiques de parcours temporel, et ce mieux que les approches existantes. Enfin, nous appliquons notre méthode à deux études longitudinales du microbiome existantes, afin de détecter les changements microbiens globaux associés aux effets alimentaires et pharmaceutiques, ainsi qu'aux modes de naissance des nourrissons.

Statistical Modelling and Computational Intelligence for High-Dimensional Biological Data
Modélisation statistique et intelligence computationnelle pour les données biologiques de haute dimension

Chair/Président: Aerambamoorthy A Thavaneswaran

Organizer/Responsable: You Liang

Room/Salle: AT 301

Date: Tuesday May 30 / mardi 30 mai

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-14:00]

Xuekui Zhang (University of Victoria)

Statistical methods for cell type annotation using single-cell genomic data

Méthodes statistiques pour l'annotation de type de cellule au moyen de données génomiques de cellule unique

Single-cell RNA-sequencing (scRNA-seq) technology enables researchers to investigate a genome at the cellular level with unprecedented resolution. An organism consists of a heterogeneous collection of cell types, each of which plays a distinct role in various biological processes. Hence, the first step of scRNA-seq data analysis is often to distinguish cell types so they can be investigated separately. In this talk, we will discuss our recent contributions for cell type annotation.

La technologie de séquençage de l'ARN à cellule unique (scRNA-seq) permet aux chercheurs d'étudier un génome au niveau cellulaire avec une résolution sans précédent. Un organisme consiste en une collection hétérogène de types de cellules, dont chacune joue un rôle dans de nombreux processus biologiques. Conséquemment, la première étape d'une analyse de données de scRNA-seq est de distinguer ces types de cellules afin de les étudier séparément. Lors de cet exposé, nous parlerons de nos récentes contributions dans l'annotation de type de cellule.

[14:00-14:30]

Li Xing (University of Saskatchewan)

grouppred: A novel machine learning approach for gene module identification and prediction via a co-expression network of single-cell sequencing data

grouppred : une nouvelle approche d'apprentissage automatique pour l'identification et la prédiction de module génétique par l'entremise d'un réseau de co-expression de données de séquençage à cellule unique

Gene co-expression network analysis is widely used in microarray and RNA sequencing data analysis. It groups genes based on their co-expression network. And genes within a group infer similarity in function or coregulation in the pathway. In literature, the approaches to group genes are mainly unsupervised, which may introduce instability and variation across different datasets. Inspired by ensemble learning, we propose a novel approach that ensemble supervised and unsupervised learning techniques and simultaneously works on two tasks, gene module identification and phenotype prediction, during the data analysis process. The identified gene modules from this approach could suggest more candidate genes to the original pathway, and those genes are potential biomarkers for pathway-related dis-

L'analyse de réseaux de co-expression génétique est couramment utilisée dans des microréseaux et des analyses de données de séquençage de l'ARN. Elle regroupe les gènes selon leur réseau de co-expression. Les gènes à l'intérieur d'un groupe présentent des similitudes dans la fonction ou la corégulation dans la voie. Dans la littérature, les approches pour grouper les gènes sont essentiellement non supervisées, ce qui peut générer de l'instabilité et des variations entre les différents jeux de données. En nous inspirant de l'apprentissage ensembliste, nous proposons une nouvelle approche qui rassemble les techniques d'apprentissage supervisées et non supervisées puis travaille simultanément sur deux tâches (l'identification de module génétique et la prédiction de phénotype) durant le processus d'analyse de données. Les modules génétiques identifiés à partir de cette approche pourraient suggérer davantage de gènes candidats pour la voie originale, et

eases. In addition, the novel approach also improves the prediction accuracy for phenotypes. The algorithm can be used as a general prediction algorithm. And, as it is specially designed to handle large samples, it is suitable for handling single-cell data with many cells. We showcased the use of the algorithm in single-cell cell-type auto-annotation.

ces gènes sont des biomarqueurs potentiels pour les maladies liées aux voies. De plus, cette nouvelle approche augmente aussi la précision de prédiction des phénotypes. L'algorithme peut servir en guise d'algorithme de prédiction général. En outre, puisqu'il est spécialement conçu pour gérer de grands échantillons, il peut traiter des données de cellule unique ayant beaucoup de cellules. Nous avons présenté l'utilisation de l'algorithme dans l'annotation automatique du type de cellule unique.

[14:30-15:00]

You Liang (Toronto Metropolitan University)

Hyperspectral microscopy-based semi-automatic segmentation of eye tissues

Segmentation semi-automatique basée sur une microscopie hyperspectrale pour les tissus oculaires

Fluorescence microscopic imaging is widely used for pathological diagnosis of diseases in tissues and biomedical research purposes. Hyperspectral imagers with increased spectral and spatial resolution have the capacity to provide greater structural and molecular information. Algorithm-based analysis platforms capable of analyzing large biomedical hyperspectral datasets are unmet needs and have the potential to extract useful spectral-spatial information from complex tissues. We present an open-source data analysis approach to exploit the potential of hyperspectral fluorescence imaging and to extract unbiased and useful spectral-spatial information from the eye. We demonstrate distinctly different autofluorescence spectra for individual eye tissue types. Furthermore, the systematic segmentation method is used to classify tissue types based on their divergent autofluorescence spectra.

L'imagerie microscopique par fluorescence est largement utilisée pour le diagnostic pathologique des maladies dans les tissus et à des fins de recherche biomédicale. Les imageurs hyperspectraux dotés d'une résolution spectrale et spatiale accrue ont la capacité de fournir davantage d'informations structurelles et moléculaires. Les plateformes d'analyse basées sur des algorithmes et capables d'analyser de grands ensembles de données hyperspectrales biomédicales sont des besoins non satisfaits et ont le potentiel d'extraire des informations spectrales et spatiales utiles à partir de tissus complexes. Nous présentons une approche d'analyse de données ouverte pour exploiter le potentiel de l'imagerie hyperspectrale de fluorescence et pour extraire des informations spectro-spatiales utiles et non biaisées de l'œil. Nous démontrons que les spectres d'autofluorescence sont nettement différents pour les différents types de tissus oculaires. En outre, la méthode de segmentation systématique est utilisée pour classer les types de tissus sur la base de leurs spectres d'autofluorescence divergents.

Recent Developments by Early Career Researchers in Bayesian Modelling for Correlated Data
Développements récents par les chercheurs en début de carrière en modélisation bayésienne pour les données corrélées

Chair/Président: Osvaldo Espin-Garcia

Organizer/Responsable: Osvaldo Espin-Garcia

Room/Salle: AT 101

Date: Tuesday May 30 / mardi 30 mai

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-14:00]

Kuan Liu (University of Toronto)
Bayesian causal analysis with clustered data
Analyse causale bayésienne à l'aide de données en grappes

Constructing causal inference methods to handle longitudinal data in observational studies is of high interest. In an observational setting, treatment assignment at each clinical visit follows a patient-adaptive decision strategy, where the clinician tailors treatment to the individual patient based on current and past clinical measurements as well as treatment histories. When patients are clustered by site, dependence between patients within the same cluster adds on to the data complexity for causal analysis. In line with the work by Wu et al. (2022), we propose an extension to the semi-parametric Bayesian marginal structural models for clustered data. Causal estimation of the average treatment effect is achieved by maximizing a posterior predictive utility function of the marginal treatment effect with importance sampling and the bag of little Bayesian bootstraps. We examine the performance of the proposed method via a simulation study and illustrate it using real-world clinical data.

L'élaboration de méthodes d'inférence causale pour traiter les données longitudinales dans les études d'observation est un sujet très intéressant. Dans un contexte d'observation, l'attribution du traitement à chaque visite clinique suit une stratégie de décision adaptée au patient : le clinicien adapte le traitement à chaque patient sur la base des mesures cliniques actuelles et passées, ainsi que des antécédents de traitement. Lorsque les patients sont regroupés par lieu, la dépendance entre les patients d'un même groupe ajoute à la complexité des données pour l'analyse causale. Dans la lignée des travaux de Wu et coll. (2022), nous proposons une extension des modèles structurels marginaux bayésiens semi-paramétriques pour les données en grappes. L'estimation causale de l'effet moyen du traitement est réalisée par la maximisation d'une fonction utilitaire prédictive a posteriori de l'effet marginal du traitement avec l'échantillonnage d'importance et le petit nombre de bootstraps bayésiens. Nous examinons l'efficacité de la méthode proposée à l'aide d'une étude de simulation et l'illustrons avec des données cliniques réelles.

[14:00-14:30]

Zihang Lu (Queen's University)
A Bayesian latent class model for clustering multiple longitudinal features
Modèle bayésien de classes latentes pour le regroupement de plusieurs variable longitudinales

Clustering longitudinal features is a common goal in clinical and epidemiological studies to identify distinct disease developmental trajectories. Compared to clustering a single longitudinal feature, integrating multiple longitudinal features allows additional information to be incorporated into the clustering process, which may reveal co-existing longitudinal patterns and gener-

Le regroupement des variables longitudinales est un objectif commun aux études cliniques et épidémiologiques afin d'identifier des trajectoires distinctes dans le développement des maladies. Par rapport au regroupement d'une seule variable longitudinale, l'intégration de plusieurs variables longitudinales permet d'intégrer des données supplémentaires au processus de regroupement, ce qui peut révéler des schémas longitudinaux coexistants

ate deeper biological insight. This talk will discuss a Bayesian latent class model and an efficient computation algorithm for clustering multiple longitudinal features with complex data structures. Results from analyzing real-life and simulated data will be presented and discussed.

et d'approfondir les connaissances biologiques. Cette présentation portera sur un modèle bayésien de classes latentes et un algorithme de calcul efficace pour le regroupement de plusieurs variables longitudinales avec des structures de données complexes. Nous présenterons et examinerons les résultats de l'analyse de données réelles et simulées.

[14:30-15:00]

Aya A. Mitani (University of Toronto) **Chen Chen** (University of Toronto) **Dipankar Bandyopadhyay** (Virginia Commonwealth University)

Fitting bivariate skew-t linear mixed models for longitudinal periodontal data: A Stan recourse

Ajustement de modèles mixtes linéaires bivariés à distribution skew-t pour des données parodontales longitudinales : une procédure Stan

To monitor the prognosis of periodontal disease, dentists measure probing pocket depth (PPD) and clinical attachment loss (CAL) on multiple sites on each tooth. PPD and CAL are positively correlated, clustered within a person, and have skewed distributions with heavy tails. Using a linear mixed effects model (LMM) framework, we can model clustered data with bivariate non-normal responses assuming the random effects and errors follow skew-t distributions. Existing work have used WinBUGS or OpenBUGS to implement the method. We extended this bivariate LMM with skew-t distribution to account for site-level and subject-level means using Stan. Stan is more efficient with MCMC sampling for complex models with large variance-covariance matrix and does not require conjugacy for multivariate priors. We illustrate our approach using data from the San Juan Overweight Adults Longitudinal Study.

Pour surveiller le pronostic de maladie parodontale, les dentistes mesurent la profondeur de la poche parodontale (PPD) et la perte d'attachement clinique (CAL) à de multiples sites sur chaque dent. La PPD et la CAL sont corrélées positivement, regroupées dans une personne et ont des distributions asymétriques à queues lourdes. En utilisant le cadre d'un modèle linéaire à effets mixtes (LMM), nous pouvons modéliser des données regroupées avec réponses non normales bivariées supposant que les effets et erreurs aléatoires suivent des distributions skew-t. Des études existantes ont utilisé WinBUGS ou OpenBUGS pour implémenter la méthode. Nous avons étendu le LMM bivarié avec distribution skew-t pour prendre en compte les moyennes au niveau du site et au niveau du sujet à l'aide de la procédure Stan. Stan est plus efficace avec un échantillonnage Monte-Carlo par chaînes de Markov (MCMC) pour les modèles complexes avec de grandes matrices de variance-covariance et ne requiert pas de loi conjuguée pour des a priori multivariés. Des données de l'étude San Juan Overweight Adults Longitudinal Study illustrent notre approche.

Chair/Président: Natalia Stepanova

Organizer/Responsable: Natalia Stepanova

Room/Salle: TB 238

Date: Tuesday May 30 / mardi 30 mai

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-14:00]

Mohamedou Ould Haye (Carleton University) **Anne Philippe** (Nantes University, France) **Caroline Robet** (Nantes University, France)

Inference for continuous-time long memory randomly sampled processes

Inference for continuous-time long memory randomly sampled processes

From a continuous-time long memory stochastic process, a discrete-time randomly sampled one is drawn using a renewal sampling process. We establish the existence of the spectral density of the sampled process, and we give its expression in terms of that of the initial process. We also investigate different aspects of the statistical inference on the sampled process. In particular, we obtain asymptotic results for the periodogram, the local Whittle estimator of the memory parameter and the long run variance of partial sums. We mainly focus on Gaussian continuous-time process. The challenge being that the randomly sampled process will no longer be jointly Gaussian. This is based on a joint paper with Anne Philippe and Caroline Robet from Nantes University, France.

À partir d'un processus stochastique à longue mémoire en temps continu, nous extrayons un processus échantillonné aléatoirement en temps discret en utilisant un processus d'échantillonnage de renouvellement. Nous établissons l'existence de la densité spectrale du processus échantillonné et nous donnons son expression en termes de celle du processus initial. Nous étudions également différents aspects de l'inférence statistique sur le processus échantillonné. En particulier, nous obtenons des résultats asymptotiques pour le périodogramme, l'estimateur local de Whittle du paramètre de mémoire et la variance à long terme des sommes partielles. Nous nous concentrons principalement sur les processus gaussiens en temps continu, le défi étant que le processus échantillonné aléatoirement ne sera plus conjointement gaussien. Ceci est basé sur un article commun avec Anne Philippe et Caroline Robet de l'Université de Nantes, France.

[14:00-14:30]

Andrei Volodin (University of Regina) **Rita Giuliano** (University of Pisa, Italy) **Manuel Ordóñez Cabrera** (University of Sevilla, Spain)

On the sub-Gaussianity of the r-correlograms

De la sous-gaussianité des r-corrélogrammes

The topic of this presentation concerns the classical problem of establishing properties of estimators for the correlation function of a stationary Gaussian process. Stationary Gaussian processes are the main objects and bases for a large variety of probability and statistical models. It is a well-known fact that a stationary Gaussian process is characterized by its autocorrelation function. This is the reason why an estimation of the autocorrelation function from the observations of some real-

Le sujet de cette présentation concerne le problème classique de l'établissement des propriétés des estimateurs de la fonction de corrélation d'un processus gaussien stationnaire. Les processus gaussiens stationnaires sont les principaux objets et bases d'une grande variété de modèles probabilistes et statistiques. Il est bien connu qu'un processus gaussien stationnaire se caractérise par sa fonction d'autocorrélation. C'est la raison pour laquelle l'estimation de la fonction d'autocorrélation à partir des observations d'une certaine réalisation du processus gaussien stationnaire joue

ization of the stationary Gaussian process plays a crucial role in the construction of an appropriate model for data. We show that the centered relay correlogram function is a sub-Gaussian random variable. This is done by a careful analysis of its Laplace transform and by estimating of the sub-Gaussian standard of the r-correlograms. This is based on a joint paper with Rita Giuliano, University of Pisa, Italy, and Manuel Ordóñez Cabrera, University of Sevilla, Spain.

un rôle crucial dans la construction d'un modèle approprié pour les données. Nous montrons que la fonction de corrélogramme de relais centré est une variable aléatoire sous-gaussienne. Pour ce faire, nous procédons à une analyse minutieuse de sa transformée de Laplace et à une estimation de la norme sous-gaussienne des r-corrélogrammes. Ceci est basé sur un article commun avec Rita Giuliano, Université de Pise, Italie, et Manuel Ordóñez Cabrera, Université de Séville, Espagne.

[14:30-15:00]

Marie Turcicova (Institute of Atmospheric Physics CAS) **Natalia Stepanova** (Carleton University)

Sparse signal recovery in regression models with Gaussian white noise

Récupération de signaux "sparses" dans les modèles de régression avec bruit blanc gaussien

We consider the problem of selecting active components of a sparse multivariate function f in the d-dimensional Gaussian white noise model. It is assumed that f belongs to a class of smooth functions in $L_2([0,1]^d)$ and that it is made up of functions of a smaller number of variables k. The integer k is either fixed or k tends to infinity with d at a slower rate than d. We establish conditions under which exact recovery of f is possible, and provide a procedure that achieves exact recovery. Our procedure is the best possible (in the asymptotically minimax sense) for selecting the active k-variate components of f. Moreover, it is adaptive to the level of sparsity of f. Our results are shown to augment previous work in this area. This is joint work with Natalia Stepanova from Carleton University, Canada.

Nous étudions le problème de la sélection des composants actifs d'une fonction f multivariée "sparse" dans un modèle de bruit blanc gaussien à d dimensions. Nous supposons que f appartient à une classe de fonctions lisses dans $L_2([0,1]^d)$ et qu'elle est composée de fonctions d'un plus petit nombre de variables k. Nous supposons que l'entier k est fixe ou qu'il tend vers l'infini avec d à un rythme plus lent que d. Nous établissons les conditions sous lesquelles la récupération exacte de f est possible, et fournissons une procédure qui permet d'obtenir une récupération exacte. Notre procédure est la meilleure possible (asymptotiquement minimax) pour sélectionner les composants actifs à k-variables de f. De plus, elle s'adapte au niveau de "sparsité"de f. Nos résultats complètent les travaux antérieurs dans ce domaine. Ces travaux sont réalisés en collaboration avec Natalia Stepanova de l'Université de Carleton, au Canada.

Statistics in transportation
Statistique dans les transports

Chair/Président: Pratheepa Jeganathan

Organizer/Responsable: Aurelie Labbe

Room/Salle: TB 210

Date: Tuesday May 30 / mardi 30 mai

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-14:00]

Lijun Sun (McGill University)

Bayesian Complementary Kernelized Learning for Multidimensional Spatiotemporal Data

Apprentissage noyauté bayésien complémentaire pour les données spatiotemporelles multidimensionnelles

Developing effective and computationally efficient statistical models to accommodate nonstationary/nonseparable processes containing both long-range and short-scale variations becomes a challenging task, especially for large-scale datasets with various corruption/missing structures. In this talk, we will introduce Bayesian Complementary Kernelized Learning (BCKL) to achieve scalable probabilistic modeling for multidimensional spatiotemporal data. BCKL integrates kernelized low-rank factorization with short-range spatiotemporal Gaussian processes (GP), in which the two components complement each other. We use a multi-linear low-rank factorization component to capture the global/long-range correlations in the data and introduce an additive short-scale GP based on compactly supported kernel functions to characterize the remaining local variabilities. Our results confirm the superior performance of BCKL in providing accurate posterior mean and high-quality uncertainty estimates.

Il est difficile de concevoir des modèles statistiques fonctionnels et efficaces sur le plan calculatoire pour s'adapter aux processus non stationnaires et non séparables contenant à la fois des variations longue distance et d'autres à courte échelle, tout particulièrement pour les ensembles de données à grande échelle comportant plusieurs structures manquantes ou corrompues. Lors de cet exposé, nous présenterons l'apprentissage noyauté bayésien complémentaire (BCKL) afin de réaliser une modélisation probabiliste extensible pour les données spatiotemporelles multidimensionnelles. BCKL intègre la factorisation à rang réduit noyauté avec les processus gaussiens (GP) spatiotemporels à courte échelle, dans lesquels les deux composantes se complémentent les unes les autres. Nous nous servons d'un composant de factorisation à rang réduit multilinéaire dans le but de saisir les corrélations globales et de longue distance dans les données et y insérer un GP additif de courte échelle basé sur des fonctions par noyaux soutenus de façon compacte afin de caractériser les variabilités locales restantes. Nos résultats confirment la performance supérieure du BCKL pour procurer une moyenne a posteriori fiable et des estimations de haute qualité sur l'incertitude.

[14:00-14:30]

Alexandra Schmidt (McGill University) **Hassan Rezaee** (HEC, Montreal) **Joshua Stipancic** (Intact Insurance) **Aurélie Labbe** (HEC, Montreal)

A process convolution model for crash count data on a network

Modèle de convolution de processus pour les données de dénombrement des accidents sur un réseau

Crash data observed on a road network often exhibit spatial correlation due to unobserved effects with inherent spatial correlation following the structure of the road network. We introduce a network process convolution (NPC) model, wherein the spatial correlation

Les données d'accidents observées sur un réseau routier présentent souvent une corrélation spatiale en raison des effets non observés et une corrélation spatiale inhérente à la structure du réseau routier. Nous introduisons un modèle de convolution de processus de réseau, dans lequel la corrélation spatiale entre les données d'ac-

among crash data is captured through a Gaussian Process (GP) approximated through a kernel convolution approach. The GP's covariance function is based on path distance computed between a limited set of knots and crash data points on the road network. The proposed model offers a straightforward approach for predicting crash frequency at unobserved locations where covariates are available, and for interpolating the GP values anywhere on the network. Inference procedure is performed following the Bayesian paradigm and is implemented in R-INLA, which offers an estimation procedure that is more efficient than a Markov Chain Monte Carlo approach. We fitted our model to synthetic data and to crash data from Ottawa, Canada.

cidents est capturée par un processus gaussien approximé par une approche de convolution de noyau. La fonction de covariance du processus gaussien est basée sur la distance de trajet calculée entre un ensemble limité de nœuds et de points correspondant à des accidents sur le réseau routier. Le modèle proposé offre une approche simple pour prédire la fréquence des accidents aux endroits non observés où les covariables sont présentes, et pour interpoler les valeurs du processus gaussien n'importe où sur le réseau. La procédure d'inférence est réalisée selon le paradigme bayésien et mise en œuvre dans R-INLA, qui offre une procédure d'estimation plus efficace qu'une approche de Monte-Carlo par chaînes de Markov. Nous avons adapté notre modèle à des données synthétiques et à des données sur les accidents survenus à Ottawa, au Canada.

[14:30-15:00]

Aurélie Labbe (HEC, Montreal) **Mengying Lei** (McGill University) **Lijun Sun** (McGill University)

BKTR - Bayesian Kernelized Tensor Regression: application to bike-sharing demand modeling

BKTR - Bayesian Kernelized Tensor Regression : application à la modélisation de la demande de vélos en libre service

As a regression technique in spatial statistics, the spatiotemporally varying coefficient model (STVC) is an important tool for discovering nonstationary and interpretable response-covariate associations over both space and time. However, it is difficult to apply STVC for large-scale spatiotemporal analyses due to its high computational cost. To address this challenge, we summarize the spatiotemporally varying coefficients using a third-order tensor structure and propose to reformulate the spatiotemporally varying coefficient model as a special low-rank tensor regression problem. The low-rank decomposition can effectively model the global patterns of the large data sets with a substantially reduced number of parameters. To further incorporate the local spatiotemporal dependencies, we use Gaussian process (GP) priors on the spatial and temporal factor matrices. We refer to the overall framework as Bayesian Kernelized Tensor Regression (BKTR). For model inference, we develop an efficient Markov chain Monte Carlo (MCMC) algorithm, which uses Gibbs sampling to update factor matrices and slice sampling to update kernel hyperparameters. We conduct extensive experiments on both synthetic and real-world data sets, and our results confirm the superior performance and efficiency of BKTR for model estimation and parameter inference.

En tant que technique de régression spatiale, le modèle à coefficients variables spatio-temporel (STVC) est un outil important permettant de détecter des associations entre une variable réponse et des covariables interprétables dans l'espace et dans le temps. Cependant, il est difficile d'appliquer le modèle STVC pour des analyses spatio-temporelles à grande échelle en raison de son coût de calcul élevé. Dans cette présentation, nous réduisons la dimension des coefficients spatio-temporels à l'aide d'une structure tensorielle de troisième ordre et proposons de reformuler le modèle comme un cas particulier de régression tensorielle de faible rang. La décomposition de faible rang permet de modéliser efficacement de grands ensembles de données avec un nombre considérablement réduit de paramètres. Afin de prendre en compte les dépendances spatio-temporelles locales, nous utilisons des lois apriori de processus gaussien (GP) sur les matrices de facteurs spatiaux et temporels. Nous référons au modèle développé sous le nom de Bayesian Kernelized Tensor Regression (BKTR). Pour l'inférence de modèle, nous développons un algorithme de chaîne de Markov Monte Carlo (MCMC) efficace, qui utilise l'échantillonnage de Gibbs pour mettre à jour les matrices factorielles et l'échantillonnage par tranches pour mettre à jour les hyperparamètres du noyau. La méthode proposée est évaluée sur un ensemble de données synthétiques et réelles sur la demande de vélos en libre-service, et nos résultats confirment les performances et l'efficacité supérieures de BKTR pour l'estimation de modèles et l'inférence de paramètres.

Chair/Président: Fangda Liu

Organizer/Responsable: Fangda Liu

Room/Salle: RB 1200

Date: Tuesday May 30 / mardi 30 mai

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-14:00]

Silvana Manuela Pesenti (University of Toronto) **Sebastian Jaimungal** (University of Toronto) **Yuri Saporito** (Getulio Vargas Foundation (FGV)) **Rodrigo Targino** (Getulio Vargas Foundation (FGV))

Risk Budgeting Allocation for Dynamic Risk Measures

Allocation de la budgétisation du risque pour des mesures de risque dynamiques

We develop an approach for risk budgeting allocation – a risk diversification portfolio strategy – where risk is measured using time-consistent dynamic risk measures. For this, we introduce a notion of dynamic risk contributions that generalise the classical Euler contributions and which allow us to obtain dynamic risk contributions in a recursive manner. Moreover, we show how the risk allocation problem may be recast as a optimisation problem and develop an actor-critic approach to solve for risk allocations using deep learning techniques.

Nous développons une approche de budgétisation du risque – une stratégie de diversification des risques dans un portefeuille – dans laquelle le risque est calculé à l'aide de mesures de risque dynamiques cohérentes dans le temps. À cette fin, nous présentons une notion de contributions de facteurs de risque dynamiques qui généralisent les contributions classiques d'Euler et permettent d'obtenir des contributions de facteurs de risque dynamiques de façon récursive. De plus, nous montrons comment le problème d'allocation des risques peut être reformulé comme problème d'optimisation et nous développons une approche acteur-critique pour résoudre les allocations des risques en utilisant des techniques d'apprentissage profond.

[14:00-14:30]

Wenjun Jiang (University of Calgary) **Tim Boonen** (University of Amsterdam)

Distributionally Robust Insurance with the Wasserstein Distance

Assurance à distribution robuste avec la distance de Wasserstein

This paper studies the optimal insurance problem within the risk minimization framework and from a decision maker (DM)'s perspective. We assume that the DM is uncertain about the underlying distribution of her loss and would consider all the distributions that closely surround a given (benchmark) distribution, where the "closeness" is measured by the Wasserstein distance. Under the expected-value premium principle, the DM picks the indemnity function that minimizes its risk exposure under the worst-case loss distribution. By assuming that the DM's preferences are given by a convex distortion risk measure, we disentangle the structures of the optimal indemnity function and worst-case loss distri-

Cette étude aborde le problème d'assurance optimale dans le cadre de minimisation de risque et selon la perspective d'un décideur. Nous supposons que le décideur n'est pas certain de la distribution sous-jacente de sa perte et il tiendrait compte de toutes les distributions étroitement liées à une distribution donnée (référence), où ≪l'étroitesse≫ est mesurée par la distance de Wasserstein. Selon le principe de la prime de valeur attendue, le décideur choisit la fonction d'indemnité qui minimise son exposition au risque selon la distribution de la pire perte. En supposant que les préférences du décideur sont données par une mesure du risque de distorsion convexe, nous délions de façon analytique les structures de la fonction d'indemnité optimale et de la distribution de la pire perte, puis donnons les formes explicites pour chacune d'elles selon les me-

bution in an analytical way, and give the explicit forms for both of them under specific distortion risk measures. We also compare the results under the first-order and second-order Wasserstein distance. Some numerical examples are given at the end to show more implications of our main results.

sures du risque de distorsion spécifique. Nous comparons aussi les résultats selon la distance de Wasserstein de premier ordre et de deuxième ordre. Nous présenterons à la fin certains exemples numériques démontrant d'autres réflexions relatives à nos résultats principaux.

[14:30-15:00]

Haiyan Liu (Michigan State University)

Worst-Case Risk with Unspecified Risk Preferences

Risque le plus défavorable avec préférences de risque non spécifiées

We study the worst-case distortion risk measure when information about distortion functions is partially available. The explicit forms of the worst-case distortion functions from several different sets of plausible distortion functions are obtained. When there is no concavity constraint on distortion functions, the worst-case distortion function is independent of the risk to be measured and the corresponding worst-case distortion risk measure is a combination of several VaR's at different confidence levels. When the concavity constraint is imposed on distortion functions and the set of concave distortion functions is defined by the riskiness of one single risk, the explicit form of the worst-case distortion function is obtained. When the set of concave distortion functions is defined by the riskiness of multiple risks, we reduce the infinite-dimensional optimization problem to a finite-dimensional optimization problem which can be solved numerically.

Nous étudions la mesure du risque de distorsion le plus défavorable lorsque les informations sur les fonctions de distorsion ne sont que partiellement disponibles. Nous obtenons les formes explicites des fonctions de distorsion les plus défavorables à partir de plusieurs ensembles de fonctions de distorsion plausibles. Lorsqu'il n'y a pas de contrainte de concavité sur les fonctions de distorsion, la fonction de distorsion la plus défavorable est indépendante du risque à mesurer et la mesure du risque de distorsion la plus défavorable correspondante est une combinaison de plusieurs VaR à différents niveaux de confiance. Lorsque la contrainte de concavité est imposée aux fonctions de distorsion et que l'ensemble des fonctions de distorsion concaves est défini par le degré de risque d'un seul risque, on obtient la forme explicite de la fonction de distorsion la plus défavorable. Lorsque l'ensemble des fonctions de distorsion concaves est défini par le degré de risque de plusieurs risques, nous réduisons le problème d'optimisation à dimension infinie à un problème d'optimisation à dimension finie qui peut être résolu numériquement.

Statistical Inference by Combining Information from Multiple Sources
Inférence statistique par combinaison d'informations provenant de sources multiples

Chair/Président: Wen Wilson Lu

Organizer/Responsable: Changbao Wu

Room/Salle: AT 102

Date: Tuesday May 30 / mardi 30 mai

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-14:00]

Peisong Han (University of Michigan)

Causal inference with improved efficiency by integrating external summary information in the possible presence of heterogeneous study populations

Inférence causale avec efficacité accrue en intégrant une information sommaire externe en présence possible de populations d'étude hétérogènes

Consider causal inference based on a small sample that meets the no-unmeasured-confounders assumption. Although causal effects can be estimated without bias, the small sample size limits the estimation efficiency. Suppose that, for the same response variable, there have been large-scale external observational studies with fewer covariates collected. These studies either studied the association between the response and the corresponding covariates using regression models, or attempted to estimate the causal effect, in which case the causal effect is not correctly estimated due to the violation of the no-unmeasured-confounders assumption. Results from these studies are available as summary statistics such as regression parameter estimates or causal effect estimates. We develop a method to integrate the external summary information into causal effect estimation for the purpose of efficiency improvement. The method can deal with population heterogeneity between the internal and external study populations.

Prenons une inférence causale basée sur un petit échantillon qui respecte l'hypothèse d'absence de confondants non mesurés. Alors que les effets causaux peuvent être estimés sans biais, la petite taille d'échantillon restreint l'efficacité d'estimation. Supposons que, pour une même variable de réponse, il existe des études observationnelles externes de grande envergure avec moins de covariables recueillies. Ces études ont soit étudié l'association entre la réponse et les covariables correspondantes au moyen de modèles de régression, ou elles ont tenté d'estimer l'effet causal, sauf qu'il est mal estimé en raison du non-respect de l'hypothèse d'absence de confondants non mesurés. Les résultats de ces études sont offerts sous forme de statistiques sommaires comme des estimations de paramètre de régression ou des estimations d'effet causal. Nous développons une méthode pour intégrer l'information sommaire externe dans l'estimation d'effet causal dans le but d'améliorer son efficacité. La méthode peut gérer l'hétérogénéité entre les populations d'études interne et externes.

[14:00-14:30]

Jae-Kwang Kim (Iowa State University) **Zhonglei Wang** (Xiamen University)

Repreducing kernel method for data integration

Méthode à noyau reproduisant pour l'intégration des données

We consider a new class of methods using reproducing kernel for data integration. The proposed method will present a unified approach to combining information from several sources and developing data integration in general. Optimal transport theory can be used

Nous étudions une nouvelle classe de méthodes faisant appel à un noyau reproduisant pour l'intégration des données. La méthode proposée présente une approche unifiée pour combiner de l'information de plusieurs sources et développer l'intégration des données en général. La théorie du transport optimal peut être uti-

to compute the reproducing kernel weights. Some statistical properties will be also investigated. Numerical results will also be presented.

lisée pour calculer les poids du noyau reproduisant. Nous discutons aussi de certaines propriétés statistiques et des résultats numériques sont présentés.

[14:30-15:00]

Changbao Wu (University of Waterloo)

Issues with Analysis of Non-Probability Survey Samples

Problèmes liés à l'analyse d'échantillons d'enquête non probabiliste

We discuss three fundamental issues in dealing with non-probability survey samples: (i) the unknown participation mechanism; (ii) the unknown population represented by the sample; and (iii) the search for external data sources on auxiliary information for the target population. We also provide brief reviews on recent methodological advances and some guidance on collecting data using non-probability surveys.

Nous abordons trois problèmes fondamentaux dans le traitement d'échantillons d'enquête non probabiliste : (i) le mécanisme inconnu de participation ; (ii) la population inconnue représentée par l'échantillon ; et (iii) la recherche de sources externes de données en matière d"information auxiliaire sur la population ciblée. Nous donnerons aussi un bref aperçu de récentes avancées méthodologiques et certains conseils sur la collecte de données utilisant des enquêtes non probabilistes.

Chair/Président: Tolu Sajobi

Organizer/Responsable: Tolu Sajobi

Room/Salle: RB 2200

Date: Tuesday May 30 / mardi 30 mai

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-15:00]

Lehana Thabane (McMaster University)

My journey as a biostatistician: "EBM=Statistical thinking in health"

Mon parcours de biostatisticien : « Médecine fondée sur des données probantes = pensée statistique dans le domaine de la santé »

Trained as a statistician in three continents, I will be narrating the story of my journey from Lesotho in Southern Africa to England, and then to Canada in the 1990s. I will share my experiences working as a biostatistician in collaborative research in health spanning across several clinical disciplines with the sole aim of advancing evidence-based care through the development and application of statistical methods. I will end the talk with some key ingredients for success in collaborative research.

Formé au métier de statisticien sur trois continents, je raconterai l'histoire de mon voyage du Lesotho, en Afrique du Sud, à l'Angleterre, puis au Canada dans les années 1990. Je présenterai mes expériences en tant que biostatisticien dans le cadre d'une recherche collaborative sur la santé couvrant plusieurs disciplines cliniques, dans le seul but de faire progresser les soins fondés sur des données probantes grâce à l'élaboration et à l'application de méthodes statistiques. Je conclurai ma présentation en expliquant quelques éléments clés de la réussite d'une recherche collaborative.

Room/Salle: RB 2220-2224

Date: Tuesday May 30 / mardi 30 mai

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-15:00]

Haoxuan Ge (University of Toronto at Mississauga) **Jue Wang** (University of Toronto)

Spatial Non-Stationarity Effects of Unhealthy Food Environments and Green Spaces for Type-2 Diabetes in Toronto

Effets de non-stationnarité spatiale des environnements alimentaires malsains et des espaces verts pour le diabète de type 2 à Toronto

Environmental factors may operate differently when relations are measured across different geographical locations, a phenomenon known as spatial non-stationarity (SNS), thus leading to biased conclusions and often ignored in past studies. This study investigates the SNS effect of unhealthy food environments and green spaces on the diabetes prevalence rate (DPR) between age groups in Toronto's neighborhoods. Geographically weighted regression models are utilized to explore the impacts of SNS. The results from this study reveal that environmental variables dissimilarly affect diabetes prevalence rates among age groups and neighborhoods in Toronto after controlling for socioeconomic factors. For example, green space density yields positive associations with DPR for elder generations but negative relationships for younger age groups around Toronto East. The observed associations will provide recommendations to support authorities in designing targeted intervention programs to control DPR.

Les facteurs environnementaux peuvent agir différemment lorsque les relations sont mesurées dans différents lieux géographiques. Ce phénomène, appelé « non-stationnarité spatiale », conduit ainsi à des conclusions biaisées et n'a souvent pas été pris en compte dans les études antérieures. Cette étude analyse l'effet des environnements alimentaires malsains et des espaces verts sur le taux de prévalence du diabète entre des groupes d'âge dans les quartiers de Toronto. Des modèles de régression pondérés géographiquement sont utilisés pour explorer les répercussions de la non-stationnarité spatiale. Les résultats de cette étude révèlent que les variables environnementales ont une incidence différente sur les taux de prévalence du diabète selon les groupes d'âge et les quartiers de Toronto, une fois les facteurs socio-économiques pris en compte. Par exemple, la densité des espaces verts présente des associations positives avec le taux de prévalence du diabète pour les générations âgées, mais des relations négatives pour les groupes d'âge plus jeunes dans la région de Toronto Est. Les associations observées permettront de formuler des recommandations pour aider les autorités à concevoir des programmes d'intervention ciblés pour contrôler le taux de prévalence du diabète.

[13:30-15:00]

Dana Ivancevic (University of Ottawa) **The Minh Luong** (Bureau of Food Surveillance and Science Integration, Food Directorate, Health Products and Food Branch, Health Canada) **Lidia Loukine** (Bureau of Food Surveillance and Science Integration, Food Directorate, Health Products and Food Branch, Health Canada) **Dominique Ibañez** (Bureau of Food Surveillance and Science Integration, Food Directorate, Health Products and Food Branch, Health Canada) **Ali Karimnezhad** (Bureau of Food Surveillance and Science Integration, Food Directorate, Health Products and Food Branch, Health Canada; University of Ottawa) **Kelly Burkett** (University of Ottawa)

Is There an Association Between Dietary Sodium Intake and Life Expectancy in the Canadian Adult Population?

Existe-t-il une association entre l'apport en sodium alimentaire et l'espérance de vie dans la population adulte canadienne ?

In this study, we investigate the association between sodium intake and mortality in the Canadian adult population and estimate life expectancy in people consum-

Dans cette étude, nous étudions l'association entre l'apport en sodium et la mortalité dans la population adulte canadienne. Nous estimons l'espérance de vie des personnes consommant du sodium

ing sodium below and above 2300 mg/day, known as the chronic disease risk reduction intake. We hypothesize that adults with sodium intake below the 2300 mg/day threshold have lower mortality rate and higher life expectancy than those with sodium intake above the threshold. We use the Canadian Community Health Survey Nutrition data linked to the Canadian Vital Statistics database using a probabilistic linkage performed by Statistics Canada. We estimate the usual dietary intake of sodium from 24-hour diet recalls using a method developed by the National Cancer Institute, known as the NCI method. We present our results investigating the association between usual sodium intake and mortality risk using a complementary log-log regression model with age and sex as confounders. We use the life table method to estimate life expectancy.

au-dessous et au-dessus de 2 300 mg par jour, ce qui correspond à l'apport permettant de réduire le risque de maladies chroniques. Nous émettons l'hypothèse que les adultes dont l'apport en sodium est inférieur au seuil de 2 300 mg par jour ont un taux de mortalité plus faible et une espérance de vie plus élevée que ceux dont l'apport en sodium est supérieur à ce seuil. Nous utilisons les données nutritionnelles de l'Enquête sur la santé dans les collectivités canadiennes couplées à la Base canadienne de données sur l'état civil au moyen d'un couplage probabiliste effectué par Statistique Canada. Nous estimons l'apport alimentaire habituel en sodium à partir de rappels alimentaires de 24 heures au moyen d'une méthode conçue par l'Institut national du cancer, connue sous le nom de méthode de l'Institut national du cancer. Nous présentons nos résultats concernant l'association entre l'apport habituel en sodium et le risque de mortalité à l'aide d'un modèle de régression log-log complémentaire avec l'âge et le sexe comme facteurs de confusion. Nous utilisons la méthode des tables de mortalité pour estimer l'espérance de vie.

[13:30-15:00]

Yan Yuan (University of Alberta) **Yutong Han** (University of Alberta)

How a Misclassified Binary Outcome Y in Training Data Affects Model Prediction Performance: a Simulation Study

Une étude en simulations relative aux effets d'un résultat binaire Y classé incorrectement dans des données d'apprentissage sur la performance de prédiction de modèle

Statistics models are used for explaining and/or predicting an outcome of interest. For explanations, the focus is on parameter estimation that describes an independent variable's effect. In this regard, the effect of a misclassified outcome variable and how to correct it has been studied extensively, with one popular method being MCSIMEX. However, a relevant question yet to be addressed is how misclassification affects predictive performance. We investigate this question through extensive simulation studies. Motivated by a real world example, we generated a binary event status Y that is subject to misclassification. We fit a logistic regression model using the misclassified Y* and assessed model performance on a test data simulated from the same underlying model without misclassification. We show that the predictive performance on test data is similar regardless of whether or not the misclassified Y* was corrected and always better than the performance on the training data.

Les modèles statistiques servent à expliquer et prédire un résultat pertinent. En guise d'explications, on se concentre sur l'estimation de paramètres qui décrivent l'effet d'une variable indépendante. Dans ce contexte, des études approfondies se sont penchées sur l'effet d'une variable de résultat incorrectement classée et comment la corriger, et une méthode bien connue à ce propos est MCSIMEX. Cependant, il reste une question pertinente à aborder : comment la mauvaise classification affecte-t-elle la performance prédictive ? Nous enquêtons sur la question par l'entremise d'études approfondies en simulations. Nous avons généré un statut d'événement binaire Y soumit à une classification incorrecte, le tout justifié par un exemple réel. Nous ajustons un modèle de régression logistique au moyen de Y incorrectement classé et de la performance de modèle évalué sur des données de test simulées à partir du même modèle sous-jacent sans erreurs de classification. Nous démontrons que la performance prédictive sur les données de test est semblable, peu importe si le Y incorrectement classé est corrigé ou non, et qu'elle surpasse à tout coup la performance des données d'apprentissage.

[13:30-15:00]

Xiao Wu (University of Toronto) **Wai Yu Amanda Ng** (University of Toronto)

Statistical Methods of Variant Calling in Next-Generation Sequence Data Analysis

Méthodes statistiques d'appel de variants dans l'analyse des données de séquençage de nouvelle génération

Next generation sequencing (NGS) is a rapidly evolving technology that has revolutionized the field of genomics. It allows the rapid and cost-effective sequencing of a large amount of DNA simultaneously, making it a powerful tool for a wide range of applications, from genome-wide association studies to the discovery of disease biomarkers. However, the analysis of massive amounts of data generated by NGS remains a major challenge and there is a need for continued development of tools to optimize data processing and analysis. Variant calling is an important process of DNA sequencing analysis, which identifies variants present in the sequence data. Recently, many variant callers have been developed based on different statistical principles, such as binary and count-outcome models, as well as Bayesian approaches. In this poster, we will review the statistical methods employed by different variant callers and discuss their applications in the study of human diseases.

Le séquençage de nouvelle génération est une technologie qui évolue rapidement et qui a révolutionné le domaine de la génomique. Il permet le séquençage rapide et rentable d'une grande quantité d'ADN simultanément, ce qui en fait un outil puissant pour de nombreuses applications, des études d'association à l'échelle du génome à la découverte de biomarqueurs de maladies. Cependant, l'analyse des quantités massives de données générées par le séquençage de nouvelle génération reste un défi majeur et il est nécessaire de poursuivre la mise au point d'outils pour optimiser le traitement et l'analyse des données. L'appel de variants est un processus important de l'analyse du séquençage de l'ADN, qui permet d'identifier les variants présents dans les données de séquence. Récemment, de nombreux outils d'appel de variants ont été conçus sur la base de différents principes statistiques, tels que les modèles à variable de réponse binaire et de comptes, ainsi que les approches bayésiennes. Nous passerons en revue les méthodes statistiques employées par différents appels de variants et discuterons de leurs applications dans l'étude des maladies humaines.

[13:30-15:00]

Yiting Chen (Simon Fraser University) **Xiaoping Shi** (University of British Columbia)

Multivariate Two-sample Test Statistics Based on Data Depth

Variables à tester pour deux échantillons multivariés basés sur la profondeur de données.

Data depth has been applied as a nonparametric measurement for ranking multivariate samples. In this talk, we focus on homogeneity tests to assess whether two multivariate samples are from the same distribution. There are many data depth-based tests for this problem, but they may not be very powerful, or have unknown asymptotic distributions, or have slow convergence rates to asymptotic distributions. Given the recent development of data depth as an important measure in quality assurance, we propose two new test statistics for multivariate two-sample homogeneity tests. The proposed test statistics have simple asymptotic half-normal distribution and chi-squared distribution with degree of freedom one. We also discuss the generalization of the proposed tests to multiple samples. The simulation study demonstrates the superior performance of the proposed tests. The test procedure is illustrated by two real data examples.

La profondeur de données a été appliquée en guise de mesure non paramétrique pour classer les échantillons multivariés. Lors de cet exposé, nous nous concentrons sur des tests d'homogénéité afin d'évaluer si deux échantillons multivariés proviennent de la même distribution. Il existe de nombreux tests basés sur la profondeur de données pour résoudre ce problème, mais ils peuvent présenter plusieurs inconvénients : faible puissance, distribution asymptotique inconnue ou taux de convergence lent relatif aux distributions asymptotiques. Compte tenu des récents développements relatifs à la profondeur de données en guise de mesure importante pour l'assurance qualité, nous proposons deux nouvelles variables à tester pour des tests d'homogénéité à deux échantillons multivariés. Les variables à tester proposées possèdent une distribution demi-normale asymptotique simple et une distribution chi carré avec un degré de liberté. Nous abordons aussi la généralisation des tests proposés pour de multiples échantillons. L'étude en simulations démontre la performance supérieure des tests proposés. La procédure de test est illustrée par deux exemples basés sur des données réelles.

[13:30-15:00]

Quynh Phuong Vu (University of Toronto)

The Conundrum of COVID-19 Self-Testing Devices

L'énigme des autotests COVID-19

The unprecedented emergence of COVID-19 upended our lives to a great extent, resulting in pressing health crises and economic fallouts on a global scale. Even after considerable effort had gone into vaccination to lift COVID-19 restrictions on our regular physical activities, many questions about COVID-19 remain unresolved, one of which is whether COVID-19 self-testing devices are science-backed and suitable for customs screening purposes. The project will provide a data perspective on this scientific problem of interest. The data set used in the project is sourced from the Harvard Dataverse and has information on the price, accuracy, time, distribution channel, and test types of COVID test kits used by 583 participants. We will conduct qualitative and quantitative analyses on the data set, starting with exploratory data analysis alongside appropriate visualizations and then developing an apt model for hypothesis testing theories, causal explanation, and prediction.

L'émergence sans précédent de la COVID-19 a bouleversé nos vies dans une large mesure, entraînant des crises sanitaires urgentes et des répercussions économiques à l'échelle mondiale. Même si des efforts considérables ont été déployés en matière de vaccination pour lever les restrictions imposées par la COVID-19 sur nos activités physiques régulières, de nombreuses questions concernant cette maladie restent en suspens, notamment celle de savoir si les dispositifs d'autotest de la COVID-19 sont scientifiquement fondés et s'ils conviennent à des fins de contrôle douanier. Le projet fournira une perspective de données sur ce problème scientifique d'intérêt. L'ensemble des données utilisées dans le projet provient du Harvard Dataverse et contient des données sur le prix, la précision, la durée, le canal de distribution et les types de tests des trousses de test COVID utilisées par 583 participants. Nous effectuerons des analyses qualitatives et quantitatives sur l'ensemble des données, en commençant par une analyse exploratoire des données accompagnée de visualisations appropriées, puis en concevant un modèle approprié pour tester des hypothèses, des théories, des explications causales et des prédictions.

[13:30-15:00]

Karen A. Kopciuk (Cancer Care Alberta, AHS)

Data and Knowledge: Considerations for Working with Data from Indigenous People

Données et savoir : les précautions à adopter en travaillant avec des données se rapportant aux peuples autochtones

Working with data from Indigenous people requires additional sensitivity and considerations, over and above the usual ethical ones. Formal data governance agreements between Indigenous groups and researchers outline the responsibilities and expectations, as well as processes to be followed that have implications for statisticians working with their data. When collaborating with First Nations (FN) groups, the principles of ownership, control, access, and possession (OCAP®) need to be adopted across all aspects of a study. How these principles were implemented by a research team in a recent co-led project will be described. Strategies for success included training in OCAP® for the statisticians as well as sharing of results in FN community gatherings to understand their meaning. Learning about Indigenous ways of knowing can expand statisticians' perspectives on knowledge from data.

Travailler avec des données portant sur les peuples autochtones exige une sensibilité et des précautions supplémentaires, au-delà des pratiques éthiques habituelles. Les accords formels de gouvernance des données entre les groupes autochtones et les chercheurs soulignent les responsabilités, les attentes et les processus à respecter qui impliquent les statisticiens se servant de leurs données. En collaborant avec les groupes des premières nations (PN), les principes de propriété, de contrôle, d'accès et de possession (PCAP®) doivent être adoptés dans l'ensemble des aspects d'une étude. Nous décrirons comment ces principes ont été mis en oeuvre par une équipe de recherche dans un récent projet coopératif. Les stratégies pour réussir comprennent une formation sur le PCAP® pour les statisticiens ainsi qu'un partage des résultats dans les regroupements de communautés de PN afin de comprendre leur signification. L'apprentissage des mœurs autochtones en matière de savoir peut élargir les perspectives d'un statisticien relatives aux découvertes tirées de données.

[13:30-15:00]

Sévérien Nkurunziza (University of Windsor) **Yunhong Lyu** (University of Windsor)

Inference in Generalized Exponential Ornstein-Uhlenbeck Processes with Change-point

Inférence dans les processus d'Ornstein-Uhlenbeck exponentiels généralisés

In this presentation, we consider inference problem

Dans cette présentation, nous étudions le problème d'inférence sur

about the drift parameter in generalized exponential Ornstein-Uhlenbeck processes with unknown change-point in the context of unknown number of base functions. In particular, we consider the case where the drift parameter may satisfy some restrictions. We propose estimators for the number of base functions and the change-point. We also derive the unrestricted estimator, the restricted estimator and shrinkage estimators as well as their asymptotic properties. Finally, to illustrate the performance of the proposed method, we present some simulation results and analyze some real financial datasets. Beyond these interesting contributions, the novelty of this approach consists in the fact that we overcome the difficulties dues to the randomness fact of the dimensions of the derived estimators.

le paramètre de dérive dans les processus d'Ornstein-Uhlenbeck exponentiels généralisés avec un point de rupture inconnu dans le contexte du nombre de fonctions de base inconnu. Plus précisement, nous considérons le cas où le paramètre de dérive pourrait satisfaire à certaines restrictions. Nous proposons des estimateurs pour le nombre de fonctions de base ainsi que pour le point de rupture. Nous établissons également l'estimateur non restreint, l'estimateur restreint et les estimateurs à rétrécissement ainsi que leurs propriétés asymptotiques. Enfin, pour illustrer la performance de la méthode proposée, nous présentons quelques résultats de simulation et analysons quelques données financières. Au-delà de ces contributions intéressantes, la nouveauté de cette approche réside dans le fait que nous surmontons les difficultés dues au fait aléatoire des dimensions des estimateurs établis.

[13:30-15:00]

Linke Li (University of Toronto: Dalla Lana School of Public Health)

Efficiently Evaluating the Economic Benefits of Group Sequential Design using Bayesian Decision Theory

Évaluation efficace des avantages économiques d'une méthodologie séquentielle par groupes au moyen de la théorie de décision bayésienne

As there is an inherent conflict between the number of viable research projects and the availability of funding, decision-makers must prioritise research questions. One method to prioritise research is to identify the study that would provide the greatest economic benefit. The expected value of sample information (EVSI) is a Bayesian decision metric that can quantify the economic benefit of a study with a pre-determined sample size. However, it is unclear how to adapt EVSI to assess the benefit of a group sequential design (GSD). Thus, we first formally define EVSI for GSD. We then employ machine learning techniques to develop novel EVSI estimation methods to efficiently compute EVSI for GSD. Their performances are compared with the conventional estimation approach, which is accurate but infeasibly complex in practice, using a real-world case study. The results show that the proposed methods can cut the computational time from two days to several seconds without a loss of accuracy.

Puisqu'il y a un conflit inhérent entre le nombre de projets de recherche viables et l'offre de fonds, les preneurs de décisions doivent prioriser les questions de recherche. Une méthode pour prioriser la recherche est de repérer l'étude qui procurerait le plus d'avantages économiques. La valeur attendue de l'information d'échantillon (EVSI) est une unité de décision bayésienne qui peut quantifier l'avantage économique d'une étude avec une taille d'échantillon prédéterminée. Cependant, on ne sait pas comment adapter l'EVSI pour évaluer les avantages d'une méthodologie séquentielle par groupes (GSD). Donc, nous définissons d'abord l'EVSI pour la GSD. Puis nous employons des techniques d'apprentissage automatique pour élaborer de nouvelles méthodes d'estimation de l'EVSI pour efficacement calculer l'EVSI pour la GSD. À partir d'une étude de cas réelle, nous comparons leurs performances avec l'approche d'estimation conventionnelle, qui est précise, mais ridiculement complexe en pratique. Les résultats démontrent que les méthodes proposées peuvent réduire le temps de calcul de deux jours à quelques secondes sans perte de précision.

[13:30-15:00]

Christopher Salim (Brock University) **S. Ejaz Ahmed** (Brock University) **Tianyu Guan** (Brock University)

A Study of Soccer Space Gain in Pass Sequences Using Logistic Regression

Étude du gain d'espace au soccer dans les séquences de passes à l'aide de la régression logistique

Some pass sequences open up more space on the pitch than others. Several studies about space gain in soccer have been performed in the past, but the relationship

Certaines séquences de passes ouvrent plus d'espace sur le terrain que d'autres. Plusieurs études sur le gain d'espace dans le soccer ont été réalisées par le passé, mais la relation entre le gain d'es-

between space gain and the ability to score a goal has not been established yet. This research aims to predict goal occurrence by using total space gain for each pass sequence as the explanatory variable in logistic regression. Combining event and tracking data from the 2019 regular season of Chinese Super League (CSL), space quality can be calculated. Space quality is defined as the product of likelihoods that a team can gain control at a given location and time and the defending team can impede scoring attempts. The research demonstrated that for every unit the total space gain increased in a given pass sequence, the scoring odds increased by 23%. This finding showcases that pass sequences that create space are more likely to help in scoring, which is consistent with real-life soccer events.

pace et la capacité à marquer un but n'a pas encore été établie. Cette recherche vise à prédire l'occurrence d'un but en utilisant le gain d'espace total pour chaque séquence de passes comme variable explicative dans la régression logistique. En combinant les données d'événements et de suivi de la saison régulière 2019 de la Chinese Super League (CSL), la qualité de l'espace peut être calculée. La qualité de l'espace est définie comme le produit des probabilités qu'une équipe puisse prendre le contrôle à un endroit et à un moment donnés et que l'équipe en défense puisse entraver les tentatives de marquage. La recherche démontre que pour chaque unité d'augmentation du gain d'espace total dans une séquence de passes donnée, les chances de marquer augmentent de 23 %. Ce résultat montre que les séquences de passes qui créent de l'espace sont plus susceptibles d'aider à marquer, ce qui correspond aux événements de la vie réelle au soccer.

[13:30-15:00]

Wai Yu Amanda Ng (University of Toronto) **Sofia Panasiuk** (University of Toronto)

A Clustering Approach to Bottom-up Theories of Subjective Well-being: Is Life Satisfaction Differentially Predicted by Levels of Domain Satisfaction

Une approche de regroupement des théories ascendantes du bien-être subjectif : La satisfaction à l'égard de la vie est-elle différemment prédite selon les niveaux des domaines de satisfaction ?

Different individuals, at various stages of their lives, place more importance on one life domain than another. This study generates a set of well-being profiles according to different combinations of aggregated elements of well-being: purpose, community, physical, financial and social well-being. To do this, we used different clustering algorithms to identify hidden clusters of individuals surveyed from 2014–2017 by the Gallup U.S Daily Poll. We tried the following clustering approaches: Gaussian finite mixture model (GFMM) clustering, K-means clustering, and hierarchical clustering. We then compared the number and characteristics of clusters given by each approach. The study can be used to gain more understanding of how changes in well-being vary by hidden clusters in their population. In this presentation, we will discuss each clustering method used and comment on the barriers we face.

Des individus différents, à différentes étapes de leur vie, accordent plus d'importance à un domaine de la vie par rapport à d'autres. Cette étude génère un ensemble de profils de bien-être selon différentes combinaisons d'éléments agrégés de bien-être reliés à : la raison d'être, la communauté, la santé, les finances et le social. Pour ce faire, nous utilisons différents algorithmes de regroupement afin de repérer les grappes cachées d'individus enquêtés de 2014 à 2017 par un sondage de Gallup. Nous avons tenté les approches de regroupement suivantes : regroupement de modèles de mélange fini gaussien, regroupement de K-moyennes et regroupement hiérarchique. Puis nous avons comparé le nombre et les caractéristiques des grappes obtenues par chaque approche. L'étude peut servir à mieux comprendre comment les changements relatifs au bien-être varient par grappes cachées dans leur population. Dans cette présentation, nous discuterons de chaque méthode de regroupement utilisée et commentons les obstacles rencontrés.

[13:30-15:00]

Alexandra Mossman (University of Waterloo)

Dynamic Treatment Regimes for Clustered Data with Between- and Within-Group Interference

Régimes de traitement dynamiques pour données en grappes avec interférence entre les groupes et à l'intérieur des groupes

Dynamic treatment regimes (DTRs) are sequences of decision rules that aim to optimize a patient's course of treatment using available information at each stage of follow-up. Much of the DTR methodology has been

Les régimes de traitement dynamiques (DTR) sont des séquences de règles de décision qui visent à optimiser le traitement d'un patient en utilisant les informations disponibles à chaque étape du suivi. Une grande partie de la méthodologie des DTR a été

developed for use within observational datasets under the stable unit treatment value assumption (SUTVA), which states that an individual's outcome is independent of the decision rules followed by others at each stage of treatment. However, this often does not hold in practice due to different forms of interference that occurs within social networks, such as in infectious disease settings where an individual may be less likely to become infected if their neighbours are fully vaccinated. While much of the biostatistical literature has addressed the issue of partial interference, where interference occurs within groups but not across groups, much work has yet to be done for modifying DTR methodology to account for both within- and between-group interference within clustered and hierarchical data. This work explores a modified version of one such DTR methodology, dynamic weighted ordinary least squares regression (dWOLS), using network weights based on propensity score functions modelling both fixed and random effects within groups of patients to account for individual- and group-specific factors affecting the nature of interference present.

développée pour être utilisée dans des ensembles de données d'observation sous l'hypothèse d'une valeur de traitement unitaire stable (SUTVA), qui stipule que le résultat d'un individu est indépendant des règles de décision suivies par d'autres à chaque étape du traitement. Toutefois, cette hypothèse n'est souvent pas valable dans la pratique en raison des différentes formes d'interférence qui se produisent au sein des réseaux sociaux, comme dans le cas des maladies infectieuses, où un individu peut être moins susceptible d'être infecté si ses voisins sont entièrement vaccinés. Alors qu'une grande partie de la littérature biostatistique a abordé la question de l'interférence partielle, où l'interférence se produit au sein des groupes mais pas entre les groupes, il reste encore beaucoup à faire pour modifier la méthodologie DTR afin de tenir compte de l'interférence au sein des groupes et entre les groupes dans les données en grappes et hiérarchiques. Ce travail explore une version modifiée de l'une de ces méthodologies DTR, la régression dynamique pondérée par les moindres carrés ordinaires (dWOLS), en utilisant des poids de réseau basées sur des fonctions de score de propension modélisant à la fois des effets fixes et aléatoires au sein des groupes de patients pour tenir compte des facteurs spécifiques aux individus et aux groupes qui affectent la nature de l'interférence présente.

[13:30-15:00]

Eman Sreeni Abbas (University of Manitoba) **Brad C. Johnson** (University of Manitoba)

A jackknife type procedure for modelling networks

Procédure de type jackknife pour la modélisation de réseaux

The jackknife resampling method has been successfully applied to many complex inference problems. At its most basic, this method involves systematically removing one observation from the data and estimating parameter(s) in the resulting sample, and then recombining these estimates. In this study, we consider a "jackknife" type procedure to network modelling by repeatedly obtaining parameter estimates using logistic regression while randomly deleting one edge from the graph network. Our goal is to determine how well the predicted networks from the model fit the observed network through goodness of fit measures. This method allows us to also comment on the robustness of this method.

La méthode de rééchantillonnage jackknife a été appliquée avec succès à de nombreux problèmes d'inférence complexe. À la base, cette méthode implique systématiquement de retirer une observation dans les données, d'estimer le ou les paramètre(s) de l'échantillon qui en résulte, puis de recombiner ces estimations. Cette étude présente une procédure de type « jackknife » de modélisation de réseau en obtenant à répétition des estimations de paramètres utilisant une régression logistique tout en supprimant de façon aléatoire une arête du réseau du graphe. Notre objectif est de déterminer à quel point les réseaux prédits à partir du modèle s'ajuste au réseau observé grâce à des mesures de qualité d'ajustement. Cette méthode nous permet également d'en commenter la robustesse.

[13:30-15:00]

Guanyu Chen (The University of British Columbia) **Ching-Lin Shih** (National Sun Yat-Sen University) **Yan Liu** (Carleton University)

Exploring How Response Time predict Student Engagement in English Language Proficiency using Computerized Adaptive Test Data

Étude du temps de réponse comme prédicteur de l'engagement des étudiants à maîtriser la langue anglaise à l'aide de données de test adaptatif informatisé

Computerized adaptive testing (CAT) is widely used, which adapts to the ability of each examinee and greatly shortens testing time. However, few studies have investigated how the response time (RT) recorded in CAT is related to student engagement in exams. Test engagement reflects a student's willingness to actively participate in the testing process. Lack of engagement can undermine test reliability and validity. This study has three goals: handling the missing-by-design issue raised by CAT, estimating student engagement using RT, and exploring how student characteristics affect their engagement. With a sample of 867 participants, preliminary results showed that RT could be used to measure test-taking behaviors via a psychometric method. Bayesian mixed effects models will be used to further explore how student characteristics influence engagement.

Largement utilisé, le test adaptatif informatisé (CAT) s'adapte à la capacité de chaque candidat à l'examen et raccourcit grandement le temps de test. Peu d'études ont cependant porté sur la façon dont le temps de réponse (RT) enregistré pendant un CAT est lié à l'engagement des étudiants pendant un examen. L'engagement d'un étudiant à un test reflète sa volonté de participer activement au processus de test. Un manque d'engagement peut saper la fiabilité et la validité du test. Cette étude a trois objectifs : traiter le problème des données manquantes par conception soulevé par le test adaptatif informatisé, estimer l'engagement des étudiants en utilisant un RT et explorer la façon dont les caractéristiques des étudiants influent sur leur engagement. Avec un échantillonnage de 867 participants, les résultats préliminaires ont montré qu'on pouvait utiliser le RT pour mesurer les comportements de participation aux tests par voie d'une méthode psychométrique. Des modèles bayésiens à effets mixtes sont utilisés pour approfondir l'influence des caractéristiques des étudiants sur leur engagement.

[13:30-15:00]

Raymond Romaniuk (Brock University) **William Marshall** (Brock University) **Mei Ling Huang** (Brock University)

Boosting Methods for Imbalanced Data Classification

Méthodes de boosting pour la classification de données déséquilibrées

Data imbalance is an important consideration when working with real world data. Over/undersampling approaches allow us to gather more insight from the limited data we have on the minority class; however, there are many proposed methods. The goal of our study is to identify the optimal approach for over/undersampling to use with Adaptive Boosting (AdaBoost). Based on a simulation study, we've found that combining AdaBoost with various sampling techniques provides an increased weighted accuracy across classes for progressively larger data imbalances. The three Synthetic Minority Oversampling Technique's (SMOTE) performed the best, with the SMOTE – Edited Nearest Neighbours (SMOTE-ENN) approach being the most accurate for all levels of data imbalance. We then applied the most effective over/undersampling methods to predict upsets (games where the lower seeded team wins) in the March Madness College Basketball Tournament.

Le déséquilibre des données est un facteur qu'il faut prendre en compte lorsqu'on travaille avec des données du monde réel. Les approches de suréchantillonnage et sous-échantillonnage permettent de nous faire une meilleure idée à partir des données limitées recueillies sur la classe minoritaire ; plusieurs méthodes sont cependant proposées. Notre étude a pour but d'identifier l'approche optimale à utiliser pour le suréchantillonnage et sous-échantillonnage avec le boosting adaptatif (AdaBoost). À partir d'une étude de simulation, nous avons trouvé que la combinaison de l'AdaBoost et de diverses techniques d'échantillonnage fournit une exactitude pondérée accrue dans l'ensemble des classes pour des déséquilibres de données progressivement plus grands. La meilleure performance a été celle des trois techniques de suréchantillonnage synthétique minoritaire (SMOTE), l'approche SMOTE – plus proches voisins modifiée (SMOTE-ENN) ayant été la plus précise pour tous les niveaux de déséquilibre des données. Nous avons ensuite appliqué les méthodes de suréchantillonnage et sous-échantillonnage les plus efficaces pour prédire les victoires inattendues (celles où une équipe classée inférieure gagne) au March Madness College Basketball Tournament.

[13:30-15:00]

Kyuson Lim (McMaster University) **Utkarsh J. Dang** (Carleton University) **Sanjeena Dang** (Carleton University)

Mixtures of contaminated multivariate Poisson-lognormal models

Mélanges de modèles multivariés Poisson-lognormale contaminés

Atypical observations or mild outliers are commonly en-

Les données réelles comportent couramment des observations

countered in real data. Modelling these together with typical observations can substantially impact the model fit. However, identifying such atypical observations can be challenging in a clustering framework. Various approaches for identifying atypical observations are available in a clustering framework for continuous data but there is a dearth of approaches for multivariate discrete data. Here, we propose a contaminated mixture model that identifies atypical observations in a clustering framework and mitigates their influence in the model fit. For this, we utilize a mixture of multivariate Poisson-lognormal (MPLN) models that accounts for both correlation among the variables and over-dispersion through a hierarchical structure. Atypical observations are identified in the latent space using a contaminated mixture framework. Parameter estimation is done using a variational variant of the expectation maximization algorithm.

atypiques ou des valeurs aberrantes légères. Modéliser de telles données avec des observations typiques peut avoir un impact marqué sur l'ajustement du modèle. Cependant, l'identification d'observations atypiques peut se révéler difficile dans un cadre de mise en grappes. Diverses approches d'identification des observations atypiques sont disponibles dans un cadre de mise en grappes pour des données continues, mais rares sont les approches pour les données discrètes multivariées. Nous proposons ici un modèle de mélange contaminé qui identifie les observations atypiques dans un cadre de mise en grappes et en atténue l'influence dans l'ajustement du modèle. Pour ce faire, nous utilisons un mélange de modèles multivariés Poisson-lognormale (MPLN) qui prend en compte la corrélation entre les variables et la surdispersion dans une structure hiérarchique. Les observations atypiques sont identifiées dans l'espace latent en utilisant un cadre de mélange contaminé. Une variante variationnelle de l'algorithme espérance-maximisation est utilisée pour l'estimation des paramètres.

[13:30-15:00]

Samopriya Basu (Simon Fraser University (Department of Statistics and Actuarial Science)) **Faezeh Yazdi** (Simon Fraser University)

Bayesian inversion and uncertainty quantification for muon tomography

Inversion bayésienne et quantification des incertitudes pour la tomographie muonique

Muon tomography is an emerging imaging technique using muons contained in natural cosmic rays to determine subsurface density in a diverse range of objects. It is often difficult to get that information from other imaging techniques and muon tomography can be carried out in a cost-effective and environmentally sound manner. The process of inference from this muon tomography data can be framed as an inverse problem, usually with non-unique solutions. In this paper, we carry out statistical inversion in a Bayesian regression framework for the subsurface density contrast based on data from muon detectors, first in the case where there is a single uniform anomaly, and extend it to multiple anomalies. Based on a Gaussian prior model we have a closed form solution for the posterior distribution that can be efficiently implemented. A posterior uncertainty representation has been developed directly from the Bayesian inverse solution which gives an uncertainty quantification for the computed solution. We also compare our Bayesian inversion approach with industry standard regularization schemes. The method is illustrated in a series of synthetic examples and a real-world application.

La tomographie muonique est une technique d'imagerie en plein développement qui utilise les μons contenus dans les rayons cosmiques naturels pour déterminer la densité du sous-sol dans une gamme variée d'objets. Cette information est souvent difficile à obtenir à partir d'autres techniques d'imagerie et la mise en œuvre de la tomographie μonique est économique et respectueuse de l'environnement. Le processus d'inférence à partir des données produites par la tomographie μonique peut être formulé comme un problème inverse, souvent sans solution unique. Dans cet article, nous utilisons l'inversion statistique dans le cadre de la régression bayésienne pour obtenir les contrastes de densité du sous-sol à partir de données de détecteurs de μons, d'abord dans le cas d'une seule anomalie uniforme, puis dans celui de multiples anomalies. En partant d'un modèle a priori gaussien, nous dérivons une solution analytique de la distribution a posteriori, qui peut être implémentée efficacement. Une représentation a posteriori des incertitudes est développée directement à partir de la solution inverse bayésienne, ce qui donne une quantification des incertitudes pour la solution calculée. Nous comparons également notre approche par inversion bayésienne aux schémas de régularisation standard de l'industrie. Notre méthode est illustrée par une série d'exemples synthétiques et une application réelle.

[13:30-15:00]

Simon Snyman (University of British Columbia) **John Braun** (University of British Columbia Okanagan) **Lengyi Han**

(University of British Columbia Okanagan)

On Data Sharpening in Nonparametric Autoregressive Models

Sur l'affinement des données dans les modèles autorégressifs non paramétriques

Data sharpening has been shown to reduce bias in nonparametric regression and density estimation. Its performance on nonlinear first order autoregressive models is studied theoretically and numerically in this paper. Although the asymptotic properties of data sharpen- ing are not as favourable in the presence of serial dependence as in bivariate regression with independent responses, it is still found to reduce bias under mild conditions on the autoregression function. Numerical comparisons with the bias reduction method of Cheng et al. (2018) indicate that data sharpening is competitive in this setting. Keywords: nonparametric regression, bias reduction, kernel, autoregressive time series

Il a été démontré que l'affinement des données réduisait le biais dans la régression non paramétrique et l'estimation de densité. Sa performance sur les modèles autorégressifs non linéaires est étudiée théoriquement et numériquement dans ce document. Bien que les propriétés asymptotiques de l'affinement de données ne soient pas aussi favorables en présence de dépendance sérielle que celle de la régression bivariée avec des réponses indépendantes, on retrouve quand même une réduction du biais sous des conditions faibles sur la fonction d'autorégression. Des comparaisons numériques avec la méthode de réduction du biais de Cheng et al. (2018) indiquent que l'affinement des données est une méthode compétitive dans ce contexte. Mots-clés : régression non paramétrique, réduction du biais, noyau, séries chronologiques autorégressives

[13:30-15:00]

Ying Miao (University of Manitoba) **Po Yang** (University of Manitoba)

Optimal single-stratum and multi-stratum designs for the dual purpose of model discrimination and parameter estimation

Conceptions optimales avec multi-strates et une seule strate visant deux buts : la discrimination du modèle et l'estimation des paramètres

Most optimal design criteria in the literature are model-based criteria that rely on having an assumed model to select an optimal design for the experiment. A specified model might not be feasible in reality. The total entropy-based criterion to address the dual problem of model discrimination and parameter estimation has been in the literature for decades. This criterion has yet to be discussed due to the computation time requested. In this research, we develop the generalized entropy-based criterion for multi-stratum design and suggest some methods to reduce computation time. Examples are provided for selecting the optimal design in single-stratum completely randomized experiments and multi-stratum experiments such as split-plot designs using the new criterion.

La plupart des critères de conception optimale dans la littérature sont des critères basés sur un modèle qui dépend d'un modèle hypothétique pour la sélection d'une conception optimale de l'expérience. Un modèle spécifié peut ne pas être faisable en réalité. Depuis des décennies, la littérature fait état du critère basé sur l'entropie totale en réponse au double problème de discrimination du modèle et de l'estimation des paramètres. Ce critère mérite d'être discuté en raison du temps computationnel qu'il requiert. Notre recherche développe un critère basé sur l'entropie généralisé pour une conception multi-strates et propose quelques méthodes pour réduire le temps computationnel. Des exemples sont fournis pour la sélection à l'aide de ce nouveau critère d'une conception optimale d'expériences à une seule strate entièrement randomisées et d'expériences multi-strates, comme les conceptions en parcelles divisées.

[13:30-15:00]

Brittany Alana Perry (Brock University) **William Marshall** (Brock University) **Mei Ling Huang** (Brock University)

Boosting Methods for Classification with Small Sample Size

Méthodes de boosting pour la classification avec de petites tailles d'échantillon

AdaBoost is an ensemble method that can be used to boost the performance of machine learning algorithms by combining several weak learners to create a single strong learner. The most popular weak learner is a decision stump (low depth decision tree). One limitation of

AdaBoost est une méthode d'ensemble qui peut être utilisée pour améliorer la performance des algorithmes d'apprentissage machine en combinant plusieurs apprenants faibles pour créer un seul apprenant fort. L'apprenant faible le plus populaire est une souche de décision (un arbre décisionnel de faible profondeur). AdaBoost

AdaBoost is its effectiveness when working with small sample sizes. This work explores variants to the AdaBoost algorithm such as Real AdaBoost, Logit Boost, and Gentle AdaBoost. These variants all follow a gradient boosting procedure like AdaBoost, with modifications to the weak learners and weights used. We are specifically interested in the accuracy of these boosting algorithms when used with small sample sizes . As an application, we study the link between functional network connectivity (as measured by EEG recordings) and Schizophrenia by testing whether the proposed methods can classify a participant as Schizophrenic or healthy control based on quantities measured from their EEG recording.

est d'une efficacité limitée avec de petites tailles d'échantillon. Ce travail explore des variants de l'algorithme AdaBoost, comme Real AdaBoost, Logit Boost et Gentle AdaBoost. Ces variants suivent tous une procédure de gradient boosting comme AdaBoost, avec des modifications aux apprenants faibles et poids utilisés. Nous nous intéressons précisément à l'exactitude de ces algorithmes de boosting lorsqu'ils sont utilisés avec de petites tailles d'échantillon. Leur application permet d'étudier le lien entre la connectivité fonctionnelle du réseau (telle que mesurée par EEG [électroencéphalographie]) et la schizophrénie en vérifiant si les méthodes proposées peuvent classifier un participant comme schizophrène ou en contrôle sain, en se basant sur les quantités mesurées à l'aide de son électroencéphalogramme.

[13:30-15:00]

Wanye Gao (University of British Columbia)

From simulation to observation – comparing spatial heterogeneity of radiation at the cellular level

De la simulation à l'observation – comparaison de l'hétérogénéité spatiale de la radiation au niveau cellulaire

Heterogeneity in the response of cancerous tumour tissue to ionizing radiation is well-established in the scientific literature, though not at all well-understood. We take some initial steps towards this understanding by considering cell-size measurements on radiochromic film (RCF) that has been exposed to varying levels of ionizing radiation, and comparing the observed spatial heterogeneity therein to what would be expected via Monte Carlo simulation under similar conditions. We describe the process of distilling high-dimensional measurements obtained through Raman spectroscopy on a grid of RCF to a low dimensional grid representation of dose deposition. Then we obtain spatial heterogeneity scores on this grid, along with the grid of simulated values, and provide arguments correlating the two results. We also provide context on experimental work currently under development that we will apply similar processes to, foreshadowing the potential of significant scientific findings on radiation response at a cellular granularity.

L'hétérogénéité dans la réponse des tissus d'une tumeur cancéreuse au rayonnement ionisant est bien établie dans la littérature scientifique, même si elle est très mal comprise. Nous faisons une première démarche pour acquérir cette compréhension en considérant des mesures de tailles de cellules sur des films radiochromiques (RCF) qui ont été exposés à divers niveaux de rayonnement ionisant et en comparant l'hétérogénéité spatiale qu'on y observe à ce qui serait attendu d'une simulation de Monte-Carlo sous des conditions similaires. Nous décrivons le processus de distillation des mesures de haute dimension obtenues par spectroscopie Raman sur une grille d'évaluation des RCF à une représentation en grille à faible dimension des dépôts de doses. Nous obtenons ensuite des scores d'hétérogénéité spatiale sur cette grille, en même temps qu'une grille de valeurs simulées et fournissons des arguments qui corrèlent les deux résultats. Nous présentons aussi le contexte d'une étude expérimentale en voie de développement à laquelle nous appliquerons des processus similaires, laissant présager un potentiel de résultats scientifiques significatifs quant à la réponse à la radiation sur le plan de la granularité cellulaire.

[13:30-15:00]

Danika M. Lipman (University of Calgary)

A Bayesian Regression model with Gaussian Process for Semi-Continuous Response

Modèle de régression bayésien avec processus gaussien pour une réponse semi-continue

Gaussian processes are kernel-based statistical methods that accommodate non-linear data structures. I will present on a two-part Gaussian process model for semi-continuous data using Bayesian methods for estimation.

Les processus gaussiens sont des méthodes statistiques basés sur un noyau qui s'adaptent aux structures de données non linéaires. Je présente un modèle de processus gaussien en deux parties pour des données semi-continues en utilisant des méthodes bayésiennes

The two parts of the model include a probit model part for the probability of nonzero response and a conditional model for the mean response given that it is nonzero. This two-part model also performs Bayesian variable selection using spike and slab priors to determine key predictors associated with the response in high dimensional data. With kernel-based methods, there is often high computational complexity in inverting high dimensional kernel matrices and performing prediction, so to circumvent this problem I have applied the Nyström method for dimension reduction. A Markov Chain Monte Carlo algorithm is used for estimation of parameters. In my presentation, I will introduce the model and present results from simulated data.

pour l'estimation. Les deux parties du modèle consiste en un modèle probit pour la probabilité de réponse non nulle et un modèle conditionnel pour la réponse moyenne, étant donné qu'elle est non nulle. Ce modèle en deux parties procède aussi à la sélection bayésienne de variables en utilisant des a priori "spike and slab" pour déterminer les prédicteurs clés associés à la réponse dans les données de haute dimension. Avec les méthodes basées sur un noyau, il y a souvent une grande complexité computationnelle à inverser les matrices de noyau et à faire une prédiction ; pour contourner ce problème, j'ai appliqué la méthode Nyström pour la réduction de la dimension. Un algorithme de Monte-Carlo par chaînes de Markov est utilisé pour l'estimation des paramètres. Je présente le modèle et les résultats avec des données simulées.

Chair/Président: Lisa M. Lix

Room/Salle: TB 240

Date: Tuesday May 30 / mardi 30 mai

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-13:45]

Yanqing Yi (Memorial University of Newfoundland)

Optimality and Efficiency of Sequential Designs of Clinical Trials

Optimalité et efficacité des plans séquentiels d'essais cliniques

This talk will discuss the use of the information sequentially collected during a clinical trial to improve design of the trial. The learning from the information will be formulated as a stochastic process as well as the decision process based on the learning. The efficiency of the learning and optimality criterion will be discussed.

Cet exposé traitera de l'utilisation des informations collectées de manière séquentielle au cours d'un essai clinique afin d'améliorer la conception de l'essai. Nous formulerons l'apprentissage tiré des informations comme un processus stochastique ainsi que le processus de décision basé sur l'apprentissage. Nous discuterons de l'efficacité de l'apprentissage et du critère d'optimalité.

[13:45-14:00]

Kehinde I. Olobatuyi (University of Victoria, Canada) **Laura Cowen** (University of Victoria) **Caroline Colijn** (Simon Fraser University) **Patrick Brown** (University of Toronto)

Hierarchical Zero-Inflated Occupancy Model for Estimating Population Abundance of Bacteria Serotypes with Low-Detection Probabilities

Modèle hiérarchique d'occupation avec excès de zéros pour l'estimation de l'abondance de la population de sérotypes bactériens avec de faibles probabilités de détection

Detection probabilities of a serotype within species of bacteria can be extremely low which implies that only the small fraction of the population abundance of serotype is detected. We propose an occupancy model called Zero-Inflated occupancy model for estimating the occupancy rates of serotypes when the non-detection probabilities are almost 1. The model provides a flexible framework enabling covariate information to be included in the model to harness the effect on the detection probabilities. We develop a hierarchical Bayesian model and compare our model to both ordinary occupancy model and Poisson occupancy model. We add augmented data of 15,000 to the 32 serotypes bacteria data detected in 3 occasions. While other models underestimate the serotype abundance in the population, our model estimates the occupancy rates of the serotypes in the population as 0.17. We estimate over 2500 serotypes in the bacteria which corroborates past studies on the serotypes, (Ohad et al., 2014).

La probabilité de détection d'un sérotype au sein d'une espèce bactérienne peut être extrêmement faible, si bien que seule une petite fraction de la population du sérotype est détectée. Nous proposons un modèle d'occupation avec excès de zéros pour estimer les taux d'occupation des sérotypes lorsque la probabilité de non-détection est proche de 1. Le modèle fournit un cadre flexible qui permet d'inclure des informations sur les covariables afin d'exploiter l'effet sur les probabilités de détection. Nous développons un modèle bayésien hiérarchique et comparons notre modèle au modèle d'occupation ordinaire et au modèle d'occupation de Poisson. Nous ajoutons des données augmentées de 15 000 aux 32 sérotypes de bactéries détectés à trois occasions. Alors que les autres modèles sous-estiment l'abondance des sérotypes dans la population, notre modèle estime les taux d'occupation des sérotypes dans la population à 0,17. Nous estimons à plus de 2500 le nombre de sérotypes dans les bactéries, ce qui corrobore des études antérieures sur les sérotypes (Ohad et al., 2014).

[14:00-14:15]

Wanhua Su (MacEwan University)

Pre-Post Test Using Partial Pairs with Missing Identities

Pre-post test utilisant des paires partielles avec identités manquantes

Either by design, carelessness or accident, the data in a pre-post study may consist of a combination of paired and independent data. Various methods have been proposed to analyze partially paired data. However, little has been done for the pre-post test using partially paired data with missing identities. Through a comprehensive simulation study, we investigate the effect of analyzing partially paired data using the two-sample t-test or its nonparametric counterpart (Wilcoxon rank-sum test). Methods based on permutation tests are proposed to deal with the missing ID variable issue. The performance of the proposed methods is evaluated on the simulated data and a couple of real-life applications.

Que ce soit à dessein, par négligence ou par accident, les données d'une pré-post étude peuvent consister en une combinaison de données appariées et indépendantes. Diverses méthodes ont été proposées pour analyser les données partiellement appariées. Cependant, peu de travaux ont été réalisés pour le pre-post test utilisant des données partiellement appariées avec des identités manquantes. Grâce à une étude de simulation complète, nous étudions l'effet de l'analyse de données partiellement appariées à l'aide du test t à deux échantillons ou de son équivalent non paramétrique (test de la somme des rangs de Wilcoxon). Des méthodes basées sur des tests par permutation sont proposées pour traiter le problème des variables d'identités manquantes. L'efficacité des méthodes proposées est évaluée avec des données simulées et dans quelques applications réelles.

[14:15-14:30]

Wenshu Dai (Binghamton University) **Yuan Fang** (Binghamton University) **Sanjeena Dang** (Carleton University)

Mixtures of Regression Models for Microbiome Data

Mélanges de modèles de régression pour des données sur le microbiome

Discrete data such as the microbiome taxa count data resulting from 16S rRNA sequencing are routinely encountered in bioinformatics. Taxa count data in microbiome studies can only reveal relative abundance, therefore are treated as compositional. Analyzing these data presents challenges because they are restricted to a simplex. Additionally, these microbiome taxa counts are affected by other biological environmental covariates such as age, diet, etc. Here, we develop logistic normal multinomial regression (FMR) models for clustering microbiome data. FMR with concomitant variables is also proposed. These models partition samples into homogeneous subpopulations and allow for the investigation of the relationship between relative abundance and covariates within each inferred group. In this presentation, we utilize an efficient framework for parameter estimation using variational Gaussian approximations (VGA). The proposed method is illustrated on simulated datasets and a real data set.

Les données discrètes comme les données de comptage des taxons du microbiome résultant du séquençage de l'ARNr 16S sont régulièrement présentes en bio-informatique. Les données de comptage des taxons dans les études sur le microbiome peuvent seulement révéler une abondance relative et sont donc traitées comme compositionnelles. L'analyse de ces données présente des difficultés parce qu'elles sont restreintes à un simplexe. De plus, ces comptages des taxons du microbiome sont affectés par d'autres covariables environnementales biologiques, comme l'âge, le régime alimentaire, etc. Nous développons ici des modèles de régression logistique multinomiale normale (FMR) pour la mise en grappes des données sur le microbiome. Une FMR avec des variables concomitantes est aussi proposée. Ces modèles partitionnent les échantillons en sous-populations homogènes et permettent d'étudier la relation entre l'abondance relative et les covariables dans chaque groupe inféré. Dans cette présentation, nous nous servons d'un cadre efficace d'estimation des paramètres utilisant des approximations gaussiennes variationnelles (VGA). La méthode proposée est illustrée à l'aide d'ensembles de données simulées et d'un ensemble de données réelles.

[14:30-14:45]

Mohammad Ehsanul Karim (The University of British Columbia) **Momenul Haque Mondol** (University of British Columbia)

Guideline for Choosing the Number of Splits for Double Cross-fit Targeted Maximum Likelihood Estimators

Lignes directrices pour le choix du nombre de fractionnements pour des estimateurs du maximum de vraisemblance ciblé par ajustement croisé double

Using complex machine learning estimators within double robust methods often leads to substantial undercoverage. Together with double robust methods, the double cross-fitting procedure allows researchers to use a wide variety of machine learning estimators. However, guidelines regarding how many times an analyst should split the data are unclear. We want to understand the implication of increasing the number of cross-fitting splits. We compared the traditional Targeted Maximum Likelihood Estimator (TMLE, with a super-learner, covariates supplied as the main effects without correctly transforming them) with 3 versions of double cross-fitting TMLE with various numbers of splits in a simulation study. Increasing split numbers (up to a certain point) is helpful in gradually reducing bias and MSE. We will present a real-world analysis as a motivating example, and provide practical recommendations for practitioners.

L'utilisation d'estimateurs d'apprentissage automatique complexes dans le cadre de méthodes doublement robustes entraîne souvent une sous-couverture importante. Alliée à des méthodes doublement robustes, la procédure d'ajustement croisé double permet aux chercheurs d'utiliser des estimateurs d'apprentissage automatique très diversifiés. Cependant, il n'y a pas de lignes directrices claires sur le nombre de fois qu'un analyste devrait fractionner les données. Nous voulons comprendre l'implication de l'accroissement du nombre de fractionnements dans l'ajustement. Une étude de simulation permet de comparer l'estimateur du maximum de vraisemblance ciblé traditionnel (TMLE, avec un super apprenant, des covariables fournies comme effets principaux sans transformation correcte) avec trois versions de TMLE par ajustement croisé double avec de différents nombres de fractionnements. L'accroissement des nombres de fractionnements (jusqu'à un certain point) est utile pour réduire progressivement le biais et l'erreur quadratique moyenne. Nous présentons également une analyse de données du monde réel à titre d'exemple motivant et faisons des recommandations pratiques pour les praticiens.

[14:45-15:00]

Nam-Anh Tran (The Hospital for Sick Children)

A Comparison of Alternative Ranking Methods in Two-Stage Clinical Trials with Multiple Interventions: An Application to the Anxiolysis for Laceration Repair in Children Trial

Une comparaison des méthodes de classement alternatives dans des essais cliniques à deux étapes avec multiples interventions : une application à l'anxiolyse pour la réparation de lacérations en essai pédiatrique

Multi-arm, multi-stage (MAMS) trials frequently use a control arm to which all interventions are compared, which hinders comparisons among the experimental arms. We developed a MAMS adaptive clinical trial, without requiring a designated control arm, using the Bayesian framework. The trial drops ineffective interventions following an interim analysis before selecting the best intervention at the final stage. We also compare two methods for ranking the interventions; the probability that each intervention is optimal (Pbest) and the surface under the cumulative ranking curve (SUCRA). Both methods are compared across 12 scenarios that vary the number of interventions and the assumed differences between the interventions. The simulated result shows the predictive power associated with SUCRA is higher than that of Pbest as the number of interventions increases, given the type 1 error of both metrics are equal. Finally, we used the proposed method to design an applied clinical trial.

Les essais à multiples branches et multiples étapes (MAMS) utilisent fréquemment une branche contrôle par rapport à laquelle toutes les interventions sont comparées, ce qui nuit aux comparaisons entre les branches expérimentales. Au moyen d'un cadre bayésien, nous avons conçu un essai clinique adaptatif MAMS qui ne requiert pas de branche contrôle. L'essai abandonne les interventions inefficaces à la suite d'une analyse intermédiaire avant de choisir la meilleure intervention à l'étape finale. Nous comparons aussi deux méthodes pour classer les interventions : la probabilité que chaque intervention soit optimale (Pbest) et la surface sous la courbe de classement cumulative (SUCRA). Les deux méthodes sont comparées à travers 12 scénarios qui varient le nombre d'interventions et les différences supposées entre les interventions. Le résultat simulé démontre que la puissance prédictive associée à SUCRA est supérieure à celle de Pbest au fur et à mesure que le nombre d'interventions augmente, étant donné que l'erreur de type 1 des deux mesures est égale. Enfin, nous utilisons les méthodes proposées pour élaborer un essai clinique appliqué.

On Goodness-of-fit and Inference Methods
Sur la qualité de l'ajustement et des méthodes d'inférence

Chair/Président: Mohamed Belalia

Room/Salle: RB 1201

Date: Tuesday May 30 / mardi 30 mai

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-13:45]

Luke Hagar (University of Waterloo) **Nathaniel T. Stevens** (University of Waterloo)

Fast Sample Size Determination for Two-Group Equivalence Tests with Unequal Variances

Détermination rapide de taille d'échantillon pour les tests d'équivalence en deux groupes avec variances inégales

Two-group equivalence tests are typically conducted using two one-sided t-tests, where the relevant test statistics jointly follow a bivariate t-distribution with singular covariance matrix. Unless we assume the two groups of data have the same variance, the degrees of freedom for this bivariate t-distribution are noninteger and unknown a priori. This makes it difficult to analytically find sample sizes that yield desired statistical power for the equivalence test. As such, functionality to compare two groups with unequal variances is not incorporated in popular R software for equivalence testing design. We propose a novel simulation-based method that uses low-dimensional randomized Sobol' sequences to recommend sample sizes for two-group equivalence tests with unequal variances. This method consistently estimates statistical power much more efficiently than traditional simulation-based approaches. We also extend this method to estimate the power curve using root-finding algorithms.

Les tests d'équivalence en deux groupes sont généralement menés à l'aide de deux tests T unilatéraux, dans lesquels les variables à tester pertinentes suivent conjointement une distribution T bivariée avec matrice de covariance singulière. À moins de considérer que les deux groupes de données ont la même variance, les degrés de liberté pour cette distribution T bivariée sont non entiers et inconnus a priori. Il est donc difficile de trouver analytiquement des tailles d'échantillon qui génèrent la puissance statistique désirée pour un test d'équivalence. Conséquemment, la fonctionnalité pour comparer deux groupes à variances inégales n'est pas intégrée dans le logiciel R pour tester l'équivalence. Nous proposons une nouvelle méthode basée sur simulations qui se sert de suites de Sobol randomisées de basse dimension pour recommander les tailles d'échantillon des tests d'équivalence en deux groupes avec variances inégales. Cette méthode estime invariablement la puissance statistique bien plus efficacement que les approches traditionnelles basées sur simulations. Nous élargissons aussi cette méthode pour estimer la courbe de puissance au moyen d'algorithmes de recherche de racine.

[13:45-14:00]

Lenin Arango-Castillo (Bank of Mexico) **Francisco J. Martínez-Ramírez** (Banco de México) **María José Orraca Corona** (Banco de México)

Univariate Measures of Persistence: A Comparative Analysis

Analyse comparative des mesures univariées de la persistance

Using Monte Carlo simulations, we compare different measures of persistence in univariate time series. Persistence is the speed with which a time series returns to the baseline after a shock. Several measures of persistence have been proposed in the literature; however, they indicate different levels of persistence in real data cases. We consider nonparametric, semiparametric, and parametric time-domain and frequency-domain persis-

À l'aide de simulations de Monte-Carlo, nous comparons différentes mesures de la persistance dans des séries chronologiques univariées. La persistance est la vitesse à laquelle une série chronologique revient à la ligne de base après un choc. Plusieurs mesures de la persistance ont été proposées dans la littérature. Cependant, elles indiquent différents niveaux de persistance dans des cas de données réelles. Nous analysons les mesures de la persistance non paramétriques, semi-paramétriques et paramétriques dans les

tence measures and investigate their performance under different anomalies found in practice, such as measurement errors, additive outliers, temporary change outliers, structural change outliers, seasonal and cyclical patterns, near unit-root and unit-root time series. Our results indicate that the nonparametric methods are less affected by the different types of time series anomalies analyzed in this work and that the time-domain measures are the most affected. Finally, we present numerical results and discuss real case data using economic time series.

domaines temporel et fréquentiel. De plus, nous étudions leur efficacité dans différentes anomalies rencontrées dans la pratique, telles que les erreurs de mesure, les valeurs aberrantes additives, les valeurs aberrantes des changements temporaires, les valeurs aberrantes des changements structurels, les modèles saisonniers et cycliques, les séries chronologiques proches de la racine unitaire et à racine unitaire. Nos résultats indiquent que les méthodes non paramétriques sont moins touchées par les différents types d'anomalies des séries chronologiques analysées dans cette étude et que les mesures du domaine temporel sont les plus touchées. Enfin, nous présentons des résultats numériques et discutons de données réelles avec des séries chronologiques économiques.

[14:00-14:15]

Wangshu Tu (Carleton University) **Sanjeena Dang** (Carleton University) **Utkarsh J. Dang** (Carleton University)

Changepoint detection using Gaussian mixture models

Détection de points de changement à l'aide de modèles de mélange gaussien

Changepoint detection aims to find abrupt changes in time series data. Such changes frequently denote substantial modifications to the process; these can vary from simple changes in location to change in the distribution. Traditional changepoint detection methods are frequently constructed for the univariate case and rely on a cost function to assess if a change occurred. Here, we investigate changepoint detection via a clustering framework, and develop a novel changepoint detection algorithm using a finite mixture of regressions with concomitant variables. Through the introduction of a label correction mechanism, the unstructured cluster labels are treated as ordered and distinct segment labels. Different kinds of change can be captured using a parsimonious family of models. The performance of our approach is illustrated by simulation studies and real data analysis.

La détection de points de changement vise à trouver des changements abrupts dans les données d'une série temporelle. De tels changements dénotent souvent des modifications marquées au processus qui peuvent aller de simples changements de la localisation à un changement dans la distribution. Les méthodes traditionnelles de détection des points de changement sont souvent construites pour le cas univarié et reposent sur une fonction de coût pour évaluer si un changement s'est produit. Nous étudions ici la détection de points de changement au moyen d'un cadre de mise en grappes et développons un nouvel algorithme de détection de points de changement en utilisant un mélange fini de régressions avec variables concomitantes. Par l'introduction d'un mécanisme de correction d'étiquetage, les étiquettes de grappes non structurées sont traitées comme des étiquettes de segments ordonnés et distincts. Différents types de changements peuvent être saisis à l'aide d'une famille parcimonieuse de modèles. Des études de simulation et une analyse de données réelles servent à illustrer notre approche.

[14:15-14:30]

Shiyu He (University of Waterloo) **Samuel W.K. Wong** (University of Waterloo)

Integrating Satellite Images and in situ Datasets for Inference and Prediction of Harmful Algal Blooms with Application to Western Lake Erie

Intégration d'images satellitaires et de données in situ pour l'inférence et la prévision d'efflorescences algales nuisibles, avec application à l'ouest du lac Érié

Harmful algal blooms (HABs) are a global environmental concern since they produce toxins and affect both the health of people and marine ecosystems. Substantial efforts have been made to track these blooms using in situ sampling and satellite imaging. However, both types of measurements do not fully capture HAB spa-

Les efflorescences algales nuisibles (HAB) constituent une préoccupation environnementale mondiale, car elles produisent des toxines et affectent à la fois la santé des personnes et les écosystèmes marins. Des efforts considérables ont été déployés pour suivre ces efflorescences à l'aide d'un échantillonnage in situ et de l'imagerie satellitaire. Cependant, ces deux types de mesures ne permettent pas

tial and temporal dynamics, as satellite imaging only measures near-surface HAB concentrations and are limited by cloud cover, and in situ sampling only collects discrete data over large areas. In this study, we propose a statistical framework to integrate the heterogeneous types of observations, while characterizing the spatio-temporal dynamics of HABs. The model provides spatio-temporal inference and prediction of HABs through the integrated nested Laplace approximation (INLA). Via an application in Western Lake Erie during 2014-2016, we show in situ samples tend to yield higher measurements than remote sensing, and that the highest HAB concentration is consistently observed along the western coastline. We also find the HAB in the summer of 2015 is abnormally higher than that of 2014 and 2016, which aligns with the literature reporting a record-high HAB intensity over Western Lake Erie in 2015.

de saisir pleinement la dynamique spatiale et temporelle des HAB, car l'imagerie satellitaire ne mesure que les concentrations d'HAB près de la surface et est limitée par la couverture nuageuse, tandis que l'échantillonnage in situ ne recueille que des données discrètes sur de vastes zones. Dans cette étude, nous proposons un cadre statistique pour intégrer ces types d'observations hétérogènes, tout en caractérisant la dynamique spatio-temporelle des HAB. Le modèle permet l'inférence spatio-temporelle et la prédiction des HAB grâce à l'approximation de Laplace intégrée et imbriquée (INLA). Par une application dans l'ouest du lac Érié au cours de la période 2014-2016, nous montrons que les échantillons in situ tendent à produire des mesures plus élevées que la télédétection, et que la concentration la plus forte de HAB est systématiquement observée le long de la côte ouest. Nous constatons également que les HAB de l'été 2015 sont anormalement plus élevés que ceux de 2014 et 2016, ce qui concorde avec la littérature qui fait état d'une intensité record des HAB dans l'ouest du lac Érié en 2015.

[14:30-14:45]

Tatiana Krikella (University of Waterloo) **Joel A. Dubin** (University of Waterloo)

Personalized Predictive Model that Jointly Optimizes Discrimination and Calibration

Un modèle prédictif personnalisé qui optimise conjointement la discrimination et la calibration

Precision medicine is accelerating rapidly in the health research field. This includes fitting predictive models for individual patients based on patient similarity in an attempt to improve model performance. We propose an algorithm which fits a personalized predictive model (PPM) using the size of a similar subpopulation that optimizes both model discrimination and calibration, as it is criticized that calibration is not assessed as often as discrimination in predictive modelling. We show that the relationship between the size of subpopulation and calibration varies depending on the (usually unknown) true model and the type of model used to fit the data. We investigate the effect of patient weighting on performance, and conclude, as expected, that the choice of the size of subpopulation has a larger effect on the PPM's performance compared to the weight function applied. A comprehensive set of simulation studies and a data analysis is presented to demonstrate our methods and findings.

La médecine de précision croît de façon fulgurante dans le domaine de recherche en santé. Parmi les avancées, on compte l'ajustement de modèles prédictif pour des patients individuels en fonction de la similarité entre patients afin de tenter d'améliorer la performance du modèle. Nous proposons un algorithme qui ajuste un modèle prédictif personnalisé (PPM) au moyen de la taille d'une sous-population similaire qui optimise à la fois la discrimination et la calibration du modèle, car l'on critique souvent le fait que la calibration n'est pas évaluée aussi souvent que la discrimination dans la modélisation prédictive. Nous démontrons que la relation entre la taille de sous-population et la calibration varie selon le vrai modèle (habituellement inconnu) et le type de modèle employé pour ajuster les données. En étudiant l'effet de la pondération de patient sur la performance, nous avons confirmé notre hypothèse : c'est-à-dire que le choix de la taille de population a un impact plus grand sur la performance du PPM par rapport à la fonction de pondération appliquée. Nous présentons des études en simulation et des analyses de données approfondies afin de démontrer nos méthodes et nos résultats.

[14:45-15:00]

Mai Ghannam (University of Windsor) **Sévérien Nkurunziza** (University of Windsor)

Inference in a tensor regression model with a possible change-point

Inférence dans un modèle de régression tensoriel avec un éventuel point de rupture

In this talk, we study an inference problem in a tensor regression model with a possible change-point. Specif-

Dans cet exposé, nous étudions un problème d'inférence dans un modèle de régression tensoriel avec un éventuel point de rupture.

ically, 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 the unrestricted estimator and the restricted estimator as well as their joint asymptotic normality. We also derive a test for testing the restriction and its asymptotic power and we prove that the proposed test is consistent. Beyond the complexity of the testing problem in the tensor model, we consider a very general case where the tensor error term and the regressors do not need to be independent and the dependence structure of the outer-product of the tensor error term and regressors is as weak as that of an L2-mixingale. Finally, we present some simulation results that corroborate the theoretical results.

Plus précisément, nous étudions un problème test d'hypothèse général, sur un paramètre tensoriel, incluant comme cas particulier le problème d'absence de point de rupture. À ce sujet, nous élaborons l'estimateur non-restreint et l'estimateur restreint ainsi que leur normalité asymptotique conjointe. De plus, nous établissons un test asymptotique pour tester la restriction. Nous établissons également la puissance asymptotique du test proposé et nous prouvons que ledit test est convergent. Au-delà de la complexité du modèle tensoriel, nous étudions un cas très général où le terme d'erreur tensorielle et les régresseurs ne sont pas nécessairement indépendants et la structure de dépendance de l'erreur tensorielle et des régresseurs est aussi faible que celle d'une L2-mixingale. En fin, nous présentons les résultats de simulation qui corroborent la théorie.

Chair/Président: Anna Heath

Room/Salle: TB 236

Date: Tuesday May 30 / mardi 30 mai

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-13:45]

Yiran Wang (University of Waterloo) **Martin Lysy** (University of Waterloo) **Audrey Béliveau** (University of Waterloo)

Genetic Mark-Recapture Methods for Estimating Population Size of Stock Populations

Méthodes de marquage et de recapture génétique pour estimer la taille de population des populations de stocks

Genetic mark-recapture (GMR) is a statistical technique used in estimating population size in ecology. By combining the relative abundance of species based on genetic data with population counts obtained for some of the species, GMR allows the estimation of the total population size and the contributions of each species. The current method uses a Lincoln-Petersen estimator and provides an asymptotically unbiased estimate for the total population size. However, the variance estimator does not account for all the uncertainty in the sampling process of the genetics data. As a result, this approach can suffer from a significantly underestimated variance, especially when the genetic sample size is small. In this work, we propose two novel Bayesian GMR models and a modified variance estimator based on the original method to address this issue. The effectiveness of the new methods is investigated via simulation studies and a real data analysis is conducted in the study.

Le marquage-recapture génétique (MRG) est une technique statistique employée dans l'estimation de taille de population en écologie. En combinant l'abondance relative des espèces basée sur des données génétiques avec des chiffres de population obtenus pour certaines des espèces, le MRG permet d'estimer la taille de population totale et la contribution de chaque espèce. La méthode actuelle se sert de l'estimateur Lincoln-Peterson et procure une estimation asymptotique non biaisée de la taille de population totale. Cependant, l'estimateur de variance ne tient pas compte de toutes les incertitudes dans le processus d'échantillonnage des données génétiques. Conséquemment, cette approche peut subir les inconvénients d'une variance grandement sous-estimée, tout particulièrement lorsque la taille d'échantillon génétique est petite. Dans le cadre de ce travail, nous proposons deux nouveaux modèles MRG bayésiens et un estimateur de variance modifié basé sur la méthode initiale pour résoudre le problème. L'efficacité de la nouvelle méthode est examinée dans le cadre d'une étude comportant des études en simulation et des analyses à partir de données réelles.

[13:45-14:00]

Fousseni Sama (Université Laval)

Comparative study of Cis-Regulatory Hubs observed from bulk organized cell data (Hi-C) to those observed in single-cell data (scHi-C)

Étude comparative des pôles cis-régulateurs observés à partir des données des cellules organisées en vrac (Hi-C) à ceux observées dans les données unicellulaires (scHi-C)

The recent development of chromatin conformation capture methods was an incredible opportunity to detect chromosome interactions at the genome scale, revealing regulatory processes involved in diseases. The application of these methods on bulk cells has allowed the development of Cis-Regulatory Hubs (CRHs), a new 3D model highlighting the interactions between distal

Le récent développement des méthodes de capture de la conformation de la chromatine a permis de détecter les interactions entre gènes et éléments de régulation à l'échelle du génome. L'application de ces méthodes a notamment permis le développement de nouveaux modèles 3D tels que les pôles Cis-régulateur, un nouveau modèle de réseaux 3D qui capture les contacts entre gènes et éléments de régulation. Ce modèle a été démontré pertinent dans

elements and promoters. CRHs built in bulk cells has been shown to be a relevant model in complex disease etiology such as schizophrenia. However, its application to single-cell data remains unexplored. Thus, the main goal of this study is to extend CRHs using single-cell Hic data. We will systematically address through extensive simulation study and data imputation the impact of clustering methods to find true interactions. Finally, method performance will be assessed based on CRH inference. Advantages and limits will be presented to application on real data.

l'étiologie de maladies complexes telles que la schizophrénie. Cependant, son application sur des données unicellulaires reste inexplorée. L'objectif ultime de l'étude est d'améliorer les évidences sur la pertinence des pôles Cis-régulateurs pour comprendre les maladies complexes, à partir des données unicellulaires Hic à haute résolution. Il s'agira donc à travers des études de simulation et d'imputation de données d'évaluer les méthodes de clustering, et leur impact sur l'inférence des réseaux. Les avantages et limites des méthodes seront discutés.

[14:00-14:15]

Devan G. Becker (Wilfrid Laurier University)

Estimating Proportions of Genetic Variants and the Variants Themselves from Wastewater Using Spatiotemporal Information

Estimation des proportions de variants génétiques et des variants mêmes dans les eaux usées à l'aide de données spatiotemporelles

Only fragments of genomes can be extracted from wastewater samples, so models are needed to estimate proportions of Variants of Concern (VOCs). This relies on known lists of mutations to define a VOC. I extend existing models with structural time series models which assume the proportion of a VOC in each location is a deviation from a city-wide proportion, using either an autoregressive structure or a Gaussian process for the temporal trend. An extension of this model estimates the proportions simultaneously with the definitions of the VOCs using unsupervised machine learning. This model takes information from multiple samples to find patterns of mutations (basis functions) that appear together, with proportions (coefficients) constrained to be smooth over time and space. I find concordance between temporal trends of estimated VOCs and known VOCs. This allows for tracking variants without clinical sequencing (which is costly and often biased).

Seuls des fragments de génomes peuvent être extraits d'échantillons d'eaux usées, de sorte que des modèles sont nécessaires pour estimer les proportions de variants préoccupants. Cette méthode s'appuie sur des listes connues de mutations pour définir un variant préoccupant. J'élargis les modèles actuels avec des modèles structurels de séries temporelles qui supposent que la proportion d'un variant préoccupant dans chaque lieu est un écart par rapport à une proportion à l'échelle de la ville, en utilisant soit une structure autorégressive, soit un processus gaussien pour la tendance temporelle. Une extension de ce modèle permet d'estimer les proportions en même temps que les définitions des variants préoccupants à l'aide d'un apprentissage automatique non supervisé. Ce modèle utilise des données provenant de plusieurs échantillons pour trouver des modèles de mutations (fonctions de base) qui apparaissent ensemble, avec des proportions (coefficients) contraintes à être lisses dans le temps et l'espace. Je constate une concordance entre les tendances temporelles des variants préoccupants estimés et des variants préoccupants connus. Ce modèle permet de suivre les variants sans séquençage clinique (lequel est coûteux et souvent biaisé).

[14:15-14:30]

Lucy Gao (University of British Columbia) **Anna Neufeld** (University of Washington) **Joshua Popp** (Johns Hopkins University) **Alexis Battle** (Johns Hopkins University) **Daniela Witten** (University of Washington)

Inference after latent variable estimation for single-cell RNA sequencing data

Inférence après estimation de la variable latente pour des données de séquençage de l'ARN unicellulaire

In the analysis of single-cell RNA sequencing data, researchers often first characterize the variation between cells by estimating a latent variable, representing some aspect of the individual cell state. They then test each gene for association with the estimated latent variable.

Dans une analyse de données de séquençage de l'ARN unicellulaire, il arrive souvent que les chercheurs caractérisent d'abord la variation entre les cellules en estimant une variable latente représentant un aspect de l'état d'une cellule individuelle. Ils évaluent ensuite chaque gène pour trouver une association avec la

If the same data are used for both of these steps, then standard methods for computing p-values and confidence intervals in the second step will fail to achieve statistical guarantees such as Type 1 error control or nominal coverage. Furthermore, approaches such as sample splitting that can be fruitfully applied to solve similar problems in other settings are not applicable in this context. We introduce count splitting, an extremely flexible framework that allows us to carry out valid inference in this setting, for virtually any latent variable estimation technique and inference approach, under a Poisson assumption. We demonstrate the Type 1 error control and power of count splitting in a simulation study, and apply count splitting to a dataset of pluripotent stem cells differentiating to cardiomyocytes.

variable latente estimée. Si les mêmes données sont utilisées aux deux étapes, alors les méthodes standards de calcul des valeurs-p et des intervalles de confiance à la seconde étape échoueront à fournir des garanties statistiques, comme le contrôle des erreurs de type 1 et la couverture nominale. De plus, des approches comme le fractionnement d'échantillons qui peuvent être appliquées avec succès à la solution de problèmes similaires dans d'autres contextes ne sont pas applicables dans celui-ci. Nous présentons un fractionnement de décompte, soit un cadre très souple qui permet de faire une inférence valide dans ce contexte, pratiquement pour toute technique d'estimation de la variable latente et toute approche par inférence, sous hypothèse Poisson. Par une étude de simulation, nous montrons le contrôle des erreurs de type 1 et le pouvoir d'un fractionnement de décompte et appliquons le fractionnement de décompte à un ensemble de données sur la pluripotence de cellules souches qui se différencient des cellules myocardiques.

[14:30-14:45]

Pingzhao Hu (Western University) **Leann Lac** (University of Manitoba) **Eric Lin** (University of Toronto) **Boyuan Liu** (University of Toronto) **Daryl L.X. Fung** (University of Manitoba) **Carson K. Leung** (University of Manitoba)

scGMM-VGAE: A Gaussian Mixture Model-based Variational Graph Autoencoder Algorithm for Clustering Single-cell RNA-seq Data

scGMM-VGAE : un algorithme à autoencodeur de graphe variationnel basé sur un modèle de mélanges gaussiens pour regrouper les données de séquençage de l'ARN à cellule unique

Traditional statistical methods have limitations in clustering high dimensional single-cell RNA sequencing (scRNA-seq) data which contain many biological and technical zeros. In this study, we propose a hybrid model which combines a Gaussian mixture model with a variational graph autoencoder algorithm as an extendable end-to-end framework to improve the cell clustering performance on scRNA-seq data. The entire algorithm is optimized simultaneously by a designed loss function. We apply it to three labeled and three simulated scRNA-seq datasets. We consider adjusted Rand index, normalized mutual information, and Silhouette coefficient as performance metrics to evaluate the clustering performance of proposed method in comparison to four selected baseline models. The results show that the proposed method outperforms the baseline methods and has great stability in cell clustering.

Les méthodes statistiques traditionnelles sont limitées pour le regroupement de données de séquençage de l'ARN à cellule unique (scRNA-seq) de grande dimension contenant plusieurs zéros biologiques et techniques. Dans le cadre de cette étude, nous proposons un modèle hybride combinant un modèle de mélanges gaussien avec un algorithme à autoencodeur de graphe variationnel en guise de cadre extensible de bout en bout afin d'améliorer l'efficacité du regroupement de cellules pour des données scRNA-seq. L'algorithme en entier est simultanément optimisé par une fonction de perte spécialement conçue. Nous l'appliquons à trois ensembles de données de scRNA-seq simulés et trois autres étiquetés. Nous tenons compte de l'indice de Rand ajusté, de l'information mutuelle normalisée et du coefficient de silhouette en guise de mesures de performance pour évaluer l'efficacité de regroupement de la méthode proposée par rapport aux quatre modèles de référence sélectionnés. Les résultats démontrent que la méthode proposée surpasse les méthodes classiques et stabilise grandement le regroupement de cellules.

Advanced statistical methods for complex data with measurement error
Méthodes statistiques avancées pour les données complexes avec erreur de mesure

Chair/Président: Li-Pang Chen

Organizer/Responsable: Li-Pang Chen

Room/Salle: AT 102

Date: Tuesday May 30 / mardi 30 mai

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-16:00]

Li-Pang Chen (National Chengchi University) **Grace Y. Yi** (University of Western Ontario)

Variable selection and estimation for the average treatment effect with error-prone confounders

Sélection de variables et estimation pour l'effet de traitement moyen avec confondants sujets à erreur

In the framework of causal inference, the inverse probability weighting estimation method and its variants have been commonly employed to estimate the average treatment effect. Such methods, however, are challenged by the presence of irrelevant pre-treatment variables and measurement error. Ignoring these features and naively applying the usual inverse probability weighting estimation procedures may typically yield biased inference results. In this article, we develop an inference method for estimating the average treatment effect with those features taken into account. We establish theoretical properties for the resulting estimator and carry out numerical studies to assess the finite sample performance of the proposed estimator.

Dans le cadre de l'inférence causale, on emploie souvent la méthode d'estimation de pondération par probabilité inverse et ses variants pour estimer l'effet de traitement moyen. Cependant, de telles méthodes font face à des défis en présence de variables prétraitement non pertinentes et des erreurs de mesure. Si on ne tient pas compte de ces facteurs et appliquons de façon naïve les procédures d'estimation de pondération par probabilité inverse, on risque d'obtenir des résultats d'inférence biaisés. Dans cet article, nous développons une méthode d'inférence pour estimer l'effet de traitement moyen en tenant compte de ces facteurs. Nous établissons des propriétés théoriques pour l'estimateur conçu et menons des études numériques afin d'évaluer la performance d'échantillon fini de l'estimateur proposé.

[16:00-16:30]

Jou-Chin Wu (National Chengchi University) **Li-Pang Chen** (National Chengchi University)

Semiparametric Estimation for Error-prone Partial Linear Single-index Models

Estimation semi-paramétrique pour les modèles linéaires partiels à indice unique et sujets à erreur

In this talk, we mainly discuss partially linear single-index models (PLSIM) subject to measurement error in both parametric and nonparametric terms. With the involvement of multivariate covariates, a few of them are informative while most of them are not. We propose the three stage procedure to eliminate measurement error effects and select important variables for linear predictor term and single-index part. To implement the computation efficiently, we develop a boosting algorithm which enables us to select variables and estimate the parameters. Theoretical results are established to justify the proposed method. In addition, we also examine sta-

Lors de cet exposé, nous aborderons principalement les modèles linéaires partiels à indice unique (PLSIM) soumis à une erreur de mesure en termes paramétriques et non paramétriques. En ajoutant de multiples covariables, on constate que bien peu d'entre elles sont informatives. Nous proposons une procédure en trois étapes pour éliminer les effets d'erreur de mesure et sélectionner les variables importantes pour le prédicteur linéaire et l'indice unique. Afin d'intégrer l'efficacité informatique, nous concevons un algorithme renforcé qui nous permet de choisir des variables et estimer les paramètres. Les résultats théoriques démontrent l'efficacité de la méthode proposée. De plus, nous étudions aussi les propriétés statistiques de la procédure renforcée, comprenant les

tistical properties of the boosting procedure, including convergence and validity of variable selection. Numerical studies, including simulation and data analysis, are conducted to assess the performance of the proposed method.

convergences et la validité de la sélection de variables. Nous menons des études numériques, y compris la simulation et des analyses de données, afin d'évaluer la performance de la méthode proposée.

[16:30-17:00]

Qihuang Zhang (McGill University) **Grace Y. Yi** (University of Western Ontario)

Generalized Network Structured Models with Mixed Responses subject to Measurement Error and Misclassification

Modèles structurés de réseau généralisés avec des réponses mixtes sujettes à l'erreur de mesure et à la mauvaise classification

Research of complex associations between gene networks and multiple responses has attracted increasing attention. A challenge is posited by the presence of a genetic network. Moreover, mismeasurement of responses introduces additional complexity to distort usual inferential procedures. In this talk, we discuss the problem with mixed binary and continuous responses that are subject to mismeasurement and associated with complex structured covariates. We first start with the case where data are precisely measured. We propose a generalized network structured model with a two-step procedure. In the first step, we employ a Gaussian graphical model to facilitate the covariates network structure. In the second step, we incorporate the estimated graphical structure of covariates and develop an estimating equation method. Furthermore, we extend the development to accommodate mismeasured responses. We analyze the mice data arising from a genome-wide association study for implementation.

La recherche d'associations complexes entre les réseaux de gènes et les réponses multiples suscite une attention croissante. La présence d'un réseau génétique pose un défi. De plus, la mauvaise mesure des réponses introduit une complexité supplémentaire qui fausse les procédures inférentielles habituelles. Dans cet exposé, nous discutons du problème des réponses mixtes binaires et continues qui sont sujettes à une mauvaise mesure et associées à des covariables structurées complexes. Nous commençons par le cas où les données sont mesurées avec précision. Nous proposons un modèle structuré en réseau généralisé avec une procédure en deux étapes. Dans la première étape, nous utilisons un modèle graphique gaussien pour faciliter la structure en réseau des covariables. Dans la deuxième étape, nous incorporons la structure graphique estimée des covariables et développons une méthode d'équation d'estimation. En outre, nous étendons le développement à la prise en compte des réponses mal mesurées. Nous analysons les données de souris provenant d'une étude d'association à l'échelle du génome pour la mise en œuvre.

Chair/Président: Johanna de Haan-Ward

Organizer/Responsable: Johanna de Haan-Ward

Room/Salle: TB 240

Date: Tuesday May 30 / mardi 30 mai

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-16:00]

Rowenna Gryba (The University of British Columbia) **Andrew VonDuyke** (North Slope Borough Department of Wildlife Management) **Henry Huntington** (Huntington Consulting) **Billy Adams** (North Slope Borough Department of Wildlife Management) **Brower Frantz** (Utqiaġvik, Alaska) **Justin Gatten** (Utqiaġvik, Alaska) **Qaiyaan Harcharek** (Utqiaġvik, Alaska) **Robert Sarren** (Utqiaġvik, Alaska) **Greg Henry** (University of British Columbia) **Marie Auger-Méthé** (University of British Columbia)

Disrupting the Traditions of Science: Indigenous Knowledge and Habitat Modelling

Perturbation des traditions scientifiques : connaissances autochtones et modélisation de l'habitat

Analysis of species-habitat relationships typically relies on 'western science' data (e.g., satellite telemetry data). Modelling species-habitat relationships can help predict impacts of climate change and provide understanding of species habitat use for conservation and management. Indigenous Knowledge (IK) embodies a wealth of ecological information and although IK is recognized as a valuable knowledge source it is rarely used in statistical methods. We provide a case study showing how IK of ringed seals can be used as a sole data source (i.e., no other species observations) to identify habitat use and important areas for this species. We estimated probabilities of habitat use from IK documented on ringed seal and used beta regression to estimate the relationships between IK ringed seal probabilities and environmental covariates, then predict spatial variation in habitat use. We show similarity between IK and satellite telemetry data from the same species but also that telemetry data may not always reflect population level habitat use. This work was completed in collaboration with Indigenous hunters who provided the IK and extensive review of results.

L'analyse des relations entre les espèces et l'habitat repose généralement sur des données « scientifiques occidentales » (par exemple, des données de télémétrie par satellite). La modélisation des relations entre les espèces et leur habitat peut aider à prévoir les effets du changement climatique et à comprendre l'utilisation de l'habitat par les espèces à des fins de conservation et de gestion. Le savoir autochtone (SA) renferme une mine d'informations écologiques et, bien qu'il soit reconnu comme une source de connaissances précieuse, il est rarement utilisé dans les méthodes statistiques. Nous présentons une étude de cas montrant comment le savoir autochtone sur les phoques annelés peut être utilisé comme seule source de données (c'est-à-dire sans autre observation de l'espèce) pour identifier l'utilisation de l'habitat et les zones importantes pour cette espèce. Nous avons estimé les probabilités d'utilisation de l'habitat à partir du savoir autochtone documenté sur le phoque annelé et utilisé la régression bêta pour estimer les relations entre les probabilités SA du phoque annelé et les covariables environnementales, puis pour prédire la variation spatiale de l'utilisation de l'habitat. Nous montrons la similitude entre le savoir autochtone et les données de télémétrie par satellite de la même espèce, mais aussi que les données de télémétrie ne reflètent pas toujours l'utilisation de l'habitat au niveau de la population. Ce travail a été réalisé en collaboration avec des chasseurs autochtones qui ont fourni le savoir autochtone et ont procédé à un examen approfondi des résultats.

[16:00-16:30]

Ethan Lawler **Chris Field** (Dalhousie University) **Kim Whoriskey** (Dalhousie University) **Hugues Benoit** (Fisheries and Oceans Canada) **Joanna Elizabeth Mills Flemming** (Dalhousie University)

Estimating Space Use from Animal Telemetry Data using Non-Parametric Langevin Diffusion

Estimation de l'utilisation de l'espace à partir de données de télémétrie animale à l'aide de la diffusion de Langevin non paramétrique

Where do animals choose to spend their time? The utilization distribution of an animal is a probability density function giving the relative probability of space use. Knowing the utilization distribution of a population allows one to, for example, probe into that population's preferred habitat and determine how often they are likely to co-occur with other species. A common data source for estimating a utilization distributions is the animal track, time-stamped location pings of a tagged animal. We develop a non-parametric version of a recent Langevin diffusion model of animal movement that allows to estimate the utilization distribution while accounting for the temporal autocorrelation inherent in a movement track. Our method uses numerous recent developments in computationally-efficient methods to model the unknown utilization distribution as spatial random field. We will showcase our model using a set of grey seal tracks spanning the waters of Atlantic Canada.

Où les animaux choisissent-ils de passer leur temps ? La distribution d'utilisation d'un animal est une fonction de densité de probabilité donnant la probabilité relative de l'utilisation de l'espace. Connaître la répartition de l'utilisation d'une population permet, par exemple, de sonder l'habitat préféré de cette population et de déterminer la fréquence de cohabitation avec d'autres espèces. Une source de données courante pour estimer une distribution d'utilisation est la trace de l'animal, c'est-à-dire les signaux de localisation horodatés d'un animal marqué. Nous concevons une version non paramétrique d'un récent modèle de diffusion de Langevin du mouvement animal qui permet d'estimer la distribution d'utilisation tout en tenant compte de l'autocorrélation temporelle inhérente à une trace de mouvement. Nous utilisons de nombreux développements récents pour des méthodes de calcul efficaces afin de modéliser la distribution de l'utilisation inconnue comme un champ aléatoire spatial. Nous présenterons notre modèle avec un ensemble de traces de phoques gris parcourant les eaux du Canada atlantique.

[16:30-17:00]

Inesh Prabuddha Munaweera Arachchilage (University of Manitoba) **Saman Muthukumarana** (University of Manitoba) **Darren Gillis** (University of Manitoba)

Bayesian modelling and simulation methods for acoustic telemetry studies

Modélisation bayésienne et méthodes de simulation pour des études de télémétrie acoustique

Bayesian models have been popular in modelling complex ecological data collected using modern animal tracking technologies such as acoustic telemetry for multiple reasons, including their extreme flexibility, ability to incorporate prior knowledge and better precision. Acoustic telemetry systems have been increasingly used to study fish movement patterns, habitat use, and estimate demographic parameters. However, the data generated using omnidirectional acoustic telemetry studies are complex, with multiple sources of variability. In this talk, we present a few applications of Bayesian hierarchical modelling for acoustic telemetry data to estimate individual-level Walleye moment paths in Lake Winnipeg and study the variations in survival probabilities of Arctic Char in the Cambridge Bay region. Furthermore, we address practical challenges in modelling such data and highlight the advantages of the Bayesian approach over classical approaches.

Les modèles bayésiens ont gagné en popularité dans la modélisation de données écologiques complexes collectées à partir de technologies de suivi des animaux comme la télémétrie acoustique pour plusieurs raisons comme : leur flexibilité extrême, leur capacité à intégrer des connaissances antérieures et leur précision supérieure. Les systèmes de télémétrie acoustiques sont de plus en plus adoptés pour l'étude de mouvement des poissons, de l'utilisation de l'habitat et pour estimer les paramètres démographiques. Cependant, les données générées par les études de télémétrie acoustique omnidirectionnelle sont complexes et contiennent plusieurs sources de variabilité. Lors de cet exposé, nous présentons quelques applications de modélisation hiérarchique bayésienne pour les données de télémétrie acoustique dans le but d'estimer les déplacements de dorés jaunes au niveau individuel dans le lac Winnipeg et d'étudier les variations dans les probabilités de survie des ombles chevalier dans la région de Cambridge Bay. De plus, nous abordons les défis pratiques de modélisation de telles données et soulignons les avantages de l'approche bayésienne par rapport aux approches classiques.

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

Room/Salle: AT 101

Date: Tuesday May 30 / mardi 30 mai

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-16:00]

Osvaldo Espin-Garcia (Western University)

Leveraging Genetic Correlation Methods for Polygenic Risk Score Construction in Multivariate Outcomes via Penalized Linear Regression Models

Tirer avantage des méthodes de corrélation génétique pour la construction de score de risque polygénique dans des résultats multivariés au moyen de modèles de régression linéaire pénalisés

Polygenic risk scores (PRS) quantify the genetic contribution of an individual's genotype to a trait, e.g. disease or phenotype. PRS can be used to group subjects into different risk strata and can thus be treated as predictors in clinical and epidemiological studies to help us better understand such traits. However, PRS methods typically focus on a single trait at a time, ignoring the potential simultaneous influence of a gene on multiple traits. To address this limitation in the existing methods, we propose a model that incorporates a genetic correlation matrix among traits into the cost function of a penalized regression framework with the objective to improve the predictive ability of multi-trait PRS models. The proposed method is evaluated against alternative approaches when both a single trait and multiple traits are of interest via comprehensive simulation studies. Lastly, an illustration using data from a study of smoking behaviour traits is presented.

Les scores de risque polygénique (PRS) quantifient la contribution génétique d'un génotype individuel à un trait (p. ex. maladie ou phénotype). Le PRS peut servir à grouper des sujets en différentes strates de risque et donc être traitées en tant que prédicteurs en essais cliniques et épidémiologiques afin de mieux comprendre de tels traits. Cependant, les méthodes de PRS se concentrent habituellement sur un seul trait à la fois, et ignorent la potentielle influence simultanée d'un gène sur plusieurs traits. Dans le but de résoudre cette restriction, nous proposons un modèle qui intègre une matrice de corrélation génétique parmi les traits dans la fonction de coût dans un cadre de régression pénalisé, avec comme objectif d'améliorer la capacité prédictive des modèles de PRS à traits multiples. Nous comparons la méthode proposée à d'autres approches dans le cadre d'études en simulations approfondies où à la fois un trait et divers traits nous intéressent. Enfin, nous illustrons les données tirées d'une étude sur les traits associés au comportement de fumeur.

[16:00-16:30]

Kevin McGregor (York University)

Dependence modelling in count-based genomic data in the presence of zero-inflation

Modélisation de la dépendance des données génomiques basée sur le comptage en présence d'un excès de zéros

Compositional data comprise vectors that describe the constituent parts of a whole. Data arising from various -omics platforms such as 16S and RNA-sequencing are compositional in nature. Correlations between features on raw counts have no meaningful interpretation. I previously developed metrics of proportionality for count-

Les données de composition comprennent des vecteurs qui décrivent les éléments constitutifs d'un tout. Les données issues de diverses plateformes omiques, telles que le séquençage de l'ARN et du gène 16S, sont de nature compositionnelle. Les corrélations entre les caractéristiques sur les comptes bruts n'ont pas d'interprétation significative. J'ai précédemment mis au point

based compositional data. Even though these metrics can handle a modest number of zeros, they do not perform as well when the proportion of zero counts is high. Estimators for Kendall's tau for a zero-inflated models have been proposed (Pimental et al. 2015). In this talk, I will discuss the development of a version of this estimator suitable for compositional data, in particular, in the context of a modified zero-inflated logit-normal multinomial model where dependence is specified through a copula structure. I will also show results from applying this estimator on microbiome metagenomic sequencing samples from a pediatric Multiple Sclerosis study.

des mesures de proportionnalité pour les données de composition basées sur le comptage. Même si ces mesures peuvent prendre en compte un petit nombre de zéros, elles ne sont pas aussi efficaces lorsque la proportion de zéros est élevée. Des estimateurs du tau de Kendall pour les modèles sans gonflement ont été proposés (Pimental et coll. 2015). Je présenterai une version de cet estimateur adaptée aux données composites, en particulier dans le contexte d'un modèle multinomial logit-normal modifié sans excès de zéros où la dépendance est spécifiée par une structure de copule. Je montrerai également les résultats de l'application de cet estimateur sur des échantillons de séquençage métagénomique du microbiome provenant d'une étude sur la sclérose en plaques chez l'enfant.

[16:30-17:00]

Elena Tuzhilina (University of Toronto) **Trevor Hastie** (Stanford University) **Mark Segal** (University of California, San Francisco)

Statistical Curve Models for Inferring 3D Chromatin Architecture

Modèles de courbes statistiques pour déduire l'architecture de la chromatine en 3D

Conformation reconstruction is an important challenge in computational biology. In this project we develop a model for the 3D spatial organization of chromatin, a crucial component of numerous cellular processes. The central object in this study is the so-called contact matrix. It represents the frequency of contacts between each pair of genomic loci; it can thus be used to infer the 3D structure. The following heuristic is usually applied to link the contact counts to the conformation: loci that are close to each other in 3D space should have a higher contact value. Most of the existing methods that operate on contact matrices are based on multidimensional scaling (MDS) and produce reconstructed 3D configurations in the form of a polygonal chain. However, none of them exploit the fact that the target solution should be a smooth curve in 3D. The smoothness attribute is either ignored or indirectly addressed via introducing highly non-convex penalties in the model. This typically leads to increased computational complexity and instability of the reconstruction algorithm. In this work we develop a novel approach for modeling chromatin directly by a smooth curve. The performance of the methods is illustrated on real Hi-C data computed for chromosome 20 and evaluated by means of orthogonal multiplex FISH imaging.

La reconstruction de la conformation est un défi important en biologie informatique. Dans ce projet, nous développons un modèle pour l'organisation spatiale en 3D de la chromatine, composant crucial de nombreux processus cellulaires. L'objet central de cette étude est la matrice de contact. Celle-ci représente la fréquence des contacts entre chaque paire de loci génomiques ; elle peut donc être utilisée pour déduire la structure 3D. L'heuristique suivante est généralement appliquée pour relier le nombre de contacts à la conformation : les loci qui sont proches l'un de l'autre dans l'espace 3D devraient avoir une valeur de contact plus élevée. La plupart des méthodes existantes qui opèrent sur les matrices de contact sont basées sur la mise à l'échelle multidimensionnelle (MDS) et produisent des configurations 3D reconstruites sous la forme d'une chaîne polygonale. Cependant, aucune d'entre elles n'exploite le fait que la solution cible doit être une courbe lisse en 3D. L'attribut de fluidité est soit ignoré, soit traité indirectement par l'introduction de pénalités hautement non convexes dans le modèle. Cela conduit généralement à une complexité de calcul accrue et à l'instabilité de l'algorithme de reconstruction. Dans ce travail, nous développons une nouvelle approche pour modéliser la chromatine directement par une courbe lisse. Les performances des méthodes sont illustrées sur des données Hi-C réelles calculées pour le chromosome 20 et évaluées au moyen de l'imagerie FISH multiplex orthogonale.

Causal inference for complex data
Inférence causale pour les données complexes

Chair/Président: Archer Yang

Organizer/Responsable: Archer Yang

Room/Salle: AT 301

Date: Tuesday May 30 / mardi 30 mai

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-15:53]

Yunan Wu (University of Texas at Dallas) **Lan Wang** (University of Miami) **Haoda Fu** (Eli Lilly and Company)

Model-Assisted Uniformly Honest Inference for Optimal Treatment Regimes in High Dimension

Inférence uniformément honnête assistée par modèle pour les régimes de traitements optimaux en grande dimension

We develop new tools to quantify uncertainty in optimal decision making and to gain insight into which variables one should collect information about given the potential cost of measuring a large number of variables. We investigate simultaneous inference to determine if a group of variables is relevant for estimating an optimal decision rule in a high-dimensional semiparametric framework. The unknown link function permits flexible modeling of the interactions between the treatment and the covariates, but leads to nonconvex estimation in high dimension and imposes significant challenges for inference. We first establish that a local restricted strong convexity condition holds with high probability and that any feasible local sparse solution of the estimation problem can achieve the near-oracle estimation error bound. We verify that a wild bootstrap procedure based on a debiased version of the local solution can provide asymptotically honest uniform inference on optimal decision making.

Nous concevons de nouveaux outils afin de quantifier l'incertitude dans la prise de décision optimale et aussi pour mieux savoir à partir de quelles variables il faut recueillir de l'information compte tenu du coût potentiel de la mesure d'un grand nombre de variables. Nous examinons une inférence simultanée dans le but de déterminer si un groupe de variables est pertinent pour l'estimation d'une règle de décision optimale dans un cadre semi-paramétrique de haute dimension. La fonction de liaison inconnue permet une modélisation flexible des interactions entre le traitement et les covariables, mais mène à une estimation non convexe à grande dimension et pose de sérieux défis pour l'inférence. Nous établissons d'abord qu'une condition locale de forte convexité restreinte reste pertinente avec haute probabilité et que toute solution éparse locale possible du problème d'estimation peut mener à une limite d'erreur d'estimation semblable au cas idéal. Nous vérifions qu'une procédure bootstrap sauvage basée sur une version non biaisée de la solution locale peut générer une inférence uniforme asymptotiquement honnête relative à la prise de décision optimale.

[15:53-16:15]

Eric Morenz (University of Washington) **Marco Carone** (University of Washington) **Charles Wolock** (University of Washington)

Debiased machine learning for survival functionals

Apprentissage machine débiaisé pour des fonctionnelles de survie

Learning causal effects on time-to-event endpoints can be challenging because survival times may be partially observed due to censoring and systematically biased due to truncation. In this work, we present nonparametric estimators of a counterfactual survival time distribution for use when the data are subject to covariate-dependent

L'apprentissage des effets causaux sur les paramètres de temps jusqu'à l'événement peuvent être problématiques parce que les temps de survie peuvent être partiellement observés en raison de la censure et systématiquement biaisés à cause de la troncature. Notre travail présente des estimateurs non paramétriques d'une distribution des temps de survie contrefactuels à utiliser lorsque les

left truncation and right censoring. Our inferential procedures explicitly allow the integration of flexible machine learning tools for nuisance estimation, and enjoy certain robustness properties. The approach we propose can be directly used to make pointwise or uniform inference on smooth summaries of the counterfactual survival time distribution, and can be valuable even in the absence of interventions, when summaries of a marginal survival distribution are of interest. We showcase how our procedures can be used to learn a variety of inferential targets and investigate their performance in simulation studies.

données sont assujetties à une troncature à gauche dépendante des covariables et à une censure à droite. Nos procédures d'inférence permettent explicitement l'intégration d'outils d'apprentissage machine flexibles pour l'estimation des nuisances et sont dotées de certaines propriétés de robustesse. L'approche que nous proposons peut être utilisée directement pour faire une inférence uniforme ou ponctuelle pour les résumés lisses de la distribution des temps de survie contrefactuels et peut être utile, même en l'absence d'interventions, lorsque les résumés d'une distribution de survie marginale revêtent de l'intérêt. Nous montrons comment nos procédures peuvent être utilisées pour apprendre des cibles inférentielles diverses et examinons leur performance dans des études par simulation.

[16:15-16:37]

Ziang Niu (University of Pennsylvania)

Inference for ATE under Unmeasured Confounding with Arbitrary Number of Invalid IVs: Beyond Majority and Plurality Voting

Inférence pour l'effet moyen du traitement en cas de confusion non mesurée avec un nombre arbitraire de VI invalides : au-delà du vote à la majorité et à la pluralité

The instrumental variable method is one of the ways to Identify and infer the average treatment effect (ATE) under the presence of unmeasured confounding. However, finding a set of valid IVs is challenging in practice. Under the recently developed GENIUS framework (Eric et al., 2021), we present a novel three-step estimation procedure allowing for potentially high-dimensional IVs which may consist of an arbitrary number of invalid IVs. We propose a new debiased estimator for estimating ATE and correspondingly develop a computationally efficient multiplier bootstrap method for constructing the confidence interval. We apply the proposed method to a Mendelian Randomization example for illustration.

La méthode des variables instrumentales est l'un des moyens d'identifier et d'inférer l'effet moyen du traitement (ATE) en présence de facteurs de confusion non mesurés. Cependant, trouver un ensemble de variables instrumentales valides est difficile dans la pratique. Dans le cadre récemment développé de GENIUS (Eric et al., 2021), nous présentons une nouvelle procédure d'estimation en trois étapes permettant de prendre en compte des VI potentiellement en haute dimension qui peuvent consister en un nombre arbitraire de VI non valides. Nous proposons un nouvel estimateur débiaisé pour estimer l'ATE et développons en conséquence une méthode bootstrap à multiplicateur efficace en termes de calcul pour construire l'intervalle de confiance. Nous appliquons, à titre d'illustration, la méthode proposée à un exemple de randomisation mendélienne.

[16:37-17:00]

Guanbo Wang (CausaLab)

Examining subgroup-specific treatment effects in multi-source data: source-specific inference and transportability to an external population

Examen des effets de traitement spécifiques à un sous-groupe dans des données multi-sources : inférence spécifique à la source et transférabilité à une population externe

One major challenge in the estimation of effect heterogeneity is that the sample size of the data used is typically not enough to precisely capture how effects vary according to the effect modifiers. Therefore, there is interest in synthesizing evidence across multi-source data to improve the precision of estimators of heterogeneous treatment efficacy. Furthermore, how to combine in-

Un défi majeur dans l'estimation de l'hétérogénéité des effets est que la taille de l'échantillon des données utilisées n'est généralement pas suffisante pour saisir avec précision comment les effets varient en fonction des modificateurs d'effet. C'est pourquoi il est intéressant de synthétiser les données provenant de sources multiples afin d'améliorer la précision des estimateurs de l'hétérogénéité de l'efficacité des traitements. En outre, la manière

formation from multi-source data in a way that is interpretable in a meaningful target population is also not clear. We develop and evaluate methods for using multi-source data to estimate subgroup treatment effects in an external target population or the populations underlying the data sources. We propose a doubly robust estimator that is non-parametrically efficient and allows for nuisance functions to be estimated using machine learning methods. We illustrate the methods in meta-analyses of randomized trials for schizophrenia and bipolar disorder.

de combiner les informations provenant de données multi-sources d'une manière qui soit interprétable dans une population cible significative n'est pas claire non plus. Nous développons et évaluons des méthodes permettant d'utiliser des données multi-sources pour estimer les effets de traitement de sous-groupes dans une population cible externe ou dans les populations sous-jacentes aux sources de données. Nous proposons un estimateur doublement robuste qui est efficace d'un point de vue non paramétrique et qui permet d'estimer les fonctions de nuisance à l'aide de méthodes d'apprentissage automatique. Nous illustrons les méthodes dans des méta-analyses d'essais randomisés pour la schizophrénie et le trouble bipolaire.

Recent Developments in High-Dimensional Inference
Développements récents en matière d'inférence en haute dimension

Chair/Président: Farouk Nathoo

Organizer/Responsable: Farouk Nathoo

Room/Salle: TB 208

Date: Tuesday May 30 / mardi 30 mai

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-16:00]

S. Ejaz Ahmed (Brock University)

Improving the Prediction Performance in High-Dimensional Data Analysis

Améliorer l'efficacité de prédiction dans l'analyse des données de grande dimension

The rapid growth in the size and scope of data sets in a host of disciplines has created a need for innovative statistical strategies for analyzing high-dimensional data. The need for novel statistical strategies to analyze such data sets is pressing. This talk focuses on the development of statistical strategies for a host of sparse regression models in the presence of weak signals. The existing penalized methods have often ignored contributions from the weak signals. However, in real scenario many predictors altogether provide useful information for prediction, although the amount of such useful information in a single predictor might be modest. We proposed a new "post-shrinkage strategy" that considers the joint impact of both strong and weak signals to improve the prediction accuracy of the selected submodel(s). We provide asymptotic properties of suggested estimators and verify it through simulation study. Some real data examples are given to demonstrate the applicability of the proposed strategies. I will present some open related research problems as well.

La croissance rapide de la taille et de la portée des ensembles de données dans un grand nombre de disciplines a créé un besoin de stratégies statistiques novatrices pour l'analyse des données de grande dimension. Il est ainsi urgent d'élaborer de nouvelles stratégies statistiques pour analyser ces ensembles de données. Cette présentation porte sur l'élaboration de stratégies statistiques pour un grand nombre de modèles de régression épars en présence de signaux faibles. Les méthodes pénalisées actuelles ont souvent ignoré les contributions des signaux faibles. Toutefois, dans la réalité, de nombreux prédicteurs fournissent ensemble des données utiles à la prédiction, même si la quantité de celles-ci dans un seul prédicteur peut être faible. Nous proposons une nouvelle « stratégie post-rétrécissement » qui prend en compte les effets conjoints des signaux forts et faibles afin d'améliorer la précision de la prédiction des sous-modèles sélectionnés. Nous fournissons les propriétés asymptotiques des estimateurs proposés et les vérifions au moyen d'une étude de simulation. Enfin, nous donnerons quelques exemples de données réelles pour démontrer l'applicabilité des stratégies proposées et présenterons également quelques problèmes de recherche en cours.

[16:00-16:30]

Linglong Kong (University of Alberta)

Regularized tensor quantile regression with applications to neuroimaging data analysis

Régression quantile tensorielle régularisée et applications à l'analyse de données de neuro-imagerie

Our work develops a regularized tensor quantile regression framework for scalar responses and enables the robust analysis of tensor-variate data. Beyond our presented application to neuroimaging for identifying regions of the hippocampus implicated in Alzheimer's disease, tensor-variate data are ubiquitous in medicine,

Dans le cadre de nos travaux, nous élaborons un cadre de régression quantile tensorielle régularisée pour les réponses scalaires, ce qui permet une analyse robuste des données à variables tensorielles. Outre l'application présentée en neuro-imagerie pour déterminer les régions de l'hippocampe liées à la maladie d'Alzheimer, les données à variables tensorielles sont omniprésentes en

ecology, and other fields where large volumes of data are collected. The tensor quantile model considered in our work stands separate from previously-established tensor regression frameworks and requires its own theoretical investigation. We establish important statistical properties of our tensor effect estimator and convergence properties of our proposed estimation algorithm. To address the high dimensionality of the tensor quantile model and the non-differentiability of the quantile loss function, we assume that the tensor effect admits a Tucker decomposition and perform estimation using smoothing techniques combined with a block relaxation algorithm. Unlike previous two-stage approaches, our methodology simultaneously considers tensor decomposition and model estimation, ensuring that the decomposition is optimally-predictive of the response.

médecine, en écologie et dans d'autres domaines où d'importants volumes de données sont collectés. Le modèle quantile tensoriel considéré dans nos travaux est distinct des cadres de régression tensorielle précédemment établis et nécessite une propre étude théorique. Nous établissons des propriétés statistiques importantes de notre estimateur de l'effet tensoriel et des propriétés de convergence de l'algorithme d'estimation proposé. Pour faire face à la haute dimensionnalité du modèle quantile tensoriel et à la non-différenciabilité de la fonction de perte quantile, nous supposons que l'effet tensoriel admet une décomposition de Tucker et effectuons l'estimation à l'aide de techniques de lissage combinées à un algorithme de relaxation par blocs. Contrairement aux approches précédentes en deux étapes, notre méthodologie prend en compte simultanément la décomposition tensorielle et l'estimation du modèle, garantissant que la décomposition prédit la réponse de manière optimale.

[16:30-17:00]

Abbas Khalili (McGill University) **Gilberto Chavez-Martinez** (McGill Univeristy)

Sparse Estimation in Markov Regime-Switching Models

Estimation éparse dans les modèles de Markov à changement de régime

Markov regime-switching vector auto-regressives are frequently used for modelling heterogeneous and complex relationships between variables in multivariate time series analysis. Applications include analyzing macroeconomic time series such as manufacturing activities, consumer price indices, and housing and asset prices. The most common method of estimation in these models is maximum likelihood estimation (MLE). However, even for moderate data dimension and number of regimes, the MLE becomes unstable. In this talk, we present regularization-based estimators when the number of regimes in the model is correctly or over-specified. We also discuss theoretical and finite-sample performances of the methods, including forecasting, and conclude with a real data analysis.

Les autorégressifs vectoriels à changement de régime de Markov sont fréquemment utilisés pour modéliser des relations hétérogènes et complexes entre variables dans l'analyse de séries chronologiques multivariées. Les applications comprennent l'analyse de séries chronologiques macroéconomiques telles que les activités manufacturières, les indices des prix à la consommation, ainsi que les prix des logements et des actifs. La méthode d'estimation la plus courante dans ces modèles est l'estimation du maximum de vraisemblance (EMV). Cependant, même pour une dimension de données et un nombre de régimes modérés, l'EMV devient instable. Dans cet exposé, nous présentons des estimateurs basés sur la régularisation lorsque le nombre de régimes dans le modèle est correctement ou sur-spécifié. Nous discutons également des performances théoriques et sur échantillon fini des méthodes, y compris les prévisions, et concluons par une analyse de données réelles.

Chair/Président: Thierry Chekouo

Room/Salle: TB 202

Date: Tuesday May 30 / mardi 30 mai

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-15:45]

Hyukjun (Jay) Gweon (Western University) **Shu Li** (Western University)

A hybrid data mining framework for variable annuity portfolio valuation

Cadre hybride d'exploration de données pour l'évaluation de portefeuilles de rentes variables

A variable annuity is a modern life insurance product that offers its policyholders participation in investment with various guarantees. To address the computational challenge of valuing large portfolios of variable annuity contracts, several data mining frameworks based on statistical learning have been proposed in the past decade. Existing methods utilize regression modeling to predict the market value of most contracts. Despite the efficiency of those methods, a regression model fitted to a small amount of data produces substantial prediction errors and thus it is challenging to rely on existing frameworks when highly accurate valuation results are desired or required. In this talk, we propose a hybrid framework that effectively chooses and assesses easy-to-predict contracts using the random forest model while leaving hard-to-predict contracts for the Monte Carlo simulation. The effectiveness of the hybrid approach is illustrated with an experimental study.

Une rente variable est un produit d'assurance-vie moderne qui offre à ses assurés une participation à l'investissement avec diverses garanties. Pour relever le défi informatique que représente l'évaluation de vastes portefeuilles de contrats de rentes variables, plusieurs cadres d'exploration de données basés sur l'apprentissage statistique ont été proposés au cours de la dernière décennie. Les méthodes actuelles utilisent la modélisation de la régression pour prédire la valeur de marché de la plupart des contrats. Malgré l'efficacité de ces méthodes, un modèle de régression adapté à une petite quantité de données produit des erreurs de prédiction substantielles, et il est donc difficile de s'appuyer sur les cadres actuels lorsque des résultats d'évaluation très précis sont souhaités ou requis. Dans cette présentation, nous proposons un cadre hybride qui choisit et évalue efficacement les contrats faciles à prévoir à l'aide du modèle de forêt aléatoire, tout en laissant les contrats difficiles à prévoir pour la simulation de Monte-Carlo. Enfin, nous illustrons l'efficacité de l'approche hybride à l'aide d'une étude expérimentale.

[15:45-16:00]

Alexandru Badescu (University of Calgary) **Maciej Augustyniak** (Université de Montréal) **Jean-François Bégin** (Simon Fraser University) **Sarath Kumar Jayaraman** (University of Calgary)

On the Weak Limits of Affine Option Pricing Models

Des limites faibles des modèles affines d'évaluation des options

This article investigates the weak convergence of discrete-time affine stochastic volatility models driven by both Gaussian and non-Gaussian innovations. Using different assumptions on the converge rates of the underlying parameters we derive various diffusion limits under the physical and risk-neutral measures, the latter being constructed based on an exponential affine pricing kernel. Our results generalize the existing limits for affine Gaussian GARCH models, and provide new insights on the

Cet article étudie la faible convergence des modèles de volatilité stochastique affine à temps discret, pilotés par des innovations gaussiennes et non gaussiennes. En utilisant différentes hypothèses sur les taux de convergence des paramètres sous-jacents, nous dérivons diverses limites de diffusion sous les mesures physiques et neutres au risque, ces dernières étant construites sur la base d'un noyau de tarification affine exponentiel. Nos résultats généralisent les limites existantes pour les modèles GARCH gaussiens affines et donnent de nouvelles indications sur les limites en

continuous-time limits of other non-Gaussian models. Numerical results on the convergence speed of European option prices for the proposed schemes are also provided.

temps continu d'autres modèles non gaussiens. Nous fournissons également des résultats numériques sur la vitesse de convergence des prix des options européennes pour les schémas proposés.

[16:00-16:15]

Maciej Augustyniak (Université de Montréal)

Basis Risk in Variable Annuities

Le risque de base dans les rentes variables

This study provides the first comprehensive analysis of basis risk in the U.S. variable annuity market. These popular personal savings and investment products contain long-term guarantees that expose life insurers to extensive financial risks. Although hedging these guarantees is critical, we find that basis risk severely impedes hedging performance at the individual fund level. We show that insurers can mitigate their exposure to this risk by applying data analytic techniques in their fund mapping process, by hedging with ETFs instead of futures, by including more index funds in their offerings, and through diversification at the portfolio level.

Cette étude fournit la première analyse complète du risque de base sur le marché américain des rentes variables. Ces produits d'épargne et d'investissement populaires contiennent des garanties à long terme qui exposent les compagnies d'assurance-vie à des risques financiers importants. Bien que la couverture de ces garanties soit essentielle, nous constatons que le risque de base entrave sérieusement l'efficacité de la couverture sur le plan des fonds individuels. Nous montrons que les assureurs peuvent atténuer leur exposition à ce risque en appliquant des techniques d'analyse de données dans leur processus de cartographie des fonds, en se couvrant avec des fonds indiciels négociables en bourse plutôt qu'avec des contrats à terme standardisés, en incluant davantage de fonds indiciels dans leur offre et en diversifiant leur portefeuille.

[16:15-16:30]

Frédéric Godin (Concordia University)

Risk allocation through Shapley decompositions, with applications to variable annuities

Allocation du risque à l'aide de décompositions de Shapley et applications aux fonds distincts

A flexible risk decomposition method for life insurance contracts embedding several risk factors is introduced. Hedging can be naturally embedded in the framework. Although the method is applied to variable annuities in this work, it is also applicable in general to other insurance or financial contracts. The approach relies on applying an allocation principle to components of a Shapley decomposition of the gain and loss. The implementation of the allocation method requires the use of a stochastic on stochastic algorithm involving nested simulations. Numerical examples studying the relative impact of equity, interest rate and mortality risk for guaranteed minimal maturity benefit (GMMB) policies conclude our analysis.

Nous présenterons une méthode de décomposition des risques flexible s'appliquant aux contrats d'assurance-vie dans une contexte de risque multi-factoriel. L'inclusion d'un portefeuille de couverture est naturelle au sein de notre méthode. Alors que notre approche est appliquée sur des fonds distincts à l'intérieur de nos travaux, celle-ci peut s'appliquer à un éventail beaucoup plus large de contrats d'assurance et de contrats financiers. L'approche s'appuie sur une décomposition de Shapley appliquée aux gains et pertes associé à un contrat. L'implémentation requiert la réalisation de simulations imbriquées. Des exemples numériques présentés étudient l'impact relatif du risque d'équité, de taux d'intérêt et de mortalité pour un contrat de type "bénéfice garanti à maturité".

[16:30-16:45]

Marie-Claude Vachon (Université du Québec à Montréal) **Anne Mackay** (Université de Sherbrooke)

Optimal stopping with discontinuous, unbounded, and time-dependent reward and applications to variable annuities

Problème de temps d'arrêt optimal avec fonction de gain discontinue, non bornée et dépendante du temps et applications aux fonds distincts

We consider a financial derivative with early exercise

Nous considérons l'évaluation d'un produit dérivé financier avec

whose reward function is time-dependent, unbounded, and presents a discontinuity at maturity. In this context, the regularity conditions required to apply the results and techniques used in the American option literature are not satisfied. We first study the existence of an optimal stopping time and observe that the problem admits a trivial solution under a simple condition on the reward function. When early exercise is optimal, we study the regularity of the value function. We observe that the problem admits another representation with a continuous reward function. This allows us to apply methods similar to those developed for American options to express the price of our derivative in terms of a free boundary value problem. We finally present an application of our results to the valuation of variable annuities with early surrenders.

exercice prématuré ayant une fonction de gain discontinue, non bornée et dépendante du temps. Dans ce contexte, les conditions usuelles pour appliquer les résultats connus sur les options américaines ne sont pas satisfaites. Tout d'abord, nous étudions l'existence d'un temps d'arrêt optimal, et nous démontrons que notre problème admet un temps d'arrêt trivial sous certaines conditions. Lorsque l'exercice prématuré est optimal, nous nous intéressons à la régularité de la fonction valeur. Nous observons que notre problème admet une autre représentation avec une fonction de gain continue. Ce dernier résultat nous permet d'utiliser les méthodes développées dans le contexte d'évaluation d'options américaines afin d'exprimer le prix de notre produit dérivé comme un problème à frontière libre. Finalement, une application de nos résultats aux problèmes liés à l'évaluation de fonds distincts avec rachats anticipés est présentée.

[16:45-17:00]

Emma Kroell (University of Toronto) **Silvana Manuela Pesenti** (University of Toronto) **Sebastian Jaimungal** (University of Toronto)

Optimal robust reinsurance with multiple insurers

Réassurance robuste optimale pour de multiples assureurs

We study a reinsurer who faces multiple sources of model uncertainty. The reinsurer offers contracts to n insurers whose claims follow different compound Poisson processes. As the reinsurer is uncertain about the insurers' claim severity distribution and frequency, they design reinsurance contracts that maximize their utility subject to an entropy penalty. Insurers also seek to maximize their utility, but without ambiguity. We solve this continuous-time Stackelberg game in the cases of proportional reinsurance and excess-of-loss insurance. We obtain explicit solutions when the reinsurer's ambiguity aversion parameter scales with its value function. Moreover, in the case without scaling, we develop an asymptotic expansion of the solution, prove it is correct to the specified accuracy, and find that the optimal solution interpolates between the barycenter of the insurers' models and the scaled entropy solution.

Nous étudions le cas d'un réassureur faisant face à des sources multiples d'incertitude de modèle. Le réassureur offre des contrats à un certain nombre d'assureurs dont les sinistres suivent divers processus de Poisson composés. Comme le réassureur est incertain de la distribution et de la fréquence de la gravité des sinistres des assureurs, il conçoit des contrats de réassurance qui maximisent leur utilité et sont assujettis à une pénalité d'entropie. Les assureurs cherchent aussi à maximiser leur utilité, mais sans ambiguïté. Nous résolvons ce jeu de Stackelberg en temps continu dans les cas de réassurance proportionnelle et d'assurance en excédent de sinistre. Nous obtenons des solutions explicites lorsque le paramètre de l'aversion à l'ambiguïté du réassureur est à l'échelle de sa fonction valeur. De plus, dans un cas sans mise à l'échelle, nous développons une expansion asymptotique de la solution, prouvant qu'elle est correcte quant à la précision spécifiée et montrons que la solution optimale interpole entre le barycentre des modèles d'assureurs et la solution d'entropie à l'échelle.

Chair/Président: Wanhua Su

Room/Salle: TB 238

Date: Tuesday May 30 / mardi 30 mai

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-15:45]

Bethany J. G. White (University of Toronto)

Are Undergraduate Life Science Students Prepared for Quantitative Research?

Les étudiants de premier cycle en sciences de la vie sont-ils préparés à la recherche quantitative ?

Statistical errors are prevalent in life sciences research (Weissgerber et al., 2016), which is contributing to reproducibility concerns in science. This emphasizes the importance of training, yet life science students often only need one statistics course toward their undergraduate programs, if that (Tong et al., 2022). A course that integrates statistics and research design is offered at the University of Toronto to improve students' quantitative preparation. Although we observed improvements in student self-efficacy to choose appropriate statistical procedures and interpret results in a previous study on this course, there were still gaps in their abilities to do both at the end of this course. To extend this work, we conducted a study across several life sciences programs this spring to explore students' abilities to engage with statistics in research. In this talk, I will describe this study, highlight key findings, and discuss implications for life science statistics educators.

Les erreurs statistiques sont fréquentes dans la recherche en sciences de la vie (Weissgerber et al., 2016), ce qui contribue aux problèmes de reproductibilité dans le domaine scientifique. Cela accentue l'importance de la formation, mais les étudiants en sciences de la vie n'ont souvent besoin que d'un seul, voire d'aucun, cours de statistique dans le cadre de leur programme de premier cycle (Tong et al., 2022). L'Université de Toronto propose un cours qui intègre les statistiques et la conception de la recherche pour améliorer la préparation quantitative des étudiants. Bien que nous ayons observé, lors d'une étude précédente de ce cours, une amélioration de la capacité des étudiants à choisir les procédures statistiques appropriées et à interpréter les résultats, il y avait encore des lacunes dans ces capacités à la fin de ce cours. Pour prolonger ce travail, nous avons mené ce printemps une étude sur plusieurs programmes de sciences de la vie afin d'explorer les capacités des étudiants à s'engager avec la recherche statistique. Dans cet exposé, je décrirai cette étude, soulignerai les principaux résultats et discuterai des implications pour les éducateurs en statistique des sciences de la vie.

[15:45-16:00]

Sohee Kang (University of Toronto Scarborough) **Justin Slater** (University of Toronto)

Authoring Peer-Reviewed Questions in Introductory Statistics

Créer des questions évaluées par les pairs en introduction aux statistiques

The essential component for achieving the subject's educational goals is student engagement. To foster engagement, we employed the online technology, Quizzical (https://quizzical.ca) that students create multiple choice questions for the lecture material that they randomly received during the term. The advantage of using Quizzical is that the creation of the question is thought to be the highest level in Bloom's taxonomy (Bloom et al., 1956). Students were also expected to identify the

Le composant essentiel pour réaliser les objectifs éducatifs d'un sujet est l'engagement étudiant. Pour favoriser l'engagement, nous avons employé la technologie en ligne Quizzical (https ://quizzical.ca) avec laquelle les étudiants créent des questions à choix multiples pour le contenu de cours et qu'ils recevront aléatoirement durant la session. L'avantage de l'utilisation de Quizzical est que la création de questions est considérée comme étant le plus haut niveau dans la taxonomie de Bloom (Bloom et al., 1956). On s'attend aussi à ce que les étudiants repèrent

correct answer and explain why it was accurate, as well as explain why the other answer choices were incorrect. In this presentation, we investigate whether the level of Quizzical authoring questions has any positive association with the following formal assessment in a large first year introductory statistics course. We will share analyses results on the engagement level of Quizzical with formal test performance after adjusting for their attitude toward statistics and survey findings.

la bonne réponse, puis expliquent pourquoi elle est correcte et aussi pourquoi les autres réponses sont incorrectes. Dans cette présentation, nous cherchons à savoir si le niveau de création de questions sur Quizzical a une association positive avec la prochaine évaluation formelle dans un grand cours d'introduction aux statistiques de première année. Nous partagerons les résultats d'analyses sur le niveau d'engagement de Quizzical avec la performance de test formel après ajustement pour leur attitude par rapport aux statistiques ainsi que des découvertes de l'enquête.

[16:00-16:15]

Kevin Mongeon (University of Ottawa) **Janique Dubois** (University of Ottawa)

Reconciling The Grade Gap Between Indigenous And Non-Indigenous Students

Réconcilier l'écart de niveau entre les élèves autochtones et non autochtones

This paper utilizes various StatsCan data files to analyze 10 years (i.e., 2010-2020) and millions of elementary and secondary student grades to estimate the discrepancy in Indigenous and non- Indigenous grades and decompose the difference into contributing factors. Based on an educational production approach, the analysis attempts to measure the contribution of socioeconomic factors, academic potential, as well as family, community, and school characteristics to the grade gap. This analysis provides detailed information to inform the development of policies and programs specifically designed to reduce, and eventually eliminate, the grade gap by improving the academic performance of Indigenous students.

Nous utilisons divers fichiers de données de Statistique Canada pour analyser dix années (2010-2020) et des millions de notes d'élèves du primaire et du secondaire. L'objectif ? Estimer l'écart entre les notes des élèves autochtones et celles des élèves non autochtones, et décomposer cet écart en facteurs contributifs. Cette analyse, reposant sur une approche de la production éducative, tente de mesurer la contribution des facteurs socio-économiques, du potentiel scolaire, ainsi que des caractéristiques familiales, communautaires et scolaires à l'écart entre les niveaux d'études. Elle fournit des données détaillées permettant d'élaborer des politiques et des programmes spécialement conçus pour réduire et, à terme, éliminer l'écart entre les niveaux scolaires en améliorant les résultats scolaires des élèves autochtones.

[16:15-16:30]

Suborna Shekhor Ahmed (University of British Columbia) **Chuyi Zheng** (University of British Columbia)

Assessing Interactive Open Access Self-Assessment Tool to Enhance Statistical Learning Capacity

Évaluation d'outil d'auto-évaluation interactif et libre accès pour augmenter les capacités d'apprentissage en statistique

An interactive assessment tool was developed in an introductory statistics course on an open-source platform to provide students an opportunity to build the capacity to apply statistical methodologies using real-life datasets. We prepared worked-out examples with detailed explanations of the procedure that allowed students to apply their learned techniques and self-evaluate themselves. To gain insights into the pedagogical value of the self-assessment tool, we collected surveys to analyze students' perceived self-efficacy, confidence in applying statistical methods, and mastery of content knowledge. In addition, other lessons in similar difficulty levels were used to compare students' perceptions where the tool was absent. We found that problem-solving experiences significantly improved confidence

Nous avons développé un outil d'évaluation interactif dans un cours de statistique à partir d'une plateforme libre-accès afin de fournir aux étudiants une occasion de renforcer leur capacité d'appliquer les méthodologies statistiques au moyen de jeux de données réelles. Nous avons préparé des exemples élaborés et accompagnés d'explications détaillées concernant la procédure qui permettent aux étudiants d'appliquer les techniques apprises et de s'autoévaluer. Pour nous donner un aperçu de la valeur pédagogique de l'outil d'auto-évaluation, nous avons recueilli des enquêtes pour analyser l'auto-efficacité perçue des étudiants, leur confiance pour appliquer les méthodes statistiques et leur maîtrise des connaissances relatives au contenu du cours. De plus, nous avons utilisé d'autres leçons dans des niveaux de difficulté similaires afin de faire une comparaison avec les perceptions des étudiants sans accès à l'outil. Nous avons constaté que

levels where self-assessment tools were available. However, prior knowledge and the year of study factors were associated with gaining mastery and developing confidence in applying methods.

les expériences de résolution de problème améliorent grandement les niveaux de confiance pour ceux ayant eu recours à l'outil. Cependant, des facteurs relatifs aux connaissances antérieures et au nombre d'années d'étude sont associés au gain en maîtrise et en confiance pour appliquer les méthodes.

NSERC – Updates and 2023 Competition Results
Mises à jour du CRSNG et résultats du concours 2023

Chair/Président: W. John Braun

Organizer/Responsable: W. John Braun

Room/Salle: RB 2200

Date: Tuesday May 30 / mardi 30 mai

Time/Heure: 17:00-18:00

Abstract/Résumé

[17:00-18:00]

Adele Ngi-Song (NSERC) **Rachel Desrochers** (Natural Sciences and Engineering Research Council of Canada) **David Haziza** (University of Ottawa)

NSERC – Updates and 2023 Competition Results

Mises à jour du CRSNG et résultats du concours 2023

NSERC – Updates and 2023 Competition Results.

CRSNG – Mise à jour et résultats du concours de 2023.

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

Chair/Président: Erica E. M. Moodie

Organizer/Responsable: Erica E. M. Moodie

Room/Salle: RB 2200

Date: Wednesday May 31 / mercredi 31 mai

Time/Heure: 08:30-09:50

Abstract/Résumé

[08:30-09:50]

Zhou Zhou (University of Toronto)

Time-frequency analysis, a non-stationary time series perspective

L'analyse temps-fréquence, une perspective de séries temporelles non stationnaires

Time-frequency analysis is a branch of signal processing aiming to analyze signals from both time and frequency perspectives simultaneously. In this presentation, I shall provide a non-stationary time series analysis perspective for noisy signals with complex oscillatory patterns so that statistical inferences can be carried out for a class of important problems in time-frequency analysis. Specific problems discussed include power spectra inference under complex temporal dynamics, asymptotic analysis for a class of nonlinear-type time-frequency analysis, and frequency detection and change point estimation for noisy signals of complex oscillation.

L'analyse temps-fréquence est une branche du traitement des signaux qui vise à analyser simultanément les signaux du point de vue du temps et de la fréquence. Dans cette présentation, je fournirai une perspective d'analyse de séries temporelles non stationnaires pour les signaux bruités avec des modèles oscillatoires complexes afin d'effectuer des inférences statistiques pour une classe de problèmes importants en analyse temps-fréquence. Les problèmes spécifiques discutés incluent l'inférence des spectres de puissance dans une dynamique temporelle complexe, l'analyse asymptotique pour une classe d'analyse temps-fréquence de type non linéaire, et la détection de fréquence et l'estimation des points de changement pour les signaux bruités d'oscillation complexe.

Securing Team-Based and Industrial Grants
Obtenir des subventions d'équipe et des subventions industrielles

Chair/Président: Linglong Kong

Organizer/Responsable: Mateen Shaikh

Room/Salle: AT 102

Date: Wednesday May 31 / mercredi 31 mai

Time/Heure: 10:20-11:50

Abstract/Résumé

[10:20-11:50]

Jim V. Zidek (The University of British Columbia) **Celia M. T. Greenwood** (McGill University)

Securing Team-Based and Industrial Grants

Obtenir des subventions d'équipe et des subventions industrielles

The New Investigators Committee provides junior academics with advice on launching their research careers. Many new investigators focus on individual grants and funding, however, there are many opportunities for team-based grants. This session will provide a discussion from several successful applicants of team-based grants and those who have grants partnering with industry. The session is 90 minutes and consists of three panelists, each of which will provide their perspectives and lessons in applying and working on team grants. This will be followed by a 30+ minute question-answer-period which may include additional discussion from the panelists and the audience.

Le Comité des nouveaux chercheurs fournit aux jeunes universitaires des conseils sur le lancement de leur carrière de chercheur. Souvent, les nouveaux chercheurs se concentrent sur les subventions et les financements individuels, alors qu'il existe de nombreuses possibilités de subventions d'équipe. Cette séance sera l'occasion d'une discussion entre plusieurs candidats ayant obtenu des subventions en équipe et ceux qui ont obtenu des subventions en partenariat avec l'industrie. La session dure 90 minutes et se compose de trois panélistes, qui présenteront chacun leur point de vue et les leçons qu'ils ont tirées de leur candidature et de leur travail dans le cadre de subventions d'équipe. Elle sera suivie d'une période de questions-réponses de plus de 30 minutes qui pourra inclure des discussions supplémentaires de la part des panélistes et du public.

Chair/Président: V. Radu Craiu

Organizer/Responsable: V. Radu Craiu

Room/Salle: AT 101

Date: Wednesday May 31 / mercredi 31 mai

Time/Heure: 10:20-11:50

Abstract/Résumé

[10:20-10:50]

Mylène Bédard (Université de Montréal) **Gabriel Boisvert-Beaudry** (Université de Montréal)

On a Generalized, Balanced Version of MALA

À propos d'une version généralisée et équilibrée de MALA

Statistical models have recently been increasing both in terms of complexity and dimensionality. These models cannot be treated analytically; MCMC methods have thus become a device of choice to sample from these target distributions. We introduce a modified Metropolis-adjusted Langevin algorithm (MALA) that features two tuning parameters: the usual step size and an interpolation parameter that accommodates the dimension of the target distribution. We theoretically study the efficiency of this sampler by making use of the local- and global-balance concepts of Zanella (2017). Although the traditional MALA is theoretically optimal in infinite-dimensional settings, in practice, the modified MALA remains superior in all contexts. Simulation studies corroborate our findings. In particular, the efficiency of the modified MALA compares favorably to that of competing algorithms in various Bayesian logistic regression contexts. There is no extra computational cost associated to the new sampler.

Les modèles statistiques ont récemment grandi en complexité et en dimension. Ces modèles ne peuvent pas être traités de façon analytique, donc les méthodes MCMC sont devenues un choix idéal pour l'échantillonnage à partir de ces distributions cibles. Nous présentons un algorithme de Langevin ajusté par Metropolis (MALA) qui comporte deux paramètres de réglage : la taille de pas habituelle et un paramètre d'interpolation qui accommode la dimension de la distribution cible. Nous étudions théoriquement l'efficacité de cet échantillonneur au moyen des concepts d'équilibre global et local de Zanella (2017). Alors que la MALA traditionnelle reste théoriquement optimale dans des cadres de dimensions infinies, en pratique la MALA modifiée lui est supérieure, peu importe la situation. Des études en simulations corroborent nos découvertes. Tout particulièrement, l'efficacité de la MALA modifiée se compare favorablement à celle des algorithmes concurrents dans plusieurs situations de régression logistique bayésienne. Enfin, ce nouvel échantillonneur ne comporte aucun coût informatique supplémentaire.

[10:50-11:20]

Morris Greenberg (University of Toronto) **Radu Craiu** (University of Toronto) **Kieran R Campbell** (University of Toronto)

Restricted Search Space MCMC with Adaptive Weighting and Sparsity Parameterization for Graph Inference

Méthode de Monte-Carlo par chaînes de Markov à espace de recherche restreint avec pondération adaptative et paramétrage de la parcimonie pour l'inférence de graphes

Inferring a directed acyclic graph (DAG) given data is computationally challenging due to the large search space needed to exhaustively consider all possible graphs. Current state-of-the-art MCMC methods for graph inference efficiently scan the space by first considering a restricted search space and iteratively expand

L'inférence d'un graphe acyclique dirigé à partir de données est un défi informatique en raison de l'espace de recherche important nécessaire pour examiner de manière exhaustive tous les graphes possibles. Les méthodes Monte-Carlo par chaînes de Markov actuelles pour l'inférence de graphes balaient efficacement l'espace en considérant d'abord un espace de recherche restreint et en

the space until a stopping criterion is met. Here, we find conditions on iterative changes to the search space that allow for posterior convergence on the unrestricted space, and develop a novel MCMC method that satisfies these conditions. Our algorithm allows for both expansion and constriction of the search space at any iteration, and allows for larger expansion steps than previous methods allow for. Our expansion procedure is dictated by parameterizing the sparsity of the graph (the maximal number of parents observed), and considering how it induces realized parent sizes for individual vertices in the graph, and our constriction procedure is dictated by associating a weight (determined by previous steps in the MCMC chain) to each edge in the space, and only considering edges with large enough weights. We present extensive simulations that characterize the performance and computational efficiency of our algorithm, contrast this with existing methods, and consider applications in the field of imaging proteomics.

élargissant itérativement l'espace jusqu'à ce qu'un critère d'arrêt soit satisfait. Dans cette étude, nous trouvons des conditions sur les changements itératifs de l'espace de recherche qui permettent une convergence postérieure sur l'espace non restreint, et nous élaborons une nouvelle méthode de Monte-Carlo par chaînes de Markov qui remplit ces conditions. Notre algorithme permet à la fois l'expansion et la restriction de l'espace de recherche à chaque itération, ainsi que des étapes d'expansion plus importantes que celles des méthodes précédentes. Notre procédure d'expansion est déterminée par le paramétrage de la parcimonie du graphe (nombre maximal de parents observés) et par la manière dont elle induit les tailles de parents réalisées pour les différents sommets du graphe. Notre procédure de constriction est déterminée par l'association d'un poids (déterminé par les étapes précédentes de la méthode de Monte-Carlo par chaînes de Markov) à chaque arête de l'espace et par la prise en compte des seules arêtes dont le poids est suffisamment élevé. Nous présentons des simulations approfondies qui caractérisent le fonctionnement et l'efficacité informatiques de notre algorithme, les comparons aux méthodes actuelles, puis nous examinons les applications dans le domaine de la protéomique d'imagerie.

[11:20-11:50]

Daniel M. Roy (University of Toronto)

Admissibility is Bayes Optimality with Infintesimals

L'admissibilité est l'optimalité de Bayes avec des infinitésimaux

In his early papers introducing statistical decision theory, Wald already recognized the close relationship between Bayes optimality and frequentist admissibility/minimaxity. One of the cleanest examples is due to Blackwell: in finite problems, admissibility is Bayes optimality with full-support priors. This relationship breaks down beyond finite problems, even in finite-dimensional spaces. For 70 years, people have been working to extend this identity in greater generality. These results—by Le Cam, Stein, Brown, Berger, etc.—have relied on generalized notions of Bayes optimality (via improper priors, etc.) and the extensive use of regularity conditions. In this talk, I'll describe a line of work, rooted in the use of techniques from mathematical logic and nonstandard analysis, where we essentially settle the relationship between admissibility and minimaxity with Bayes optimality, without regularity conditions, by allowing priors to assign infinitesimal mass to events.

Dans ses premiers articles introduisant la théorie de la décision statistique, Wald avait déjà reconnu la relation étroite entre l'optimalité de Bayes et l'admissibilité/minimaxité fréquentiste. L'un des exemples les plus clairs est dû à Blackwell : dans les problèmes finis, l'admissibilité est l'optimalité de Bayes avec des probabilités a priori à support complet. Cette relation s'effondre au-delà des problèmes finis, même dans les espaces à dimension finie. Depuis 70 ans, on s'efforce d'étendre cette identité à une plus grande généralité. Ces résultats, obtenus par Le Cam, Stein, Brown, Berger, etc., s'appuient sur des notions généralisées de l'optimalité de Bayes (par le biais de probabilités a priori impropres, etc.) et sur l'utilisation extensive de conditions de régularité. Dans cet exposé, je décrirai une ligne de travail, enracinée dans l'utilisation de techniques de la logique mathématique et de l'analyse non standard, où nous réglons essentiellement la relation entre l'admissibilité et la minimaxité avec l'optimalité de Bayes, sans conditions de régularité, en permettant aux probabilités a priori d'attribuer une masse infinitésimale aux événements.

Measuring Statistical Evidence
Mesure de la preuve statistique

Chair/Président: Éric P. Marchand

Organizer/Responsable: Éric P. Marchand

Room/Salle: AT 301

Date: Wednesday May 31 / mercredi 31 mai

Time/Heure: 10:20-11:50

Abstract/Résumé

[10:20-10:50]

Michael J. Evans (University of Toronto)

The Role of Frequentism in Bayesian Inference

Le rôle du fréquentisme dans l'inférence bayésienne

Frequentist inference and Bayesian inference are often considered as antagonistic and to a certain extent that is true. The basic idea underlying frequentism is the assessment of the reliability of inferences through the consideration of repeated sampling performance. On the other hand Bayesian inference considers basic logical principles or axioms to justify the methodology. As is discussed here, these two seemingly very different viewpoints can live together and, as such, strengthen our overall confidence that statistical methodology is sound. The frequentist concept of confidence plays a key role in this while a Bayesian formulation permits a clear specification of what is meant by the fundamental concept of statistical evidence which drives the inferences. There is a price for both to pay: frequentists must accept that inference is Bayesian while Bayesians must accept that the reliability of the inferences must be measured and that is frequentist in nature.

L'inférence fréquentiste et l'inférence bayésienne sont souvent considérées comme opposées, ce qui est vrai dans une certaine mesure. L'idée générale à la base du fréquentisme est l'évaluation de la fiabilité des inférences en tenant compte de la performance d'échantillonnage répété. Tandis que l'inférence bayésienne se fonde sur les principes logiques de base ou les axiomes pour justifier la méthodologie. Comme nous l'aborderons ici, ces deux visions très différentes peuvent coexister, et ainsi renforcer notre confiance générale que la méthodologie statistique est solide. Le concept fréquentiste de confiance joue un rôle clé, alors que la formulation bayésienne procure une spécification claire de la compréhension du concept fondamental de la preuve statistique au cœur des inférences. Les deux camps doivent par contre faire un compromis : les fréquentistes doivent accepter que l'inférence soit bayésienne, tandis que les bayésiens doivent accepter de réaliser la tâche de nature fréquentiste qu'est la mesure de la fiabilité des inférences.

[10:50-11:20]

Minge Xie (Rutgers University)

Repro Samples Method for Uncertainty Quantification in Irregular Inference Problems and More

Méthode des échantillons repro pour la quantification de l'incertitude dans les problèmes d'inférence irréguliers et autres

Increasingly more complex and diverse problems in data science demand us to have new inferential frameworks to tackle highly non-trivial irregular inference problems, for instance, those involving discrete or non-numerical parameters and those involving non-numerical data, etc. In this talk, we present a novel, wide-reaching and effective simulation-inspired framework, called repro samples method, to conduct statistical inference and quan-

Les problèmes de plus en plus complexes et diversifiés de la science des données exigent que nous disposions de nouveaux cadres inférentiels pour traiter les problèmes d'inférence irréguliers hautement non triviaux, par exemple, ceux qui impliquent des paramètres discrets ou non numériques et ceux qui impliquent des données non numériques, etc. Dans cet exposé, nous présentons un cadre novateur, étendu et efficace inspiré de la simulation, appelé méthode des échantillons repro, pour ef-

tify uncertainty for the irregular problems plus more. We systemically develop both exact and approximate (asymptotic) theories to support the development. An attractive feature is that the development doesn't need to rely on a likelihood or use the large sample central limit theorem, and thus is especially effective for complicated and irregular inference problems often encountered in machine learning and data science. The effectiveness of the proposed method is illustrated through two case study examples that provide solutions to two open inference problems in the statistics: (a) In a Gaussian mixture model, how to construct a confidence set for the unknown number of components? (b) In a high dimensional regression model, how to construct confidence sets for either the unknown true model, a single or a collection of regression coefficients, or both model and regression coefficients jointly? Comparison studies demonstrate that the proposed methods have far superior performance to existing Bayesian, frequentist, and fiducial attempts. Although the case studies pertain to the traditional statistics models, the framework also has direct extensions to more complex machine learning models involving non-numerical data, such as graphical network, voice data, etc.

fectuer une inférence statistique et quantifier l'incertitude pour les problèmes irréguliers et bien d'autres. Nous développons systématiquement des théories exactes et approximatives (asymptotiques) pour en soutenir le développement. Une caractéristique attrayante est que le développement n'a pas besoin de s'appuyer sur une vraisemblance ou d'utiliser le théorème central limite des grands échantillons, et est donc particulièrement efficace pour les problèmes d'inférence compliqués et irréguliers souvent rencontrés dans l'apprentissage automatique et la science des données. L'efficacité de la méthode proposée est illustrée par deux études de cas qui fournissent des solutions à deux problèmes d'inférence ouverts en statistique : (a) Dans un modèle de mélange gaussien, comment construire un ensemble de confiance pour le nombre inconnu de composants ? (b) Dans un modèle de régression en haute dimension, comment construire des ensembles de confiance soit pour le vrai modèle inconnu, soit pour un seul ou un ensemble de coefficients de régression, soit pour le modèle et les coefficients de régression conjointement ? Des études comparatives démontrent que les méthodes proposées sont bien plus performantes que les tentatives bayésiennes, fréquentistes et fiduciaires existantes. Bien que les études de cas concernent les modèles statistiques traditionnels, le cadre a également des extensions directes à des modèles d'apprentissage automatique plus complexes impliquant des données non numériques, telles que réseaux graphiques, données vocales, etc.

[11:20-11:50]

Éric P. Marchand (Université de Sherbrooke)

On the behaviour of Bayesian credible intervals for some restricted parameter space problems

Sur le comportement d'intervalles de crédibilité bayésiens en présence de contraintes paramétriques

Bayesian credible intervals are not designed to attain a given level of frequentist coverage probability (fcp), but there are number of well-known instances where there is a matching of Bayesian credibility and the fcp. An example arises with a sample from a $N(\mu, \sigma)$ population where the HPD credible set for μ associated prior density $\frac{1}{\sigma}$ and of credibility $1-\alpha$ matches the standard Student-t CI with exact fcp=$1-\alpha$. Such a matching breaks down in presence of a parametric restriction when the original prior density is truncated onto the restricted parameter space. The analysis of the fcp in such situations is thus of interest and we review findings that describe fcp, namely in terms of lower bounds. General findings with wide applicability are presented. The choice of the Bayesian credible set bounds is discussed. Finally, we present contrasting results for compact restrictions, and then analogous results for two ordered normal means.

Même s'il y a plusieurs exemples où la crédibilité bayésienne coïncide avec la probabilité de recouvrement fréquentiste (prf), ce n'est pas le cas en général. Un exemple instructif est le cas d'un échantillon tiré d'une loi N(mu, sigma) avec densité a priori 1/sigma, où il y a correspondance entre les intervalles bayésiens et l'intervalle standard de Student. Cependant, cette correspondance échoue lorsque la densité a priori est tronquée sur un espace paramétrique contraint. L'analyse de la prf est donc d'intérêt dans de telles situations et nous passons en revue des résultats décrivant le comportement de la prf, notamment sous forme de bornes inférieures. Des propriétés générales et applicables pour un grand nombre de situations sont établies. Il s'avère que le choix de bornes de l'intervalle bayésien peut jouer un rôle important. Pour conclure, nous mettons de l'avant des propriétés pour des espaces paramétriques compacts qui font contraste et nous abordons un cas de contrainte d'ordre de deux moyennes.

Mentorship for Statisticians
Mentorat pour statisticiens et statisticiennes

Chair/Président: Thérèse A. Stukel

Organizer/Responsable: Thérèse A. Stukel

Room/Salle: TB 238

Date: Wednesday May 31 / mercredi 31 mai

Time/Heure: 10:20-11:50

Abstract/Résumé

[10:20-11:50]

Thérèse A. Stukel (ICES/ University of Toronto) **Erica E. M. Moodie** (McGill University) **Béatrice Baribeau** (Statistics Canada) **Jennifer McNichol** (University of Victoria) **Kathryn Mills** (Canada Revenue Agency)

Mentorship for Statisticians

Mentorat pour statisticiens

Mentorship for Statisticians.

Mentorat pour statisticiens.

CJS Award Address
Allocution du récipiendaire du Prix de la RCS

Chair/Président: Andrei Volodin

Organizer/Responsable: Andrei Volodin

Room/Salle: RB 2200

Date: Wednesday May 31 / mercredi 31 mai

Time/Heure: 10:20-11:50

Abstract/Résumé

[10:20-11:50]

Evan Sidrow (The University of British Columbia)

Modelling multi-scale, state-switching functional data with hidden Markov models

Modélisation de données fonctionnelles multi-échelles et à changement d'état à l'aide de modèles de Markov cachés

Data sets composed of sequences of curves sampled at high frequencies in time are increasingly common in practice, but they can exhibit complicated dependence structures that cannot be modelled using common methods in functional data analysis. We detail a hierarchical approach that treats the curves as observations from a hidden Markov model. The distribution of each curve is defined by a fine-scale model that may involve autoregression and require data transformations using moving-window summary statistics or Fourier analysis. This approach is broadly applicable to sequences of curves exhibiting intricate dependence structures. As a case study, we use this framework to model the fine-scale kinematic movements of a northern resident killer whale (Orcinus orca) off the west coast of Canada. Through simulations, we show that our model produces more interpretable state estimation and more accurate parameter estimates compared to existing methods. We also consider computational challenges that arise for high-dimensional and big-data scenarios.

Les ensembles de données composés de séquences de courbes échantillonnées à haute fréquence sont de plus en plus courants dans la pratique, mais ils peuvent présenter des structures de dépendance compliquées impossibles à modéliser par les méthodes courantes d'analyse des données fonctionnelles. Nous détaillons une approche hiérarchique qui traite les courbes comme des observations d'un modèle de Markov caché. La distribution de chaque courbe est définie par un modèle à petite échelle qui peut inclure une autorégression et nécessiter des transformations de données à l'aide de statistiques sommaires à fenêtre mobile ou d'une analyse de Fourier. Cette approche est largement applicable aux séquences de courbes présentant des structures de dépendance complexes. Dans une étude de cas, nous utilisons ce cadre pour modéliser les mouvements cinématiques à petite échelle d'un orque résident du Nord (Orcinus orca) au large de la côte ouest du Canada. Grâce à des simulations, nous montrons que notre modèle produit des estimations d'état plus interprétables et des estimations de paramètres plus précises que les méthodes existantes. Nous examinons également les défis informatiques qui se posent pour les scénarios en haute dimension et à données volumineuses.

Chair/Président: Mohamed Belalia

Room/Salle: TB 208

Date: Wednesday May 31 / mercredi 31 mai

Time/Heure: 10:20-11:50

Abstract/Résumé

[10:20-10:35]

Hélène Guérin (Université du Québec à Montréal)

Asymptotic distribution of the Elephant Random Walk

Distribution asymptotique de la marche aléatoire de l'éléphant

The Elephant Random Walk (ERW) has been introduced in 2004 by Schütz and Trimper in order to investigate the long-term memory effects in non-Markovian random walks. Since 2013, this random walk has been studied with different techniques such as Martingale theory, random trees and Polya-type urns. The ERW has three asymptotic regimes. In this talk, we will focus on the superdiffusive regime, where the renormalized ERW converges a.s. to a non-Gaussian random variable. The distribution of this asymptotic variable is unknown. In this talk, we will present new results on this distribution obtained using the different ways to study the ERW. It is interesting to see how the combination of different approaches (martingale, random trees, Polya urn) allows to progress in the knowledge of this asymptotic variable. This is a joint work with Lucile Laulin and Kilian Raschel.

La marche aléatoire de l'éléphant (ERW) a été introduite en 2004 par Schütz et Trimper afin d'étudier les effets de mémoire à long terme dans les marches aléatoires non-markoviennes. Depuis 2013, cette marche aléatoire a été largement étudiée avec différentes techniques telles que la théorie de Martingale, les arbres aléatoires et les urnes type urnee de Polya. L'ERW a trois régimes asymptotiques. Dans cet exposé, nous nous concentrerons sur le régime superdiffusif, où la marche de l'éléphant renormalisée converge p.s. vers une variable aléatoire non-gaussienne. La distribution de cette variable asymptotique est inconnue. Dans cet exposé, nous présenterons de nouveaux résultats sur cette distribution obtenus en utilisant les différentes manières d'étudier cette marche aléatoire (martingale, arbres aléatoires, urne de Polya). Il s'agit d'un travail en collaboration avec Lucile Laulin et Kilian Raschel.

[10:35-10:50]

Camila P. E. de Souza (University of Western Ontario) **Pedro H. T. O. Souza** (Universidade de Campinas) **Ronaldo Dias** (Universidade de Campinas)

Bayesian adaptive selection of basis functions for functional data representation

Sélection adaptative bayésienne de fonctions de base pour la représentation de données fonctionnelles

We developed and applied a new Bayesian approach via the Gibbs sampler to select basis functions for a finite representation of functional data. The proposed methodology uses Bernoulli latent variables to assign zero to some of the basis function coefficients with a positive probability. This procedure allows for an adaptive basis selection since it can determine the number of bases and which ones should be selected to represent functional data. Moreover, the proposed procedure measures the uncertainty of the selection process and can

Nous avons développé et appliqué une nouvelle approche bayésienne par échantillonneur de Gibbs afin de sélectionner les fonctions de base pour une représentation finie de données fonctionnelles. La méthodologie proposée fait appel aux variables latentes de Bernoulli pour attribuer zéro aux coefficients de certaines fonctions de base avec une probabilité positive. Cette procédure permet une sélection de base adaptative puisqu'elle peut déterminer le nombre de bases et lesquelles doivent être sélectionnées pour représenter les données fonctionnelles. De plus, la procédure proposée mesure l'incertitude du processus de sélection et peut être appliquée

be applied to multiple curves simultaneously. We apply the proposed methodology to a real dataset involving daily numbers of COVID-19 cases in Brazil. Simulation studies show the main properties of the proposed method, such as its accuracy in estimating the coefficients and the strength of the procedure to find the true set of basis functions. We also compared the proposed model via simulation with the well-established LASSO and Bayesian LASSO.

simultanément à de multiples courbes. La méthodologie proposée est appliquée à un ensemble de données réelles recensant les nombres quotidiens de cas de COVID-19 au Brésil. Des études de simulation montrent les principales propriétés de la méthode proposée, comme son exactitude pour l'estimation des coefficients et la force de la procédure pour trouver le vrai ensemble de fonctions de base. Nous avons aussi comparé, à l'aide de simulations, le modèle proposé au LASSO et au LASSO bayésien qui sont deux méthodes bien établies.

[10:50-11:05]

Klaus Peter Herrmann (Université de Sherbrooke) **Marius Hofert** (The University of Hong Kong) **Johanna G. Nešlehová** (McGill University)

On a Class of Distortions that Transform GEV Distributions into GEV Distributions

D'une classe de distorsions qui transforment les distributions GEV en distributions GEV

Non-degenerate weak limits of maxima for a sequence of iid random variables belong necessarily to the class of generalized extreme value (GEV) distributions by the Fisher-Tippett-Gnedenko theorem. When considering dependent sequences, the resulting limiting distributions can often be expressed as distortions of the associated iid limit. A prominent example are power distortions that arise in the case of weakly dependent time series. Noting that power distortions of GEV distributions stay in the GEV class raises the question of which general class of distortions transform GEV distributions into GEV distributions. In this talk we answer this question by establishing and solving a related functional equation. We discuss properties of the solution and connect our result to the weak limiting behavior of maxima for different dependence models.

Les limites faibles non dégénérées des maxima pour une séquence de variables aléatoires iid appartiennent nécessairement à la classe des distributions de valeurs extrêmes généralisées (GEV) en vertu du théorème de Fisher-Tippett-Gnedenko. Lorsque l'on considère des séquences dépendantes, les distributions limites résultantes peuvent souvent être exprimées comme des distorsions de la limite iid associée. Les distorsions de puissance qui apparaissent dans le cas de séries temporelles faiblement dépendantes en sont un exemple frappant. Le fait que les distorsions de puissance de distributions GEV restent dans la classe GEV soulève la question de savoir quelle classe générale de distorsions transforme les distributions GEV en distributions GEV. Dans cet exposé, nous répondons à cette question en établissant et en résolvant une équation fonctionnelle connexe. Nous discutons des propriétés de la solution et établissons un lien entre notre résultat et le comportement limite faible des maxima pour différents modèles de dépendance.

[11:05-11:20]

Serge B. Provost (The University of Western Ontario)

Certain Results Pertaining to Distributional Moments

Certains résultats relatifs aux moments de distribution

Based on a theorem that states that a sample can be recovered from its moments, a technique for determining the relative information contained in each moment of a distribution is proposed. On applying this result, one could, for instance, select the appropriate number of moments to be utilized when applying moment-based density estimation methodologies such as those that have been advocated by this author. Illustrative numerical examples will be presented. Additionally, it will be explained that a set of multivariate observations can be exactly retrieved from its marginal sample moments and associated empirical copula. Time permitting, cer-

Nous proposons une technique pour déterminer l'information relative contenue dans chaque moment d'une distribution, en se fondant sur un théorème qui dicte qu'un échantillon peut être récupéré à partir de ses moments. En appliquant ce résultat, on pourrait, par exemple, choisir le nombre adéquat de moments à utiliser lorsqu'il faut appliquer des méthodologies d'estimation de densité à base de moments comme celles que nous préconisons. Nous présenterons des exemples numériques illustratifs. En outre, nous expliquerons qu'un ensemble d'observations multivariées peut être récupéré précisément à partir de ses moments échantillonnaux marginaux et de ses copules empiriques associées. Si le temps le permet, certains résultats distributionnels impliquant des estimations de

tain distributional results involving copula density estimates and their joint moments shall also be presented.

densité de copule et leurs moments conjoints seront également présentés.

[11:20-11:35]

Zilin Wang (Wilfrid Laurier University) **Mary Thompson** (University of Waterloo)

Modelling Missing-Not-At-Random in Complex Surveys

Modélisation de données omises de façon non aléatoire dans des enquêtes complexes

In the analysis of incomplete data in complex surveys, the probability of being non-missing (propensity score) could depend on the study variable itself, the so-called missing-not-at-random (MNAR) mechanism. In this paper, we model the propensity score under MNAR when it is not homogeneous across the support of the study variable. Conditional on values being in subsets of support of the study variable, the proportions of non-missingness are used to model the propensity scores. Using a weighted likelihood function and survey estimating equation methods, we estimate the parameters of the distribution of the study variable and the propensity score within each subset. Simulation studies show the method provides more accurate estimators of the study variable parameters than when missingness is ignored; estimation of the propensity scores can be improved with prior information on the non-homogeneity of support of the study variable. We apply this method in large scale longitudinal surveys.

Dans l'analyse de données incomplètes dans des enquêtes complexes, la probabilité que les données soient non manquantes (score de propension) pourrait dépendre de la variable à l'étude elle-même, on parle alors d'un mécanisme d'omissions non aléatoire des données (MNAR). Nous modélisons le score de propension sous l'hypothèse MNAR lorsqu'il n'est pas homogène dans le support de la variable à l'étude. Conditionnellement à ce que les valeurs soient dans les sous-ensembles du support de la variable à l'étude, les proportions de données non manquantes sont utilisées pour modéliser les scores de propension. À l'aide d'une fonction de vraisemblance pondérée et des méthodes d'équations estimantes de l'enquête, nous estimons les paramètres de distribution des variables de l'étude et le score de propension dans chaque sous-ensemble. Des études par simulation montrent que la méthode fournit des estimateurs plus exacts des paramètres de la variable à l'étude lorsque les données manquantes sont ignorées ; l'estimation des scores de propension peut être améliorée avec de l'information préalable sur la non-homogénéité du support de la variable à l'étude. Nous appliquons la méthode à des enquêtes longitudinales à grande échelle.

[11:35-11:50]

Bailey Trenton Janeczko (University of Manitoba) **Yuliya V. Martsynyuk** (University of Manitoba)

Functional CLT Based Confidence Intervals for the Slope in Linear Structural Errors-in-Variables Models

Intervalles de confiance basés sur le théorème central limite fonctionnel (FCLT) pour la pente dans les modèles structurels linéaires d'erreurs de variables

Asymptotic normality of modified least squares estimators (MLSE) yields confidence intervals (CI) for the slope in linear structural errors-in-variables models (LSEIVM) with univariate observations. More asymptotic CI's can be obtained from a Functional Central Limit Theorem (FCLT) for modified least squares processes (MLSP) for the slope, which is convergence in distribution of appropriate functionals of MLSP's to the same functionals of a standard Wiener process. We study finite-sample performance of asymptotic CI's for the slope that are based on MLSE's and MLSP's. Evaluation and comparison of finite-sample properties of these CI's are done via their empirical coverage probabilities and expected lengths for a variety of LSEIVM's. We make a connection between our study and the one

La normalité asymptotique des estimateurs des moindres carrés modifiés (MLSE) fournit des intervalles de confiance (CI) pour la pente dans les modèles structurels linéaires d'erreurs de variables (LSEIVM) avec observations multivariées. Avec le théorème central limite fonctionnel (FCLT), qui donne la convergence en distribution des fonctionnelles appropriées des MLSP vers les mêmes fonctionnelles que dans un processus standard de Wiener, on peut obtenir un plus grand nombre d'intervalles de confiance asymptotiques pour des processus des moindres carrés modifiés (MLSP) pour la pente. Nous étudions la performance pour des échantillons de taille finie des CI asymptotiques pour la pente qui sont basés sur les MLSE et les MLSP. L'évaluation et la comparaison des propriétés pour des échantillons de taille finie de ces CI sont faites au moyen de leur probabilité de couverture empirique et de leur longueur attendue pour divers LSEIVM. Nous établissons un lien

from the literature on FCLT based CI's for the mean of i.i.d. random variables, where for some distributions such CI's were shown to be desirable alternatives to a classical CI obtained simply from the CLT.

entre notre étude et celle dans la documentation sur les CI basés sur le FCLT pour la moyenne de variables aléatoires indépendantes et identiquement distribuées (i.i.d.) dans un contexte où on a montré que pour certaines distributions de tels CI sont des solutions de rechange désirables aux CI classiques obtenus simplement du théorème central limite.

Chair/Président: Nan Zheng

Room/Salle: TB 210

Date: Wednesday May 31 / mercredi 31 mai

Time/Heure: 10:20-11:50

Abstract/Résumé

[10:20-10:35]

Simon J. Bonner (Western University) **Braedan Walker** (Western University) **Ron Bonner**

Incorporating Isotope Patterns in the Analysis of Mass-Spectrometry Data

Intégration des motifs isotopiques dans l'analyse des données de spectrométrie de masse

Mass spectrometry (MS) is a chemical technique that identifies and quantifies chemicals in a mixture based on their molecular masses. MS is complicated because different molecules may have indistinguishable masses. Moreover, natural isotopes produce forms of each molecule with different masses. We have developed a method to improve the analysis of MS data by incorporating the isotope patterns of candidate molecules directly into models of the counts within an observed spectrum. Using a Bayesian implementation of a Poisson generalized linear model with an identity link, we implemented spike-and-slab priors to determine the most likely combination of candidates in a mixture while ensuring that the fitted means are non-negative. Further analysis allows the signals of unknown chemicals that might be present to be extracted. We present an example of our approach to quantify the amounts of different lipids within a human sample.

La spectrométrie de masse (MS) est une technique chimique qui permet d'identifier et de quantifier les substances chimiques d'un mélange en fonction de leur masse moléculaire. La spectrométrie de masse est compliquée car différentes molécules peuvent avoir des masses impossibles à distinguer. De plus, les isotopes naturels produisent des formes de chaque molécule avec des masses différentes. Nous avons développé une méthode pour améliorer l'analyse des données MS en incorporant les schémas isotopiques des molécules candidates directement dans les modèles de comptage au sein d'un spectre observé. En utilisant une implémentation bayésienne d'un modèle linéaire généralisé de Poisson avec un lien d'identité, nous avons mis en œuvre des priorités de type « spike-and-slab » pour déterminer la combinaison la plus probable de candidats dans un mélange tout en garantissant que les moyennes ajustées ne sont pas négatives. Une analyse plus poussée permet d'extraire les signaux de produits chimiques inconnus qui pourraient être présents. Nous présentons un exemple de notre approche pour quantifier les quantités de différents lipides dans un échantillon humain.

[10:35-10:50]

Hao He (University of Ottawa) **David Haziza** (University of Ottawa) **Song Cai** (Carleton University)

Efficient and Robust Empirical Likelihood Approach for Estimating Equations with Missing Data

Approche efficace et robuste de la vraisemblance empirique pour l'estimation d'équations avec données manquantes

Empirical likelihood (EL) is a flexible nonparametric method for various inference problems that does not require the specification of data distribution. We propose an EL approach when data are missing at random. This approach employs a kernel-based estimator of the condition expectation of the estimating function, given the fully observed variables for the units with missing values. We show that the maximum empirical likelihood estimator based on the proposed method is doubly ro-

La vraisemblance empirique (EL) est une méthode non paramétrique flexible pour divers problèmes d'inférence qui ne nécessite pas de spécifier la distribution des données. Nous proposons une approche de la vraisemblance empirique lorsque certaines données manquent au hasard. Cette approche utilise un estimateur à base de noyau de l'espérance conditionnelle de la fonction d'estimation, qui tient compte des variables entièrement observées pour les unités ayant des valeurs manquantes. Nous montrons que l'estimateur du maximum de vraisemblance empirique

bust and efficient in the sense that it attains the semiparametric efficiency bound. Moreover, the log-EL ratio of the proposed approach is asymptotically chi-square distributed when evaluated at true parameter values. We present the results from a simulation study to evaluate the performance of the proposed method in terms of bias and efficiency.

basé sur la méthode proposée est doublement robuste et efficace dans le sens où il atteint la limite d'efficacité semi-paramétrique. En outre, le ratio log-EL de l'approche proposée est asymptotiquement distribué selon une distribution chi-carré lorsqu'il est évalué aux vraies valeurs des paramètres. Nous présentons les résultats d'une étude de simulation pour évaluer la performance de la méthode proposée en termes de biais et d'efficacité.

[10:50-11:05]

Shaun McDonald (Simon Fraser University) **Dave Campbell** (Carleton University)

FRODO: A Novel Approach to Micro-Macro Multilevel Regression

FRODO : Une nouvelle approche de la régression micro-macro multiniveaux

Within the field of hierarchical modelling, little attention is paid to micro-macro models: those in which group-level outcomes are dependent on covariates measured at the level of individuals within groups. Although such models are perhaps underrepresented in the literature, they have applications in economics, epidemiology, and the social sciences. Here, we present a new empirical Bayesian technique for fitting such models called FRODO. The method jointly infers group-specific densities for multilevel covariates and uses them as functional predictors in a functional linear model. The power and versatility of FRODO are demonstrated on several simulated datasets, showcasing its ability to accommodate a wide variety of covariate distributions and regression models.

Dans le domaine de la modélisation hiérarchique, peu d'attention est accordée aux modèles micro-macro : ceux dans lesquels les résultats au niveau du groupe dépendent de covariables mesurées au niveau des individus au sein des groupes. Bien que de tels modèles soient peut-être sous-représentés dans la littérature, ils ont des applications en économie, en épidémiologie et en sciences sociales. Ici, nous présentons une nouvelle technique bayésienne empirique pour ajuster de tels modèles appelée FRODO. La méthode déduit conjointement les densités spécifiques au groupe pour les covariables à plusieurs niveaux et les utilise comme prédicteurs fonctionnels dans un modèle linéaire fonctionnel. La puissance et la polyvalence de FRODO sont démontrées sur plusieurs ensembles de données simulées, démontrant sa capacité à s'adapter à une grande variété de distributions de covariables et de modèles de régression.

[11:05-11:20]

Jonathan Jalbert (Polytechnique Montreal) **Gabriel Gobeil** (Polytechnique Montréal) **Philippe Roy** (Institut de recherche d'Hydro-Québec)

Post-processing the bulk and the tail of simulated precipitations

Post-traitement du coeur et de la queue des précipitations simulées

Intense precipitations are likely to increase in intensity and frequency in Canada. The study of the impacts these changes is generally done using simulated precipitation from climate models. Those simulated precipitations are generally biased and must be post-process before using them in impact studies. The common post-processing method modify the distribution of simulated precipitations with respect to the observed precipitations. The empirical cumulative distribution function (cdf) is often used but the accuracy of these estimates decreases in the tail of the distribution, and it is not suitable for the post-processing of extreme values. Instead of the empirical cdf, we propose to use the extended generalized Pareto distribution (EGPD) for post-processing simulated precipitations. The EGPD is able

Les précipitations intenses sont susceptibles d'augmenter en intensité et en fréquence au Canada. L'étude des impacts de ces changements se fait généralement à partir de précipitations simulées par des modèles climatiques. Ces dernières sont généralement biaisées et doivent être post-traitées avant d'être utilisées dans les études d'impact. La méthode de post-traitement la plus courante consiste à modifier la distribution des précipitations simulées par rapport aux précipitations observées. La fonction de répartition (cdf) empirique est souvent utilisée mais la précision de cette estimation diminue dans la queue de la distribution. Elle n'est pas donc pas adaptée au post-traitement des valeurs extrêmes. Nous proposons d'utiliser la distribution de Pareto généralisée étendue (EGPD) pour le post-traitement des précipitations simulées. L'EGPD permet de modéliser le coeur et la queue de façon cohérente avec la théorie des valeurs extrêmes.

to model the bulk and the tail consistently with respect to the extreme value theory. The method is applied to post-process simulated precipitations of several Canadian cities.

[11:20-11:35]

Jeffrey D. Picka (University of New Brunswick)

Is scientific realism compatible with statistical inference?

Le réalisme scientifique est-il compatible avec l'inférence statistique ?

Antirealism is a core aspect of modern applied statistics, from its development by Karl Pearson through George Box's statement that all models are wrong, but some models are useful. The importance of taking antirealistic stances about scientific knowledge and the costs of ignoring implicitly realist stances will be discussed in the context of uncertainty quantification, statistical education, and the current controversy regarding the best use of the p-value.

L'antiréalisme est un aspect essentiel de la statistique appliquée moderne, depuis son développement par Karl Pearson jusqu'à la déclaration de George Box selon laquelle tous les modèles sont faux, mais certains modèles sont utiles. L'importance d'adopter des positions antiréalistes sur la connaissance scientifique et les conséquences de l'ignorance des positions implicitement réalistes seront discutées dans le contexte de la quantification de l'incertitude, de l'éducation statistique et de la controverse actuelle concernant la meilleure utilisation de la valeur-p.

[11:35-11:50]

Paritosh Kumar Roy (McGill University) **Alexandra Schmidt** (McGill University) **Thaís C. O. Fonseca** (Federal University of Rio de Janeiro, Brazil)

Dynamical Non-Gaussian Modeling of Multivariate Spatial Processes

Modélisation dynamique non gaussienne de processus spatiaux multivariés

Commonly environmental monitoring networks are designed to measure various processes simultaneously, and data analysis postulates a multivariate model to capture their association. Observed data often suggest that these processes do not follow a normal distribution. We extend the well-known multivariate dynamic linear model to accommodate heavy-tailed distributions. The proposed model assumes a scale mixture that follows a linear model of coregionalization. We discuss the properties of the resultant cross covariance function. Inference is performed following the Bayesian paradigm, and the MCMC algorithm is implemented in Stan. Artificial data analysis suggests we can recover the parameters used to generate the data. Analysis of environmental processes observed across the UK illustrates the effectiveness of the proposed model in quantifying prediction uncertainty.

Les réseaux de surveillance de l'environnement sont généralement conçus pour mesurer simultanément différents processus, et l'analyse des données postule un modèle multivarié pour saisir leur association. Or les données observées suggèrent souvent que ces processus ne suivent pas une distribution normale. Nous étendons le modèle linéaire dynamique multivarié bien connu pour tenir compte des distributions à queue lourde. Le modèle proposé suppose un mélange d'échelles qui suit un modèle linéaire de corégionalisation. Nous discutons des propriétés de la fonction de covariance croisée résultante. L'inférence est effectuée selon le paradigme bayésien et l'algorithme MCMC est implémenté dans Stan. Une analyse de données artificielles suggère que nous pouvons récupérer les paramètres utilisés pour générer les données. Via une analyse des processus environnementaux observés au Royaume-Uni, nous illustrons l'efficacité du modèle proposé pour quantifier l'incertitude des prévisions.

Chair/Président: Suborna Shekhor Ahmed

Room/Salle: TB 240

Date: Wednesday May 31 / mercredi 31 mai

Time/Heure: 10:20-11:50

Abstract/Résumé

[10:20-10:35]

Liyuan Zheng (University of Waterloo) **Yilei Wu** (Pinterest) **Yingli Qin** (University of Waterloo) **Mu Zhu** (University of Waterloo)

Quadratic Discriminant Analysis for High-Dimensional Data

Analyse discriminante quadratique pour les données de grande dimension

In this talk, we consider Quadratic Discriminant Analysis (QDA) and joint estimation of precision matrices for high-dimensional data. Our data-driven QDA rule incorporates precision matrices which are jointly estimated under the 'Partially Common Diagonal + Low-rank' structural assumption. We provide an explicit convergence rate for the classification error of our proposed QDA rule in a similar spirit to that in Cai and Zhang (2021) but without the sparsity assumption on the difference between two precision matrices. Regarding numerical studies, we first demonstrate that the proposed QDA rule performs well, compared to other high-dimensional QDA rules, in both sparse and non-sparse cases; we then present real-data analyses based on some micro-array data sets.

Lors de cet exposé, nous examinons l'analyse discriminante quadratique (QDA) et l'estimation conjointe de matrices de précision pour des données de grande dimension. Notre cadre de QDA basée sur les données comporte des matrices de précision qui sont conjointement estimées sous l'hypothèse structurelle ≪diagonale partiellement commune et faible rang≫. Nous fournissons un taux de convergence explicite pour l'erreur de classification de la QDA proposée de manière similaire à celui de Cai et Zhang (2021), mais sans hypothèse d'éparsité dans la différence entre les deux matrices de précision. En ce qui concerne les études numériques, nous démontrons d'abord que la QDA proposée fonctionne bien dans les cas épars et non épars par rapport aux autres QDA de grande dimension. Nous présentons ensuite des analyses de données réelles fondées sur des ensembles de données de microréseaux.

[10:35-10:50]

Nicholas Mitsakakis (Children's Hospital of Eastern Ontario Research Institute) **Samer El Kababji** (Children's Hospital of Eastern Ontario Research Institute) **Khaled El Emam** (Children's Hospital of Eastern Ontario Research Institute)

Determining Adequate Sample Size for Studies Using Machine Learning

Détermination de la taille adéquate de l'échantillon pour les études utilisant l'apprentissage automatique

Studies that employ traditional statistical methods can rely on established techniques and guidelines for determining adequate sample size for achieving sufficient power, precision, or developing accurate prediction models. However, there are no well-developed methods for determining sufficient sample size for machine learning (ML) applications. Previous simulation studies suffer by simplistic data generation mechanisms, unrealistic assumptions, lack of tuning, limited range of sample sizes and performance measures used, among others. We address this gap, using a comprehensive

Les études qui utilisent des méthodes statistiques traditionnelles peuvent s'appuyer sur des techniques et des lignes directrices bien établies pour déterminer une taille d'échantillon qui permette d'obtenir une puissance et une précision suffisantes ou de développer des modèles de prédiction précis. Cependant, il n'existe pas de méthodes bien développées pour déterminer cette taille d'échantillon pour les applications d'apprentissage automatique (AA). Les études de simulation précédentes souffrent, entre autres, de mécanismes de génération de données simplistes, d'hypothèses irréalistes, d'un manque de réglage, d'une gamme limitée de tailles d'échantillons et de mesures de performance uti-

simulation design focusing on binary classification. We extend a published methodology for simulating benchmark population datasets, and use varying sized samples, nested cross validation and two performance measures (AUC, Brier score) to train, tune and evaluate models of three ML methods (random forests, neural nets, SVM). Results can guide sample size determination and the design of studies using ML methods.

lisées. Nous comblons cette lacune en utilisant un modèle de simulation complet axé sur la classification binaire. Nous étendons une méthodologie publiée à la simulation d'ensembles de données sur une population de référence et utilisons des échantillons de taille variable, une validation croisée imbriquée et deux mesures de performance (AUC, score de Brier) pour former, mettre au point et évaluer des modèles de trois méthodes d'AA (forêts aléatoires, réseaux neuronaux, SVM). Les résultats peuvent guider la détermination de la taille de l'échantillon et la conception d'études utilisant des méthodes d'AA.

[10:50-11:05]

Steve Ferreira Guerra (McGill University) **Steve Ferreira Guerra** (McGill University) **Robert Platt** (McGill University) **Michal Abrahamowicz** (McGill University)

SIMEX Adaptation to Correct for Misclassification in Binary Time-Varying Prescription-Based Exposures

SIMEX adapté pour corriger les erreurs de classification dans des expositions binaires basées sur des ordonnances et qui varient dans le temps

Prescription claims have been increasingly used in drug safety and effectiveness studies to determine individual drug exposures. However, exposures based on prescriptions do not represent actual drug intake. This is known as misclassification which results in biased inference in naive analyses. We developed methods specifically tailored to correct for misclassification in the analyses of time-varying prescription-based exposures. The approach relies on a pragmatic adaptation of the misclassification simulation-extrapolation (MC-SIMEX) method. Simulation results indicate a substantial reduction bias, and this for various exposure metrics that might be of interest in different applications. In conclusion, accounting for misclassification in time-varying prescription-based exposures is challenging, but crucial to avoid biased analyses. Pragmatic methods, such as the one proposed, are a promising avenue to more appropriately account for complex types of misclassification.

Les demandes de remboursement d'ordonnances sont de plus en plus utilisées dans les études de sécurité et d'efficacité des médicaments pour déterminer l'exposition individuelle aux médicaments. Cependant, les expositions basées sur des ordonnances ne représentent pas la prise réelle de médicaments. C'est ce que l'on appelle une erreur de classification, qui entraîne une inférence biaisée dans les analyses naïves. Nous avons développé des méthodes spécialement conçues pour corriger ces erreurs de classification dans les analyses des expositions basées sur des ordonnances et qui varient dans le temps. L'approche repose sur une adaptation pragmatique de la méthode de simulation-extrapolation des erreurs de classification (MC-SIMEX). Les résultats de la simulation indiquent une réduction substantielle du biais, et ce pour diverses mesures d'exposition qui pourraient être utiles dans différentes applications. En conclusion, la prise en compte des erreurs de classification dans les expositions basées sur des ordonnances et qui varient dans le temps est un défi, mais elle est cruciale pour éviter les analyses biaisées. Les méthodes pragmatiques, telles que celle proposée, constituent une voie prometteuse pour prendre en compte de manière plus appropriée les types complexes d'erreurs de classification.

[11:05-11:20]

Yanglei Song (Queen's University) **Na Li** (Queen's University) **Chunfang Devon Lin** (Queen's University) **Dongsheng Tu** (Queen's University)

Bootstrap Adjusted Predictive Classification under Generalized Linear Models

Classification prédictive ajustée par bootstrap selon des modèles linéaires généralisés

Predictive classification concerns identifying subgroups based on a continuous biomarker through the estimation of an unknown cutpoint and assessing whether these subgroups differ in treatment effect. The problem is

La classification prédictive consiste à définir des sous-groupes en se basant sur un biomarqueur continu à travers l'estimation d'un point d'interruption inconnu, et à évaluer si ces sous-groupes ont des effets de traitement différents. Nous abordons le problème se-

considered under a generalized linear model framework for clinical outcomes and formulated as testing the significance of the interaction between the treatment and the subgroup indicator. When the main effect does not exist, the cutpoint is non-identifiable under the null. We propose profile score-type test statistics, and further m-out-of-n bootstrap techniques to obtain their critical values. The proposed procedures do not rely on the knowledge about the model identifiability, and we establish their asymptotic size validity and study the power under local alternatives in both cases. Further, we show the inconsistency of the standard bootstrap for the non-identifiable case. The proposed method is applied to a dataset from a clinical trial on advanced colorectal cancer.

lon un cadre de modèle linéaire généralisé pour des résultats cliniques, puis le formulons pour tester la pertinence de l'interaction entre le traitement et l'indicateur de sous-groupe. Lorsque l'effet principal n'existe pas, le point d'interruption est non identifiable. Pour obtenir ces valeurs cruciales, nous proposons donc des variables à tester de type score de profil, et des techniques par bootstrap ≪m-out-of-n≫. Les procédures proposées ne dépendent pas des connaissances relatives aux identifications de modèle. De plus, nous prouvons leur validité de taille asymptotique et étudions la puissance selon des alternatives locales dans les deux cas. En outre, nous démontrons l'incohérence de la méthode bootstrap standard dans le cas d'absence d'identification. La méthode proposée est appliquée à un jeu de données tiré d'un essai clinique sur le cancer colorectal avancé.

[11:20-11:35]

Qingrun Zhang (University of Calgary) **Jingni He** (University of Calgary) **Qing Li** (University of Calgary) **Qingrun Zhang** (University of Calgary)

rvTWAS: identifying rare variants underlying complex traits using a data-bridge model

rvTWAS : identification de variants rares sous-jacents à des traits complexes au moyen d'un modèle de pont de données

Transcriptome-wide association study (TWAS) has been successful in identifying genetic basis of complex diseases. In a typical TWAS framework, one first trains a gene expression predictive model using genotype; and then in a genotype-disease cohort, one predicts the gene expression and associates the predicted with phenotype. However, rare variants (defined as their population minor allele frequency ¡ 0.5%), which contain ¿90% sites of genetic diversity and proven to be functional critical, cannot be included, due to the rare variants' inability to predict. Herein, based upon our previous work disentangling feature selection and feature aggregation in TWAS, we propose rvTWAS, a protocol that uses a Bayesian feature selection to form a set of rare variants, which are in turn aggregated using a kernel machine to be associated to diseases. We showed the superior performance of rvTWAS in both simulated and real data and discovered additional genes underlying multiple complex diseases.

L'étude d'association à l'échelle du transcriptome (TWAS) a réussi à identifier les bases génétiques de maladies complexes. Dans un cadre de TWAS général, on doit d'abord former un modèle prédictif d'expression génique, puis, dans une cohorte génotype-maladie, on doit prédire l'expression génique et associer la prévision avec le phénotype. Cependant, les variants rares (définis par la fréquence allélique mineure inférieure à 0,5 % dans leur population), qui contiennent plus de 90 % des sites de diversité génétique et qui sont prouvés comme étant fonctionnellement cruciaux, ne peuvent pas être inclus en raison de leur incapacité à faire des prédictions. Conséquemment, en nous fondant sur notre travail antérieur ayant séparé la sélection de variables et l'agrégation de variables dans un TWAS, nous proposons rvTWAS : un protocole qui emploie une sélection de variables bayésienne pour former un ensemble de variants rares, qui sont ensuite agrégés au moyen d'une machine à noyaux afin de les associer aux maladies. Nous avons démontré la performance supérieure de la rvTWAS avec des données réelles et simulées, et avons découvert des gènes supplémentaires à la base de plusieurs maladies complexes.

The modeling approaches for the proportional/count data with extra variations
Approches de modélisation pour les données proportionnelles/de comptage avec variations supplémentaires

Chair/Président: Dianliang Deng

Organizer/Responsable: Dianliang Deng

Room/Salle: AT 301

Date: Wednesday May 31 / mercredi 31 mai

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-14:00]

Cindy Feng (Dalhousie University)

Zero-Inflated Models for Adjusting Varying Exposures: A Cautionary Note on the Pitfalls of Using Offset

Modèles à surreprésentation de zéros pour l'ajustement d'expositions variant avec le temps ; mise en garde contre les pièges d'une utilisation d'un "offset"

Zero-inflated count data are frequently encountered. Two-parts model is often used to model excessive zeros, which are a mixture of two components. When the rate of events per unit exposure is of interest, offset is commonly used to account for the varying extent of exposure, which is essentially a predictor whose regression coefficient is fixed at one. Such an assumption of exposure effect is, however, quite restrictive. Further, for zero-inflated models, offset is often only included in the count component of the model. However, the probability of the excessive zero component could also be affected by the amount of "exposure". We proposed incorporating the varying exposure as a covariate rather than an offset term in both the probability of excessive zeros and conditional counts components of the zero-inflated model. A real example is used to illustrate the usage of the proposed method, and simulation studies are conducted to assess the performance of the proposed methods.

Les données de comptage avec une surreprésentation de zéros sont fréquentes. Un modèle en deux parties est souvent utilisé pour la modélisation d'un excès de zéros qui est un mélange de deux composantes. Quand le taux d'événements par exposition d'unité est d'intérêt, un "offset" est couramment utilisé pour prendre en compte de l'étendue de la variable d'exposition qui est essentiellement un prédicteur dont le coefficient de régression est fixe à un. Une telle hypothèse d'effet d'exposition est toutefois plutôt restrictive. De plus, dans les modèles à surreprésentation de zéros, le "offset" est souvent compris dans la composante de comptage du modèle. Cependant, la probabilité d'une composante à surreprésentation de zéros peut aussi être affectée par la quantité d'« exposition ». Nous proposons d'incorporer l'exposition variant avec le temps comme une covariable plutôt que comme un terme "offset" à la fois dans la probabilité d'un excès de zéros et dans les composantes de comptage conditionnelles du modèle à surreprésentation de zéros. Un exemple réel est utilisé pour illustrer l'usage de la méthode proposée et des études de simulation sont menées pour évaluer la performance des méthodes proposées.

[14:00-14:30]

Guohua Yan (University of New Brunswick) **Xiaolei Zhang** (Yunnan Normal University) **Renjun Ma** (University of New Brunswick) **Jiaxiu Li** (University of New Brunswick)

Analysis of longitudinal binomial data with positive association between the number of successes and the number of failures: an application to stock instability study

Analyse de données binomiales longitudinales avec association positive entre le nombre de réussites et le nombre d'échecs : une application à l'étude de l'instabilité des stocks

Numerous methods have been developed for longitudinal binomial data in the literature. These traditional methods are reasonable for longitudinal binomial data

De nombreuses méthodes ont été développées pour les données binomiales longitudinales dans la littérature. Ces méthodes traditionnelles sont raisonnables pour des données binomiales longitu-

with a negative association between the number of successes and the number of failures over time; however, a positive association may occur between the number of successes and the number of failures over time in some behaviour, economic, disease aggregation and toxicological studies as the numbers of trials are often random. In this talk, we present a joint Poisson mixed modelling approach to longitudinal binomial data with a positive association between longitudinal counts of successes and longitudinal counts of failures. This approach can accommodate both a random and zero number of trials. It can also accommodate overdispersion and zero inflation in the number of successes and the number of failures. An optimal estimation method for our model has been developed using the orthodox best linear unbiased predictors. Our approach provides robust inference against misspecified random effects distributions and consolidates the subject-specific and population-averaged inferences. The usefulness of our approach is illustrated with an analysis of quarterly bivariate count data of stock daily limit-ups and limit-downs.

dinales avec une association négative entre le nombre de réussites et le nombre d'échecs au fil du temps ; une association positive peut toutefois se produire entre le nombre de réussites et le nombre d'échecs au fil du temps dans certaines études comportementales, économiques, d'agrégation de maladies et toxicologiques, étant donné que le nombre d'essais est souvent aléatoire. Dans cette présentation, nous présentons une approche de modélisation conjointe mixte de Poisson pour les données binomiales longitudinales avec une association positive entre les décomptes longitudinaux des réussites et les décomptes longitudinaux des échecs. Cette approche peut s'adapter aussi bien à un nombre aléatoire qu'à un nombre nul d'essais de même qu'à une surdispersion et à une inflation nulle du nombre de réussites et du nombre d'échecs. Une méthode d'estimation optimale pour notre modèle a été développée en utilisant les meilleurs prédicteurs linéaires orthodoxes sans biais. En plus de fournir une inférence robuste aux distributions d'effets aléatoires mal spécifiées, notre approche consolide également les inférences spécifiques au sujet et à la moyenne de la population. L'utilité de notre approche est illustrée par une analyse des données trimestrielles de comptage bivarié des limites quotidiennes des stocks à la hausse et à la baisse.

[14:30-15:00]

Xiaoqing Zhang (University of Regina) **Dianliang Deng** (University of Regina)

The Generalized Lindley Binomial Model for Analyzing the Proportions with End-point Inflation

Le modèle binomial généralisé de Lindley pour l'analyse de proportions avec inflation terminale

In this paper, a new three-parameter probability distribution called Generalized Lindley-binomial (GLB-3) distribution is proposed to analyze the proportional data with end-point inflation. The probabilistic properties of the distribution such as the probability mass function, the r-th factorial moment function are derived. The Fisher scoring algorithm and the EM algorithm are presented for computing estimations of parameters in the proposed GLB-3 model with/without covariates. A limited simulation study is performed to indicate the performance of derived EM algorithms for estimating parameters in the generalized lindley-binomial model with/without covariates. The whitefly dataset is used to present the application to the proposed model.

Dans cet article, nous proposons une nouvelle distribution de probabilités à trois paramètres baptisée distribution binomiale généralisée de Lindley (GLB-3) dans le but d'analyser les données proportionnelles avec inflation terminale. Nous avons dérivé les propriétés probabilistes de la distribution comme la fonction de masse de probabilités et la fonction de moment factoriel d'ordre r. Nous présentons l'algorithme de dénombrement de Fisher et l'algorithme EM pour le calcul des estimations des paramètres dans le modèle GLB-3 avec ou sans covariables. Une étude en simulations limitées a été réalisée pour démontrer la performance des algorithmes EM dérivés pour estimer les paramètres dans le modèle binomial généralisée de Lindley avec ou sans covariables. Nous nous servons de l'ensemble de données sur les aleurodes pour illustrer l'application du modèle proposé.

Modelling heterogeneity in complex high-dimensional data
Modélisation de l'hétérogénéité dans les données complexes à haute dimension

Chair/Président: Abbas Khalili

Organizer/Responsable: Abbas Khalili

Room/Salle: AT 102

Date: Wednesday May 31 / mercredi 31 mai

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-14:00]

Tudor A. Manole (Carnegie Mellon University) **Patrick Bryant** (Carnegie Mellon University) **John Alison** (Carnegie Mellon University) **Mikael Kuusela** (Carnegie Mellon University) **Larry Wasserman** (Carnegie Mellon University)

Homogeneity Testing in the Search for Pairs of Higgs Bosons at the Large Hadron Collider

Test d'homogénéité à la recherche de paires des bosons de Higgs dans le Grand collisionneur de hadrons

The Higgs boson is a particle in the Standard Model of particle physics, which was theorized in the 1960s as being partly responsible for the mechanism of particle mass generation. In July 2012, two independent experiments at the Large Hadron Collider (LHC) announced the observation of a new particle consistent with the Higgs boson. One of the long-term goals at the LHC is now to search for the production of pairs of Higgs bosons. In this talk, we describe some of the statistical challenges posed by such a measurement. We begin by describing the statistical procedure used for carrying out searches at the LHC, which amounts to a complex test for homogeneity. We then describe some difficulties in applying this procedure to the search for pairs of Higgs bosons, and we propose a new solution based on optimal transport theory. This talk is based on joint work with Patrick Bryant, John Alison, Mikael Kuusela, and Larry Wasserman.

Le boson de Higgs est une particule dans le modèle standard de la physique des particules. Une hypothèse dans les années 60 suggérait qu'il était partiellement responsable du mécanisme de génération de masse de particules. En juillet 2012, deux expériences indépendantes avec le Grand collisionneur de hadrons (LHC) ont annoncé l'observation d'une nouvelle particule cohérente avec le boson de Higgs. L'un des objectifs à long terme au LHC est maintenant de rechercher la production de paires de bosons de Higgs. Lors de cet exposé, nous décrivons certains des défis statistiques que pose une telle mesure. Nous commençons en décrivant la procédure statistique employée pour mener nos recherches au LHC, qui se résume à un test complexe d'homogénéité. Nous décrivons ensuite certaines difficultés d'application de la procédure à la recherche des paires de bosons de Higgs, et proposons une nouvelle solution basée sur la théorie du transport optimal. Cet exposé est fondé sur le travail conjoint de Patrick Bryant, John Alison, Mikael Kuusela, et Larry Wasserman.

[14:00-14:30]

Pengqi Liu (McGill University)

Regularization in Finite Mixture of Sparse GLMs with Ultra-High Dimensionality

Régularisation dans les mélanges finis de GLM épars avec une dimensionnalité ultra-élevée

Finite mixture of generalized linear regression models (FM-GLM) are used for analyzing data that arise from populations with unobserved heterogeneity. In recent applications of FM-GLM, data are often collected on a large number of features. However, fitting an FM-GLM to such high-dimensional data is numerically challenging. To cope with the high-dimensionality in estimation,

Les mélanges finis de modèles de régression linéaire généralisée (FM-GLM) sont utilisés pour analyser des données provenant de populations présentant une hétérogénéité non observée. Dans les applications récentes des FM-GLM, les données sont souvent collectées sur un grand nombre de caractéristiques. Cependant, l'ajustement d'un FM-GLM à ces données en haute dimension est numériquement difficile. Pour faire face à la haute dimension-

it is often assumed that the model is sparse and only a handful of features are relevant to the analysis. Most of the existing development on sparse estimation is in the context of homogeneous regression or supervised learning problems. In this talk, we discuss some of the challenges and also recent computational and theoretical developments for sparse estimation in FM-GLM when the number of features can be in exponential order of the sample size. We also discuss our results in finite-sample settings via simulations and a real data analysis.

nalité de l'estimation, on suppose souvent que le modèle est peu dense et que seules quelques caractéristiques sont pertinentes pour l'analyse. La plupart des développements existants sur l'estimation éparse couvrent le contexte de la régression homogène ou des problèmes d'apprentissage supervisé. Dans cet exposé, nous discutons de certains défis ainsi que des développements calculatoires et théoriques récents pour l'estimation peu dense dans les FM-GLM lorsque le nombre de caractéristiques peut être de l'ordre exponentiel de la taille de l'échantillon. Nous discutons également de nos résultats sur des échantillons finis par le biais de simulations et d'une analyse de données réelles.

Isobel Loutit Invited Address
Allocution Isobel-Loutit

Chair/Président: Reza Ramezan

Organizer/Responsable: Reza Ramezan

Room/Salle: AT 101

Date: Wednesday May 31 / mercredi 31 mai

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-15:00]

Derek Bingham (Simon Fraser University)

Fast Emulation and Modular Calibration for Simulators with Functional Response

Émulation rapide et étalonnage modulaire pour simulateurs à réponse fonctionnelle

Scalable surrogate models facilitate computationally tractable emulation of computer models (or simulators) when large ensembles of runs are available. Gaussian Process (GP) models are commonly used for computer model emulation; however they cannot scale to truly large datasets. Dense functional output such as spatial or time-series data adds an additional layer of complexity that must be handled carefully for fast emulation. In this work we develop a highly scalable emulator for functional data motivated by Kennedy and O'Hagan (2001) and Higdon et al. (2008), but built upon Local Approximate Gaussian Process (Gramacy, 2016). We apply our emulator to model calibration using global GP lengthscale parameter estimates to scale the input space, which dramatically increases the speed of MCMC. We show that our fast approximation based emulator can be a suitable alternative to the methods in Higdon et al. (2008) for functional response at a fraction of the computational cost.

Les modèles de substitution évolutifs facilitent l'émulation calculable de modèles informatiques (ou de simulateurs) lorsque de grands ensembles d'exécutions sont disponibles. Les modèles de processus gaussiens (GP) sont couramment utilisés pour l'émulation de modèles informatiques, mais ils ne peuvent pas s'adapter à des ensembles de données vraiment importants. Les données fonctionnelles denses, telles que les données spatiales ou temporelles, ajoutent une couche de complexité supplémentaire qui doit être traitée avec soin pour une émulation rapide. Dans ce travail, nous développons un émulateur hautement évolutif pour les données fonctionnelles, motivé par Kennedy et O'Hagan (2001) et Higdon et al. (2008), mais basé sur le processus gaussien approximatif local (Gramacy, 2016). Nous appliquons notre émulateur à la calibration de modèles en utilisant des estimations de paramètres GP globales à l'échelle de longueur pour mettre à l'échelle l'espace d'entrée, ce qui augmente considérablement la vitesse du MCMC. Nous montrons que notre émulateur basé sur l'approximation rapide peut être une alternative appropriée aux méthodes de Higdon et al. (2008) pour la réponse fonctionnelle à une fraction du coût de calcul.

Pierre Robillard Invited Address
Allocution du récipiendaire du Prix Pierre-Robillard

Chair/Président: Yingwei (Paul) Peng

Organizer/Responsable: Yingwei (Paul) Peng

Room/Salle: RB 2200

Date: Wednesday May 31 / mercredi 31 mai

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-15:00]

Saifuddin Syed (The University of British Columbia)

Scalable Bayesian inference with non-reversible parallel tempering

Inférence bayésienne évolutive avec atténuation parallèle non réversible

Markov chain Monte Carlo (MCMC) methods are the most widely used tools in Bayesian statistics for making inferences from complex posterior distributions. For challenging problems where the posterior is high-dimensional with well-separated modes, MCMC algorithms can get trapped exploring local regions of high probability. Parallel tempering (PT) tackles this problem by delegating the task of global exploration to a tractable reference distribution (e.g. prior) which communicates to the target (e.g. posterior) through a sequence of parallel MCMC algorithms targeting distributions of increasing complexity to the target. The classical approach to designing PT algorithms relied on a reversibility assumption, making PT challenging to tune and even deteriorating performance when introducing too many parallel chains. This talk will introduce a new non-reversible paradigm for PT (NRPT) that dominates its reversible counterpart while avoiding the performance collapse endemic to reversible PT methods. We will then establish near-optimal tuning guidelines, an efficient black-box methodology scalable to GPUs. Researchers have applied NRPT at scale to study the evolutionary structure of cancer, model plasma dynamics in nuclear fusion reactors, discover magnetic polarization in supermassive black hole M87, and generate the first image of Sagittarius A*, the supermassive black hole at the center of the Milky Way galaxy.

Les méthodes de Monte Carlo par chaînes de Markov (MCMC) sont les outils les plus utilisés en statistique bayésienne pour effectuer des inférences à partir de distributions postérieures complexes. Pour les problèmes difficiles où la distribution postérieure est en haute dimension avec des modes bien séparés, les algorithmes MCMC peuvent être piégés en explorant des régions locales de haute probabilité. L'atténuation parallèle (parallel tempering, PT) s'attaque à ce problème en déléguant la tâche d'exploration globale à une distribution de référence traitable (p. ex. a priori) qui communique à la cible (p. ex. postérieur) par une séquence d'algorithmes MCMC parallèles ciblant des distributions de complexité croissante à la cible. L'approche classique de la conception d'algorithmes de PT repose sur une hypothèse de réversibilité, ce qui rend le PT difficile à régler et entraîne même une détérioration des performances lorsqu'on introduit un trop grand nombre de chaînes parallèles. Cet exposé présentera un nouveau paradigme non réversible pour le PT (NRPT) qui surperforme son homologue réversible tout en évitant l'effondrement des performances endémique aux méthodes de PT réversibles. Nous établirons ensuite des directives de réglage quasi-optimales, une méthodologie efficace de boîte noire extensible aux GPU. Les chercheurs ont appliqué le NRPT à grande échelle pour étudier la structure évolutive du cancer, modéliser la dynamique du plasma dans les réacteurs à fusion nucléaire, découvrir la polarisation magnétique dans le trou noir supermassif M87 et générer la première image de Sagittarius A*, le trou noir supermassif au centre de la galaxie de la Voie lactée.

Methods for Non-Normal and Misclassified Data
Méthodes pour les données non normales et mal classées

Chair/Président: Zelalem Firisa Negeri

Room/Salle: TB 208

Date: Wednesday May 31 / mercredi 31 mai

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-13:45]

Rick E. Danielson (Fisheries and Oceans Canada)

Modelling Data and Phenomena in Complementary Ways

Modéliser des données et des phénomènes de façons complémentaires

Notable advances have been made in parametric (McCullagh 2002) and algorithmic (Breiman 2001) data modelling, but even prior to the digital data explosion, phenomenological models like Newton's equations of motion were being reworked into what would become the weather and climate models of today. We propose that a complementary derivation, and correspondingly detailed interpretation of individual terms, is possible for a core family of data models. Such a derivation for the canonical linear regression model is not known to exist, and this reworking involves treating representation and equation errors as signal terms with their own interpretation. We also seek to distinguish data and phenomena models, highlight their complementary aspects, and link these to ongoing efforts in parametric and algorithmic data modelling.

La modélisation de données algorithmiques (Breiman 2001) et paramétriques (McCullagh 2002) a connu des progrès significatifs. Cependant, même avant l'explosion des données numériques, des modèles phénoménologiques comme l'équation de Newton ont été retravaillés pour devenir les modèles climatiques et météorologiques d'aujourd'hui. Nous proposons qu'une dérivation complémentaire, et une interprétation détaillée correspondante des termes individuels soient possibles pour une famille clé de modèles de données. On sait qu'une telle dérivation existe pour le modèle de régression linéaire canonique, et cette réadaptation implique de considérer les erreurs d'équation et de représentation comme des termes de signaux ayant leur propre interprétation. Nous cherchons aussi à distinguer les données et les modèles de phénomène, souligner leurs aspects complémentaires et le relier à des travaux actuels en modélisation de données algorithmiques et paramétriques.

[13:45-14:00]

Jean-François Bégin (Simon Fraser University)

Ensemble Economic Scenario Generators: Unity Makes Strength

Générateurs de scénarios économiques ensemblistes : L'union fait la force

Over the last 40 years, various frameworks have been proposed to model economic and financial variables relevant to actuaries. These models are helpful but searching for a unique model that gives optimal forecasting performance can be frustrating and ultimately futile. This study therefore investigates whether we can create better economic scenario generators by combining them. We first consider eight prominent economic scenario generators and apply Bayesian estimation techniques to them, thus allowing us to account for parameter uncertainty. We then rely on predictive distribution stacking to obtain optimal model weights that pre-

Au cours des 40 dernières années, divers modèles ont été proposés pour modéliser les variables économiques et financières pertinentes pour les actuaires. Ces modèles sont utiles mais la recherche d'un modèle unique donnant des performances optimales en matière de prévision peut être frustrante et futile. Cette étude cherche donc à savoir si nous pouvons créer de meilleurs générateurs de scénarios économiques (GSEs) en les combinant. Nous examinons d'abord huit GSEs et les estimons en appliquant des techniques d'estimation bayésiennes. Nous utilisons une méthode basée sur l'empilement des distributions prédictives pour obtenir les poids optimaux des modèles, prescrivant la manière dont les modèles doivent être combinés. Une étude empirique

scribe how the models should be averaged. An empirical study based on three economies is performed. We find that the optimal weights change over time and differ from one economy to another. The out-of-sample behaviour of the ensemble model compares favourably to the other eight models. Creating ensembles is thus beneficial from an out-of-sample perspective.

basée sur trois pays est réalisée. Nous constatons que les poids optimaux évoluent dans le temps et diffèrent d'un pays à l'autre. Le comportement hors échantillon du modèle ensembliste se compare favorablement aux huit autres modèles.

[14:00-14:15]

Xintong Li (University of Waterloo) **Ben Mingbin Feng** (University of Waterloo)

Machine learning proxies for estimating tail risk measures of dynamic hedging errors via nested simulation

Proxys dans l'apprentissage automatique pour l'estimation des mesures de risque de queue des erreurs de couverture dynamique par simulation imbriquée

Variable Annuities (VAs) are equity-linked insurance contracts with embedded guarantees. Insurers typically use dynamic hedging programs, such as delta hedging, to manage market risk exposures in VAs. Dynamic hedging uses a portfolio of assets whose value moves in tandem in the direction of the value of liability associated with the guaranteed minimum benefits. As a result, the gain or loss from the hedging portfolio is expected to offset, at least to some extent, the liability under the guaranteed minimum benefit. In practice, there will always be gaps between the evolution of the hedge portfolio and the hedge position required for the liability, which makes slippages in hedge portfolios inevitable. Estimating the tail risk measures of VA liabilities, including the hedging error, is of prime interest to risk management and regulatory capital assessment. This is a multi-period nested estimation problem: The outer-level simulation generates the underlying risk factors' plausible realizations. For each outer scenario, an inner-level simulation is used to construct the hedging portfolios at all rebalancing points. Such a nested simulation procedure is computationally burdensome. In this study, we propose a two-stage simulation procedure that first uses a generic machine learning proxy to identify scenarios that are likely to have large hedging errors. Then we concentrate the remaining simulation budget to run extensive inner simulations on these potential tail scenarios. Our numerical experiments show that a generic machine learning proxy results in substantial computational savings, and it generalizes well for different VA contracts.

Les rentes variables sont des contrats d'assurance liée aux valeurs boursières comportant des garanties intégrées. Les assureurs utilisent en général des programmes de couverture dynamique, tels que la couverture delta, pour gérer l'exposition au risque du marché dans le cadre des rentes variables. La couverture dynamique utilise un portefeuille d'actifs dont la valeur évolue parallèlement à la valeur du passif associé aux prestations minimales garanties. Par conséquent, le gain ou la perte du portefeuille de couverture devrait compenser, au moins dans une certaine mesure, le passif au titre des prestations minimales garanties. Dans la pratique, il existera toujours des écarts entre l'évolution du portefeuille de couverture et la position de couverture requise pour le passif, ce qui rend inévitables les dérapages dans les portefeuilles de couverture. L'estimation des mesures du risque de queue du passif des rentes variables, y compris l'erreur de couverture, est primordiale pour la gestion des risques et l'évaluation du capital réglementaire. Il s'agit d'un problème d'estimation imbriquée sur plusieurs périodes : la simulation externe permet de créer de manière plausible les facteurs de risque sous-jacents. Pour chaque scénario externe, une simulation interne est utilisée pour bâtir les portefeuilles de couverture à tous les points de rééquilibrage. Une telle procédure de simulation imbriquée est lourde sur le plan informatique. Dans cette étude, nous proposons une procédure de simulation en deux étapes qui utilise d'abord un proxy générique dans le cadre de l'apprentissage automatique pour repérer les scénarios propices à d'importantes erreurs de couverture. Nous concentrons ensuite le budget de simulation restant pour effectuer des simulations internes approfondies sur ces scénarios de queue potentiels. Nos expériences numériques montrent qu'un proxy générique dans le cadre de l'apprentissage automatique permet de réaliser des économies informatiques substantielles et qu'il se généralise bien pour différents contrats de rentes variables.

[14:15-14:30]

Ziming Chen (University of Toronto) **Ewan Goligher** (Toronto General Hospital, University Health Network, Canada) **Eddy Fan** (Toronto General Hospital, University Health Network, Canada) **Anders Granholm** (Copenhagen University Hospitals,

Denmark) **Michael Harhay** (Epidemiology and Medicine, University of Pennsylvania, USA) **Danny Mcauley** (Queen's University Belfast, UK) **Martin Urner** (Toronto General Hospital, University Health Network, University of Toronto, Canada) **Christopher Yarnell** (Toronto General Hospital, Canada) **Anna Heath** (The Hospital for Sick Children, University of Toronto, Canada, University College London, London,UK)

Comparing Analysis Methods for Combining Mortality and Ventilator Time in Critical Care Trials: A Simulation Study

Comparaison de méthodes d'analyse pour combiner la mortalité et le temps de respirateur dans les essais en soins intensifs : une étude de simulations

Critical care trials often aim to evaluate the effects of treatment on mortality and time spent on a ventilator. There is no standard method that combines both these effects and maintains good interpretability. Thus, this simulation study aims to determine the most appropriate analysis method that combines these two outcomes. We evaluate conventional frequentist methods, hurdle models, proportional odds models and survival models fitted in a Bayesian framework. We generate data from two different mechanisms to guide overall method selection. We also propose a novel method to make decisions in Bayesian trials that evaluate two treatment effects. Methods were compared using statistical power and Type 1 error. T-tests have power above 0.9 but result in Type 1 error inflation above 0.05. Proportional odds models are less powerful but maintain power above 0.85 and penalize negative effects on mortality. Survival models with two treatment effects are more interpretable and maintain power.

Les essais en soins intensifs cherchent souvent à évaluer les effets d'un traitement sur la mortalité et le temps passé avec un respirateur. Il n'existe aucune méthode standard qui combine ces deux effets tout en restant facile à interpréter. C'est pourquoi cette étude de simulations cherche à déterminer la meilleure méthode d'analyse qui pourrait combiner ces deux réponses. Nous évaluons les méthodes fréquentistes conventionnelles, les modèles à barrière, les modèles de cotes proportionnelles et les modèles de survie ajustés dans un cadre bayésien. Nous générons des données à partir de deux différents mécanismes pour orienter la sélection de méthode globale. Nous proposons aussi une nouvelle méthode pour prendre des décisions dans des essais bayésiens qui évaluent deux effets de traitement. Nous avons comparé les méthodes au moyen de la puissance statistique et de l'erreur de type 1. Les tests-t ont une puissance supérieure à 0,9, mais entraînent une hausse de l'erreur de type 1 supérieure à 0,05. Les modèles de cotes proportionnelles ont une puissance inférieure, mais réussissent à maintenir la puissance en haut de 0,85 et pénalisent les effets négatifs sur la mortalité. Les modèles de survie sont plus faciles à interpréter et maintiennent la puissance.

[14:30-14:45]

Lin Liu (University of California, San Diego) **Ruohui Chen** (University of California, San Diego) **Loki Natarajan** (University of California, San Diego)

A Linear Mixed Model Approach for Measurement Error Adjustment

Approche de modèle linéaire mixte pour l'ajustement des erreurs de mesure

Wearable devices have been widely used to capture health behaviors such as physical activity. These sensor-based measures are deemed to be objective and less prone to self-report biases; while there can still be measurement errors from device recordings, which pose serious challenges for statistical analysis and obtaining unbiased risk estimates. We propose a two-stage linear mixed effect model based approach to correct bias caused by measurement errors when using multiple recordings for subjects from wearable devices. We provide theoretical proof of the debiasing process using the Best Linear Unbiased Predictors (BLUP), and use both simulation and a cohort study to demonstrate the performance of the proposed approach, while comparing to the naïve approach that directly uses device mea-

Les dispositifs portables sont largement utilisés pour mesurer les comportements en matière de santé, tels que l'activité physique. Ces mesures par capteurs sont considérées comme objectives et moins sujettes à des biais d'auto-déclaration ; cependant, il peut toujours y avoir des erreurs de mesure provenant des enregistrements des appareils, ce qui pose de sérieux problèmes pour l'analyse statistique et l'obtention d'estimations de risque non biaisées. Nous proposons une approche basée sur un modèle linéaire à effets mixtes en deux étapes pour corriger les biais causés par les erreurs de mesure lors de l'utilisation d'enregistrements multiples à partir de dispositifs portables. Nous fournissons une preuve théorique du processus de réduction du biais en utilisant les meilleurs prédicteurs linéaires non biaisés (BLUP), et utilisons à la fois la simulation et une étude de cohorte pour démontrer la performance de l'approche proposée, tout en la com-

sures without appropriately adjusting measurement errors. Our results show that our BLUP correction method can greatly reduce biases in risk estimates and thus enhance validity of study findings.

parant à l'approche naïve qui utilise directement les mesures du dispositif sans tenir compte des erreurs de mesure de manière appropriée. Nos résultats montrent que notre méthode de correction BLUP peut réduire considérablement les biais dans les estimations du risque et donc améliorer la validité des résultats de l'étude.

[14:45-15:00]

Philip J. Schmidt (University of Waterloo) **Monica Emelko** (University of Waterloo)

Probabilistic Modelling: A Not-so-new Way to Fix Problems with Least Squares Regression of qPCR Standard Curves

Modélisation probabiliste : une pas si nouvelle façon de régler des problèmes avec la régression des moindres carrés de courbes standard à qPCR

qPCR has been used in wastewater monitoring to track disease prevalence in populations throughout the COVID-19 pandemic. A standard curve is used to calibrate data (Cq values) to log-concentrations of target genes via least squares regression. This becomes unreliable at low concentrations and cannot explain non-detects that do not yield Cq values at all. Century-old fundamentals of quantitative microbiology have not been reflected in qPCR data analysis and explain patterns of Cq values and non-detects at low concentrations. We developed a novel hierarchical probabilistic modelling approach to draw inference from qPCR data at low concentrations, including interpretation of non-detects and concentration estimation from several technical replicates. Additionally, Bayesian uncertainty analysis was used to characterize uncertainty in concentration estimates. This realizes the value in data that might otherwise be discarded, facilitating public health decision-making from trace detections.

On se sert de la qPCR dans le contrôle des eaux usées afin de dépister la prévalence de maladie dans les populations durant la pandémie COVID-19. Une courbe standard est utilisée pour calibrer les données (valeurs Cq) aux concentrations logarithmiques des gènes cibles au moyen d'une régression par moindres carrés. Elle peut par contre être inconstante à des concentrations faibles et ne peut pas expliquer l'absence de détections qui ne génèrent aucune valeur Cq. Les fondements datant d'un siècle sur la microbiologie quantitative ne sont nullement reflétés dans l'analyse de données à qPCR et expliquent aussi les schémas des valeurs Cq et l'absence de détection à des concentrations faibles. Nous avons développé une nouvelle approche de modélisation probabiliste hiérarchique pour dessiner une inférence à partir des données de qPCR à des concentrations faibles, sans oublier l'interprétation des absences de détection et l'estimation de la concentration à partir de plusieurs répétitions techniques. De plus, nous avons employé une analyse d'incertitude bayésienne dans les estimations de concentration. De cette façon, nous réalisons la valeur des données qui auraient autrement été abandonnées et facilitons la prise de décision en santé publique grâce aux détections de trace de maladie.

Improved Methods for Linear and Non-Linear Models
Méthodes améliorées pour les modèles linéaires et non linéaires

Chair/Président: Denis Talbot

Room/Salle: TB 210

Date: Wednesday May 31 / mercredi 31 mai

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-13:45]

Alwell Oyet (Memorial University of Newfoundland) **Brajendra Sutradhar** (Memorial University of Newfoundland) **R. Prabhakar Rao** (Sri Sathya Sai Institute of Higher Learning)

Estimation of Multinomial Dynamic Logit Models

Estimation de modèles logit dynamiques multinomiaux

When categorical responses are collected repeatedly over a small period of time from a large number of independent individuals, the responses of each individual at a given time are likely to be influenced by the individual's fixed covariates. Also, these responses are likely to be correlated with the individual's past responses. We will develop a multinomial dynamic logits (MDL) model that accommodates the longitudinal correlations among the repeated multinomial responses. In addition to the influence of the fixed covariates, we will consider a situation where the MDL model is conditional on individual specific random effects in order to also accommodate the influence of unobservable random effects on the responses. We will discuss likelihood approaches for estimating the model parameters and examine its performance through simulation studies.

Lorsque des réponses catégorielles sont collectées de manière répétée sur une courte période auprès d'un grand nombre d'individus indépendants, les réponses de chaque individu à un moment donné sont susceptibles d'être influencées par les covariables fixes de celui-ci. En outre, ces réponses sont susceptibles d'être corrélées avec les réponses antérieures de l'individu. Nous développerons un modèle de logits dynamiques multinomiaux (MDL) qui prend en compte les corrélations longitudinales entre les réponses multinomiales répétées. Outre l'influence des covariables fixes, nous envisagerons une situation dans laquelle le modèle MDL est conditionnel à des effets aléatoires spécifiques à l'individu afin de tenir compte de l'influence d'effets aléatoires non observables sur les réponses. Nous discuterons des approches de vraisemblance pour l'estimation des paramètres du modèle et nous examinerons ses performances à l'aide d'études de simulation.

[13:45-14:00]

Michael John Ilagan (McGill University)

Using latent class analysis to counter trolling of chatbot safety judgments

Utilisation d'une analyse de classes latentes pour neutraliser le trollage des jugements des robots conversationnels en matière de sécurité

Amid interactions with live users, chatbots risk learning offensive behavior, such as racist remarks. To inform the chatbot's judgment, training examples of "safe" and "unsafe" utterances may be sourced continuously from live users. However, trolls are known to provide examples with incorrect class labels. To clean up trolled training data, previous work proposed removal of utterance/label pairs that have high cross-validation error. However, cross-validation is expensive and assumes that most users are honest. In this work, I aim to address both

Certaines interactions avec des utilisateurs humains risquent d'enseigner aux robots conversationnels des comportements offensants, comme les remarques racistes. Pour informer le jugement des robots conversationnels, des exemples d'entraînement aux énonciations « sécuritaires » et non sécuritaires » peuvent leur être fournis continuellement par des utilisateurs humains. Cependant, les trolleurs sont connus pour fournir des exemples d'étiquettes de classes incorrectes. Pour nettoyer les données d'entraînement entachées par le trollage, un travail antérieur proposait de supprimer les paires énonciation/étiquette ayant une erreur de validation

limitations. Drawing inspiration from the automated essay scoring literature, I propose collecting multiple user labels per utterance and replacing cross-validation with latent class analysis (LCA). LCA is relatively inexpensive, and in a simulation study, LCA-based predicted class labels maintained high accuracy even when majority of the users were trolls. Toward safer AI, algorithms need to be robust to trolls.

croisée élevée. Cependant, la validation croisée est coûteuse et part du principe que la plupart des utilisateurs sont honnêtes. Notre travail vise à traiter de ces deux limites. M'inspirant de la documentation sur le système automatisé de notation des essais, je propose la collecte d'étiquettes d'utilisateurs multiples par énonciation et le remplacement de la validation croisée par une analyse de classes latentes (LCA), relativement peu coûteuse. Dans une étude de simulations, l'exactitude des étiquettes de classes prédites basées sur la LCA est restée élevée, lorsque la majorité des utilisateurs étaient des trolleurs. Pour une intelligence artificielle plus sécuritaire, les algorithmes doivent être robustes face aux trolleurs.

[14:00-14:15]

Simon Lévesque (Université de Sherbrooke) **Félix Camirand Lemyre** (Université de Sherbrooke) **Jean-François Ethier** (Université de Sherbrooke)

Distributed Inference for Generalized Linear Models

Inférence distribuée pour les modèles linéaires généralisés

When all the data is available at a central location, many methods exist to fit a model and infer characteristics of the sample to the population. However, in many fields of application, researchers must work with many datasets which cannot be centralized. This is the case in health research, where confidential data may not cross borders or even leave hospitals. This situation lead to the use of so-called distributed methods for model fitting and inference for a variety of applied parametric models. In parallel, a theoretical framework to construct distributed estimators has been developed. Theses approaches make use of rounds of communication (i.e. sharing of summary statistics or estimates) between a trusted site and data sites. This presentation will focus on the link between the theoretical framework and the methods in use for generalized linear models, with an illustration using real data from the MIMIC-IV database.

Lorsque toutes les données requises sont disponibles à un point central, il existe une multitude de méthodes pour ajuster un modèle et inférer des caractéristiques à la population. Cependant, dans plusieurs champs d'application, les données sont divisées en plusieurs sous-ensembles qui ne peuvent être amalgamés, notamment en santé, où les données confidentielles ne peuvent souvent pas traverser les frontières. Ceci a amené à l'utilisation de méthodes d'ajustement et d'inférence dites distribuées pour une variété de modèles appliqués. En parallèle, un cadre théorique a été développé pour construire des estimateurs distribués. Ces approches font appel à des partages de statistiques sommaires entre un site de confiance et des sites de données. Dans cette présentation, on fera le lien entre le cadre théorique et les méthodes utilisées en pratique pour les modèles linéaires généralisés et on illustrera les résultats avec des données réelles provenant de la base de données MIMIC-IV.

[14:15-14:30]

Guanjie Lyu (University of Windsor) **Mohamed Belalia** (University of Windsor)

Smooth estimation of conditional quantile function using Bernstein polynomials

Estimation lisse de fonction quantile conditionnelle au moyen des polynômes de Bernstein

The conditional quantile function is an essential tool for understanding the relationship between covariates and the response variable in data analysis. We propose two approaches for smoothing Nadaraya-Watson's discrete conditional quantile estimator using Bernstein polynomials. Consistency and normality are established, a simulation study is carried out to assess the performance of these estimators.

La fonction quantile conditionnelle est un outil essentiel pour comprendre la relation entre les covariables et la variable de réponse dans l'analyse de données. Nous proposons deux approches pour lisser l'estimateur conditionnel discret de Nadaraya-Watson au moyen des polynômes de Bernstein. Nous établissons la convergence et la normalité et menons une étude de simulations pour évaluer la performance de ces estimateurs.

[14:30-14:45]

Anita Shahrokhian (Queen's University) **Chunfang Devon Lin** (Queen's University)

Adaptive Designs for Contour Estimation from Computer Experiments with Quantitative and Qualitative Inputs

Devis adaptatifs pour une estimation de contour tirée d'expériences informatiques avec entrées qualitatives et quantitatives

Computer experiments with quantitative and qualitative inputs are widely used to study scientific and engineering processes. Much of the work has focused on design and building surrogate models or process optimization. In this work, we propose an adaptive design approach for estimating a contour from computer experiments with mixed inputs by combining predictive mean and variance based on the EzGP to balance exploration and exploitation. The approach starts with an initial design and adds points sequentially. In each stage of the sequential procedure, the proposed approach divides the design points in the search space into three groups using confidence bounds, and within each group, an acquisition function is used to select a candidate point, and among the three selected points, we aim to choose the best solution for the contour of interest. The approach provides empirically more accurate contour estimation than the existing approaches. Theoretical justification of special cases is given.

On se sert couramment d'expériences informatiques avec entrées quantitatives et qualitatives pour étudier les processus scientifiques et d'ingénierie. La majorité du travail s'est concentré sur les devis et la construction de modèles de substitution ou l'optimisation de modèles. Dans le cadre de ce travail, nous proposons une approche de devis adaptatif pour estimer un contour à partir d'expériences informatiques avec des entrées mixtes en combinant une moyenne prédictive et une variance basée sur le EzGP pour équilibrer l'exploration et l'exploitation. L'approche commence par un devis initial puis ajoute des points de façon séquentielle. À chaque étape de la procédure séquentielle, l'approche proposée divise les points du devis dans l'espace de recherche en trois groupes au moyen de bornes de confiance. Ensuite, on utilise une fonction d'acquisition à l'intérieur de chacun de ces groupes afin de sélectionner un point candidat. Enfin, nous tentons de choisir parmi les trois points sélectionnés la meilleure solution pour le contour désiré. L'approche procure une estimation de contour empiriquement plus précise que celle des approches actuelles. Nous fournissons des justifications théoriques pour des cas spéciaux.

[14:45-15:00]

Fateh Chebana (Institut national de la recherche scientifique)

Multivariate L-moments and copula, Statistical inference with hydrological applications

L-moments et copules multivariés, inférence statistique et applications hydrologiques

L-moments are widely and commonly used in hydrological applications within a variety of statistical tools in the univariate framework. In this talk, we present new developments of statistical tools and methods based on the multivariate L-moments. L-moments (univariate and multivariate) have very interesting statistical features, especially useful and adapted to hydrology among other application fields. On the other hand, extreme events, such as in hydrology with floods or droughts, are characterized by several dependent random variables. Hence, the corresponding risk requires to be better evaluated in the multivariate setting and by making the most of all advantages of the L-moments and copula. First, we present an important and crucial result showing the connection between copulas and multivariate L-moments. Second, we propose a new method to estimate parameters of multiparameter copula along with goodness-of-fit tests specific for multiparameter copulas. Finally, we present multivariate homogeneity tests based on multivariate L-moments. Some asymptotic results, simula-

Les L-moments sont souvent utilisés dans les applications hydrologiques dans un grand nombre d'outils statistiques dans le cadre d'une analyse univariée. Nous présentons de nouveaux outils et de nouvelles méthodes statistiques basés sur les L-moments multivariés. Les L-moments (univariés et multivariés) présentent des caractéristiques statistiques très intéressantes, particulièrement utiles et adaptées à l'hydrologie, parmi d'autres domaines d'application. Par ailleurs, les événements extrêmes, tels que les inondations ou les sécheresses en hydrologie, sont caractérisés par plusieurs variables aléatoires dépendantes. Par conséquent, le risque correspondant doit être mieux évalué dans un cadre multivarié et il faut exploiter au mieux tous les avantages des L-moments et une copule. Tout d'abord, nous présentons des résultats importants et cruciaux montrant le lien entre les copules et les L-moments multivariés. Ensuite, nous proposons une nouvelle méthode pour estimer les paramètres des copules multiparamétriques, ainsi que des tests d'adéquation spécifiques aux copules multiparamétriques. Enfin, nous présentons des tests d'homogénéité multivariés en fonction des L-moments multivariés, des résultats asymptotiques, des études de simulation, ainsi que des études de cas, en particulier

tion studies as well as case studies are presented, especially in the hydrological context.

dans le contexte hydrologique.

Advances in Clustering
Progrès en regroupement

Chair/Président: Jeffrey L. Andrews

Organizer/Responsable: Paul David McNicholas

Room/Salle: AT 102

Date: Wednesday May 31 / mercredi 31 mai

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-15:52]

Mackenzie Neal (McMaster University) **Paul David McNicholas** (McMaster University)

Variable Selection for Skewed Clustering and Classification

Sélection de variables pour regroupement asymétrique et classification

As datasets from virtually all fields of endeavour continue to grow in size and complexity, the curse of dimensionality cannot be overlooked. Researchers in model-based clustering have recognized the need for effective dimension reduction techniques; as a result, many such algorithms exist to date. These algorithms, however, are often specific to Gaussian clustering problems and break down in the presence of skewness. We present a novel skewed variable selection algorithm that utilizes the Manly transformation mixture model to select variables based on their ability to separate clusters. We compare our approach with other asymmetric and normal variable selection methods using simulated and real-world datasets. We find that the proposed algorithm is suitable for dimension reduction in the presence of skewness.

Comme la taille et la complexité des ensembles de données dans pratiquement tous les domaines d'activité continuent d'augmenter, le fléau de la dimension ne peut pas être ignoré. Les chercheurs s'intéressant au problème de regroupement basé sur des modèles ont reconnu la nécessité de techniques efficaces de réduction de la dimension ; par conséquent, bon nombre d'algorithmes de cette nature existent aujourd'hui. Ces algorithmes sont toutefois souvent spécifiques au problème de regroupement gaussien et leur performance se dégrade en présence d'asymétrie. Nous présentons un nouvel algorithme de sélection de variables asymétriques qui fait appel au modèle de mélange par transformation de Manly pour la sélection de variables en se basant sur leur capacité à séparer les groupes. Nous comparons notre approche à d'autres méthodes de sélection asymétriques et à variables normales à l'aide d'ensembles de données simulées et de vraies données. Nous voyons que l'algorithme proposé convient pour la réduction de la dimension en présence d'une asymétrie.

[15:52-16:15]

Michael Gallaugher (Baylor University) **Eman Alamer** (McMaster University) **Paul David McNicholas** (McMaster University)

Mixture Modeling of Skewed Mixed-Type Data

Modélisation de mélanges de données de type mixte asymétriques

Data collected today is oftentimes of mixed type, meaning different types of variables are collected on the same subject. This poses some issues when using traditional statistical methods, especially in the area of clustering and classification. There is a relative paucity of methods presented in the literature thus far for clustering mixed type data. Moreover the methodology presented in the literature assumes that the variables treated as continu-

Comme les données collectées de nos jours sont souvent de type mixte, différents types de variables sont collectées sur un même sujet. Par conséquent, l'utilisation de méthodes statistiques traditionnelles pose certains problèmes, en particulier sur le plan de la mise en grappes et de la classification. Jusqu'à maintenant, la littérature présente relativement peu de méthodes pour la mise en grappes de données de type mixte. De plus, la méthodologie présentée dans la documentation suppose que les variables traitées

ous follow a Gaussian distribution, which can be a problem if that data is skewed or contains outliers. Specifically, if the Gaussian assumption is violated then this may lead to overfitting the number of groups and increased misclassification rate. This talk will present a mixture modelling method for clustering mixed type data that incorporates skewed distributions to help alleviate the aforementioned issues. Both simulated and real data will be used for illustration.

comme étant continues suivent une distribution gaussienne, ce qui peut être problématique si ces données sont asymétriques ou contiennent des valeurs aberrantes. Plus précisément, si l'hypothèse gaussienne est enfreinte, il peut s'ensuivre un surajustement du nombre de groupes et un taux plus élevé d'erreurs de classification. Notre exposé présente une méthode de modélisation de mélanges pour la mise en grappes de données de type mixte qui incorpore des distributions asymétriques dans le but d'aider à atténuer les problèmes mentionnés plus haut. Des données simulées et réelles servent à illustrer la méthode.

[16:15-16:38]

Brian Franczak (MacEwan University)

Recent developments in using mixtures of multivariate asymmetric distributions for classification

Avancées récentes dans l'utilisation de mélanges de distributions asymétriques multivariées pour la classification

We define classification as the process of assigning group labels to unlabelled observations. Classification is performed in multiple ways; for example, one can utilize unsupervised, semi-supervised, or fully supervised techniques. When a finite mixture model is used for classification, we can call the process model-based classification. This talk will discuss recent advances in the development of mixtures with multivariate density functions capable of modeling skewness directly. In the context of classification applications, topics may include one or more strategies for handling data with missing values, working with high-dimensional data sets, rectifying issues with typical parameter estimation schemes, parameterizing tail-weight separately in each dimension of the data, or accounting for outlying or spurious observations. The proposed models will be demonstrated using both simulated and real data. Model performance will be assessed using standard metrics and by comparison to popular publicly available methods.

Nous définissons la classification comme le processus d'attribution d'étiquettes de groupe à des observations non étiquetées. La classification est réalisée de plusieurs manières. Par exemple, il est possible d'utiliser des techniques non supervisées, semi-supervisées ou entièrement supervisées. Lorsqu'un modèle de mélange fini est utilisé pour la classification, on peut parler de classification basée sur un modèle. Au cours de cette présentation, nous discuterons des avancées récentes dans le développement de mélanges avec des fonctions de densité multivariées capables de modéliser directement l'asymétrie. Dans le contexte des applications de classification, les sujets peuvent inclure une ou plusieurs stratégies pour traiter les données avec des valeurs manquantes : travailler avec des ensembles de données de haute dimension, rectifier les problèmes avec les schémas typiques d'estimation des paramètres, paramétrer le poids de queue séparément dans chaque dimension des données, ou prendre en compte les observations aberrantes ou fallacieuses. Nous démontrerons les modèles proposés à l'aide de données simulées et réelles. Enfin, nous évaluerons l'efficacité du modèle à l'aide de mesures standard en le comparant à des méthodes populaires accessibles au public.

[16:38-17:00]

Paul David McNicholas (McMaster University)

Some Current Challenges in Clustering

Certains défis actuels relatifs au regroupement

A selection of some of the current challenges in the field of clustering are presented along with thoughts on how they might be tackled.

Nous présenterons certains défis actuels liés au domaine du regroupement et discuterons de solutions pour potentiellement les surmonter.

How should we conduct sensitivity analysis when making inferences that have policy implications? Learning from our colleagues in health economics.
Comment mener une analyse de sensibilité lorsque l'on effectue des inférences qui ont des implications politiques ? Leçons de nos collègues en économie de la santé

Chair/Président: Kuan Liu

Organizer/Responsable: Eleanor M. Pullenayegum

Room/Salle: AT 301

Date: Wednesday May 31 / mercredi 31 mai

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-15:30]

Eleanor M. Pullenayegum (Hospital for Sick Children) **Petros Pechlivanoglou** (The Hospital for Sick Children)

Roundtable: How should we conduct sensitivity analysis when making inferences that have policy implications? Learning from our colleagues in health economics.

Table ronde : comment réaliser une analyse de sensibilité lorsque l'élaboration d'inférences a des incidences politiques ? Apprendre de nos collègues en économie de la santé.

We increasingly use data collected as part of usual societal operations to make inferences that have policy implications, either at the local or national level. Because the data is not usually collected with research in mind, analyses typically make many assumptions. The sensitivity of our inferences to these assumptions is key to giving decision-makers a full description of the evidence. However, sensitivity analysis is often an afterthought. Health economists have been undertaking analyses with clear policy implications for many years, and have developed a structured approach to sensitivity analysis. The goal of this session is to learn from their approach, with a view to formulating best practices for sensitivity analysis when using data to inform policy decisions. We will use a case study and a draft proposal for best practices based on guidance documents from regulatory and health technology agencies to invite discussion and debate.

Nous utilisons de plus en plus les données recueillies par l'entremise des opérations sociales habituelles pour faire des inférences qui ont des incidences politiques, que ce soit au niveau local ou national. Puisque la plupart des données ne sont généralement pas recueillies dans le but de la recherche, les analyses vont souvent faire beaucoup d'hypothèses. La sensibilité de nos inférences relative à ces hypothèses est la clé pour fournir aux décideurs une description complète des résultats. Cependant, l'analyse de sensibilité n'est souvent qu'une idée après coup. Les économistes de la santé réalisent des analyses avec des incidences politiques claires depuis plusieurs années, et ont conçu une approche structurée pour l'analyse de sensibilité. L'objectif de cette séance est d'apprendre de leur approche et chercher à formuler les meilleures pratiques d'analyse de sensibilité lorsqu'on se sert de données pour informer les décisions politiques. Nous utiliserons une étude de cas et une ébauche de proposition des meilleures pratiques fondées sur des documents de références tirés des agences de technologie médicale et de réglementation afin d'inviter la discussion et le débat.

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

Chair/Président: Wen Wilson Lu

Organizer/Responsable: Wen Wilson Lu

Room/Salle: AT 101

Date: Wednesday May 31 / mercredi 31 mai

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-17:00]

J. Michael Brick (Westat Inc) **Jill M DeMatteis** (Westat)

Efficient Sample Designs for Multimode Surveys

Des plans efficaces pour des enquêtes multimodes

Multimode data collection is being used to deal with survey response rates and costs. Face-to-face collection may improve response rates and reduce nonresponse bias but is more costly than the Web. With face-to-face data collection, the sample is typically geographically clustered within PSUs. An efficient approach begins with the less costly modes followed by the more expensive face-to-face mode for a subsample of the units within each PSU. We present alternatives to this traditional design. One alternative subsamples PSUs rather than units to constrain costs. All Web nonrespondents in subsampled PSUs are followed up with face-to-face, with no follow-up in the not subsampled PSUs. Another alternative is a hybrid design that includes both a two-stage sample and an independent, unclustered sample. The face-to-face follow-up is used for all Web nonrespondents in the clustered sample only. We simulate several estimators and designs and show some advantages of the hybrid design.

On emploie la collecte de données multimode pour gérer les taux de réponse et les coûts d'une enquête. La collection faite en personne pourrait améliorer les taux de réponse et réduire le biais d'absence de réponse, mais c'est beaucoup plus dispendieux qu'en ligne. Au moyen d'une collecte faite en personne, l'échantillon est regroupé géographiquement à l'intérieur d'UPE. Une approche efficace débute par des méthodes de collecte abordables suivies par des rencontres en personne plus coûteuses afin de générer un sous-échantillon des unités à l'intérieur de chaque UPE. Nous présentons des solutions de rechange à cette méthode traditionnelle. Une option serait de sous-échantillonner les UPE plutôt que les unités pour réduire les coûts. Ensuite, tous les non répondants en ligne des UPE sous-échantillonnées seraient rencontrés en personne, alors qu'il n'y aurait pas de suivi effectué au sein des autres UPE. Une autre option serait une méthode hybride comprenant à la fois un échantillon à deux degrés et un échantillon indépendant non grappé. La rencontre en personne ne s'appliquerait que pour les non-répondants dans l'échantillon en grappes. Nous simulons divers estimateurs et méthodes et illustrons certains avantages du plan hybride.

Chair/Président: Yifan Li

Room/Salle: TB 238

Date: Wednesday May 31 / mercredi 31 mai

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-15:45]

Yun-hee Choi (Western University) **Seungwoo Lee** (Western University) **Laurent Briollais** (Lunenfeld-Tanenbaum Research Institute; Dalla Lana School of Public Health, University of Toronto)

A Correlated Competing-risks Model with Time-varying Covariates: An Application to Hereditary Breast and Ovarian Cancer Families

Modèle de risques concurrents corrélés avec covariables variables dans le temps : application aux familles touchées par les cancers héréditaires du sein et de l'ovaire

Time-to-event data arising from family-based studies is often complex due to various factors such as time varying covariates, competing risks, and familial correlation. Hereditary breast and ovarian cancer families suffer from high risks of both breast and ovarian cancers and are recommended to undergo frequent screenings or prophylactic surgery for prevention or early detection. We develop a correlated competing risks model to evaluate time-dependent effects of interventions on breast cancer risks in the presence of competing risks. Our proposed model accounts for time-varying covariates and their time-dependent effects and provides cause-specific penetrance estimates of cancer for family data. We apply our approach to 498 BRCA1 families recruited through the Breast Cancer Family Registry and illustrate the importance of our approach accounted for both competing risks and time varying effects when estimating cause-specific penetrance of breast cancer in the presence of ovarian cancer.

Les données de durée de vie issues d'études familiales sont souvent complexes en raison de divers facteurs tels que les covariables variables dans le temps, les risques concurrents et la corrélation familiale. Les familles touchées par le cancer héréditaire du sein et de l'ovaire présentent des risques élevés de ces deux types de cancer et il leur est recommandé de se soumettre à des dépistages fréquents ou à une chirurgie prophylactique à des fins de prévention ou de détection précoce. Nous développons un modèle de risques concurrents corrélés pour évaluer les effets temporels des interventions sur les risques de cancer du sein en présence de risques concurrents. Le modèle que nous proposons tient compte des covariables variables dans le temps et de leurs effets dépendants du temps, et fournit des estimations de la pénétrance du cancer en fonction de la cause pour les données familiales. Nous appliquons notre approche à 498 familles BRCA1 recrutées par le biais du Breast Cancer Family Registry et nous illustrons l'importance de notre approche, qui tient compte à la fois des risques concurrents et des effets variables dans le temps, lors de l'estimation de la pénétrance spécifique à la cause du cancer du sein en présence d'un cancer de l'ovaire.

[15:45-16:00]

Ayesha Madhushani Rathnayake Nayaka Bandaralage (Memorial University of Newfoundland) **Ashokan Mulayath Variyath** (Memorial University Newfoundland) **Zhaozhi Fan** (Memorial University Newfoundland)

Weighted Accelerated Failure Time Model

Modèle de temps de défaillance accéléré pondéré

The accelerated failure time (AFT) model is widely used in survival analysis and auxiliary information can be used to improve the efficiency of the model. We developed a weighted AFT model by using empirical like-

Le modèle de temps de défaillance accéléré est largement utilisé dans l'analyse de survie, et des données auxiliaires peuvent être utilisées pour améliorer l'efficacité du modèle. Nous avons conçu un modèle de temps de défaillance accéléré pondéré en utilisant

lihood probabilities as weights based on information from past studies. Theoretical justifications of the proposed model are developed. Simulation studies were conducted to compare the effectiveness of the proposed weighted AFT model and the standard AFT model.

des probabilités de vraisemblance empiriques comme poids, sur la base de données provenant d'études antérieures. Nous présentons des justifications théoriques du modèle proposé. Des études de simulation ont été menées pour comparer l'efficacité du modèle de temps de défaillance accéléré pondéré proposé et du modèle de temps de défaillance accéléré standard.

[16:00-16:15]

Rob Deardon (University of Calgary)

Population-dependent time-varying transmission rates in epidemic models

Taux de transmission variant dans le temps dépendant de la population dans les modèles d'épidémie

One of the many difficulties in modelling epidemic spread is that caused by behavioural change in the underlying population. Such changes in the underlying population can result in major changes in transmission dynamics of the disease, making the modelling challenging. We propose a model formulation where time-varying transmission is captured by the level of alarm in the population and specified as a function of the past epidemic trajectory. The model is set in a data-augmented Bayesian framework as epidemic data are often only partially observed, and we can utilize prior information to help with parameter identifiability. We investigate the estimability of the population alarm across a wide range of scenarios, using both parametric functions and non-parametric Gaussian process and splines. The benefit and utility of the proposed approach is illustrated through an application to COVID-19 data from New York City.

Une des nombreuses difficultés de modéliser une propagation épidémique est causée par des changements de comportement dans la population sous-jacente. De tels changements peuvent entraîner des changements importants dans la dynamique de transmission de la maladie, ce qui rend la modélisation problématique. Nous proposons la formulation d'un modèle dans lequel la transmission variant dans le temps est capturée par le niveau d'alerte dans la population et spécifiée comme fonction de la trajectoire épidémique passée. Le modèle est établi dans un cadre bayésien de données augmentées, puisque les données épidémiques sont souvent partiellement observées et nous pouvons utiliser l'information préalable pour aider à l'identifiabilité des paramètres. Nous étudions l'estimabilité de l'alerte dans la population à l'aide d'un large éventail de scénarios, en utilisant à la fois des fonctions paramétriques et des processus Gaussien non paramétriques ainsi que des splines. Une application à des données de la ville de New York sur la COVID-19 sert à illustrer l'avantage et l'utilité de l'approche proposée.

[16:15-16:30]

Chinmoy Roy Rahul (University of Calgary) **Rob Deardon** (University of Calgary)

Semi-parametric spatial infectious disease modeling

Modélisation spatiale semi-paramétrique pour les maladies infectieuses

Individual-level models (ILMs) fitted in a Bayesian Markov chain Monte Carlo (MCMC) framework can be used to understand the underlying mechanisms responsible for spreading infectious diseases, considering population heterogeneity via various individual-level covariates. However, these ILMs typically rely on strong parametric assumptions, and using inappropriate assumptions might lead to faulty conclusions. This research aim is to explore the use of semi-parametric spatial infectious disease transmission models in a Bayesian MCMC framework, which will help us to estimate the relationships between explanatory variables and the risk of infection more robustly. In this talk, we present results from simulated epidemics for constant

Les modèles de niveau individuel (ILM) ajustés dans un cadre bayésien de méthode de Monte-Carlo par chaînes de Markov (MCMC) peuvent être utilisés pour comprendre les mécanismes sous-jacents responsables de la propagation de maladies infectieuses, en prenant en compte l'hétérogénéité de la population à l'aide de diverses covariables de niveau individuel. Cependant, ces ILM reposent généralement sur des hypothèses paramétriques fortes et l'utilisation d'hypothèses inappropriées peut entraîner des conclusions erronées. Cette recherche vise à explorer l'utilisation de modèles spatiaux semi-paramétriques de transmission de maladies infectieuses dans un cadre bayésien de MCMC qui nous aideront à obtenir une estimation plus robuste de la relation entre les variables explicatives et le risque d'infection. Nous présentons des résultats à l'aide données épidémiologiques simulées pour des

and linear piecewise distance functions with both fixed and estimated change points. We also show how we can choose the number of change points via the Deviance Information Criteria (DIC).

fonctions de distance constantes et linéaires par segments, avec points de changement fixes et estimés. Nous montrons la possibilité de choisir le nombre de points de changement en utilisant le critère d'information de déviation (DIC).

[16:30-16:45]

Sayantee Jana (Indian Institute of Technology Hyderabad) **Sze Fu** (Dalla Lana School of Public Health, Centre for Global Health Research, St. Michael's Hospital, University of Toronto) **Helen Gellband** (Dalla Lana School of Public Health, Centre for Global Health Research, St. Michael's Hospital, University of Toronto) **Prabhat Jha** (Dalla Lana School of Public Health, Centre for Global Health Research, St. Michael's Hospital, University of Toronto) **Patrick Brown** (Dalla Lana School of Public Health, Centre for Global Health Research, St. Michael's Hospital, University of Toronto)

Spatio-temporal modelling of malaria mortality in India from the Million Death Study

Modélisation spatio-temporelle de la mortalité causée par le paludisme en Inde à partir de l'étude Million Death Study

Spatio-temporal modelling of malaria mortality is crucial for effective resource allocation and disease control policy-making, for the country's progress towards UN 2030 SDG. A spatio-temporal model to examine spatial distribution of malaria mortality rates and rate of temporal decline, across India, has been constructed. A "random slope" model was used, with malaria risk depending on a spatial relative risk surface and a linear time effect with a spatially-varying coefficient. The models were adjusted for urban/rural status, using 2004-13 Million Death Study data with nationwide geographic coverage. The rural population had twice the risk of death due to malaria compared to urban population. Malaria mortality in some of the highest-risk regions, are declining faster than other areas; however, rate of decline was not uniformly correlated with level of risk. Overall decline was faster after 2010. The results suggest a need for increased attention in high-risk rural populations.

La modélisation spatio-temporelle de la mortalité causée par le paludisme est cruciale pour l'attribution efficace des ressources et l'élaboration de politiques de lutte contre la maladie, ainsi que pour permettre au pays de poursuivre ses progrès en vue d'atteindre les objectifs de développement durable 2030 de l'ONU. Un modèle spatio-temporel a été créé pour analyser la distribution spatiale des taux de mortalité causés par le paludisme et le taux de déclin temporel dans toute l'Inde. Un modèle de « pente aléatoire » a été utilisé avec un risque de paludisme en fonction d'une surface de risque relative spatiale et d'un effet temporel linéaire avec un coefficient variant dans l'espace. Les modèles ont été adaptés en fonction des statuts urbain et rural avec les données de l'étude Million Death Study 2004-13 qui couvre l'ensemble du territoire national. Le risque de décès causé par le paludisme était deux fois plus élevé dans la population rurale que dans la population urbaine. La mortalité causée par le paludisme dans certaines des régions les plus à risque diminue plus rapidement que dans d'autres régions. Cependant, le taux de diminution n'est pas uniformément corrélé avec le niveau de risque. Le déclin global s'est accéléré après 2010. Les résultats suggèrent qu'il est nécessaire d'accorder une attention accrue aux populations rurales à risque élevé.

[16:45-17:00]

Wenling Zhang (University of Waterloo) **Cecilia A. Cotton** (University of Waterloo)

Causal Inference for Recurrent Events via Aggregated Marginal Odds Ratio

Inférence causale pour les événements récurrents par l'entremise du rapport de cotes marginal agrégé

In clinical and medical studies, researchers often work with longitudinal data. While there are various causal effect estimands designed for point treatment, the causal effect measures for time-varying treatment and recurrent events are relatively underdeveloped. In this presentation, we introduce a new causal measure that quantifies the effect of a time-varying treatment on recurrent events, and propose estimators with robust standard er-

En études cliniques et médicales, les chercheurs travaillent souvent avec des données longitudinales. Bien qu'il existe plusieurs estimateurs d'effet causal conçus pour les traitements ponctuels, les mesures d'effet causal pour les traitements à durée variable et les événements récurrents restent relativement sous-développées. Lors de cet exposé, nous présentons une nouvelle mesure d'effet causal qui quantifie l'effet d'un traitement à durée variable sur des événements récurrents. Nous proposons aussi des estimateurs avec

rors based on different weight models for both conventional causal measures and the proposed measure in different time settings. We demonstrate that the proposed causal estimand can be consistently estimated for study periods of moderate length, and the estimation results are compared under various treatment settings with different weight models. We also find that the proposed method is suitable for both absorbing and nonabsorbing treatments. The methods are applied to the 1997 National Longitudinal Study of Youth as an illustrative example.

des erreurs-types robustes basés sur différents modèles de poids pour les mesures causales conventionnelles et pour la mesure proposée dans différents cadres temporels. Nous démontrons que l'estimateur causal proposé peut être estimé de façon cohérente pour des périodes d'études de durée moyenne. Puis les résultats d'estimation sont comparés selon différents cadres de traitement et avec différents modèles de poids. Nous découvrons aussi que la méthode proposée est adéquate pour les traitements absorbants et non absorbants. Les méthodes sont appliquées à la National Longitudinal Study of Youth de 1997 en guise d'exemple.

Chair/Président: Guanjie Lyu

Room/Salle: TB 210

Date: Wednesday May 31 / mercredi 31 mai

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-15:45]

Félix Camirand Lemyre (Université de Sherbrooke) **Jean-François Quessy** (Université du Québec à Trois-Rivières)

Change-Point Tests and Estimators for Gradually Changing Dependence Structures Based on Kendall's Tau

Procédures inférentielles pour la détection de changements graduels au sein de structures de dépendance, basées sur le tau de Kendall

Suppose that pairs (X1,Y1),...,(Xn,Yn) of independent observations are subject to a gradual change in their stochastic behaviour in the sense that for given K1 ¡ K2 ∈ 1, . . . , n, the underlying joint distribution of a given pair is F before K1, G after K2 ¿ K1, and gradually moving from F to G between the two times of change K1 and K2. This set-up elegantly generalizes the usual abrupt change model which is usually assumed in change-point analysis. Under this configuration, asymptotically unbiased estimators of Kendall's tau up to the change and after the change are derived, as well as tests and estimators of gradual change points related to these measures of association. The asymptotic behaviour of the introduced estimators and test statistics as n $\to \infty$ is rigorously investigated, in particular by demonstrating a general result on weighted indexed U-statistics computed under heterogeneous data. The good sampling properties of the proposed methodologies are studied in a simulation study that considers various scenarios of gradual changes in dependence. The usefulness of the introduced tools is illustrated on time series of monthly atmospheric carbon dioxide concentrations and global temperature for the period 1959–2015.

Supposons qu'une séquence d'observations (X1,Y1),...,(Xn,Yn) indépendantes soit sujette à l'occurence d'un changement distributionnel graduel, au sens où pour K1 ¡ K2 ∈ 1, . . . , n, la distribution jointe d'une paire est donnée par F avant K1, G après K2, et changeante de K1 à K2 entre les deux temps de changement K1 et K2. Ce cadre généralise celui plus habituel de la détection de changement abrupte. Suivant ce modèle, des estimateurs asymptotiquement sans biais, des tests et des estimateurs du temps de changement sont développés. Le comportement asymptotique des estimateurs introduits et des tests statistics sont rigoureusement étudiés, qui sont basés sur un résultat général portant sur les U-statistiques pondérées reposant sur des observations hétérogènes. Les propriétés échantillonnages sont étudiées au moyen de simulations, et l'utilité est illustrée sur des données sérielles atmosphériques pour la période 1959-2015.

[15:45-16:00]

Ana Carolina da Cruz (Western University)

Bayesian Clustering Approach Based on Constant-Wise Change-Point Models

Approche bayésienne de regroupement basée sur des modèles de points de changement constants

Motivated by the research question of clustering abnormal (tumour) single cells based on their copy-number profiles, we propose a model-based approach for clustering observations based on their change-point pro-

Motivés par la question de la recherche sur le regroupement de cellules individuelles anormales (tumeurs) en fonction de leurs profils de nombre de copies, nous proposons une approche basée sur un modèle pour regrouper les observations en fonction de leurs

files via Gibbs sampler. Our model assumes a Dirichlet process on the constant-wise change-point structures to cluster observations that share the same change-point profiles. A Dirichlet process on the change-point structures allows the proposed nonparametric Bayesian approach to control the number of clusters in the model, not requiring the pre-specification of the number of clusters, as opposed to other clustering techniques. We evaluate the proposed model under various simulated scenarios and via a real data application. As a result, the proposed technique is among the earliest to infer the change-point positions and perform clustering simultaneously.

profils de points de changement au moyen de l'échantillonneur de Gibbs. Notre modèle repose sur l'hypothèse d'un processus de Dirichlet sur les structures de points de changement constants pour regrouper les observations qui présentent les mêmes profils de points de changement. Un processus de Dirichlet sur les structures des points de changement permet à l'approche bayésienne non paramétrique proposée de contrôler le nombre de groupes dans le modèle, sans qu'il soit nécessaire de le spécifier au préalable, contrairement à d'autres techniques de regroupement. Nous évaluons le modèle proposé dans le cadre de divers scénarios simulés et au moyen d'une application de données réelles. Par conséquent, la technique proposée est l'une des premières à déduire les positions des points de changement et à effectuer un regroupement simultanément.

[16:00-16:15]

Junwei Shen (McGill University) **Erica E. M. Moodie** (McGill University) **Shirin Golchi** (McGill University)

Privacy-Preservation Through Two-Stage Bayesian Meta-Analysis of Individualized Treatment Rules

Préservation de la confidentialité grâce à une méta-analyse bayésienne en deux étapes des règles de traitement individualisé

Individualized treatment rules (ITRs) tailor treatments to patients based on individual characteristics. Estimating ITRs requires identifying individuals who benefit most from the particular treatments, thus detecting treatment-covariate interactions. Large datasets from multisite studies may be required to develop an effective ITR due to the low power provided by smaller datasets for detecting the often small treatment-covariate interactions. However, sharing of individual-level data may be prohibited for confidentiality reasons. To address this issue, we adapt a two-stage Bayesian meta-analysis approach to estimate the optimal ITRs which optimize the expected patient outcomes using multisite data without disclosing the individual-level data beyond the site at which each individual's data is collected. We demonstrate the use of the proposed method via simulations and a real data application.

Les règles de traitement individualisé (RTI) permettent d'adapter les traitements aux patients en fonction de caractéristiques individuelles. L'estimation des RTI nécessite l'identification des individus qui bénéficient le plus des traitements particuliers, ce qui permet de détecter les interactions entre le traitement et les covariables. De vastes ensembles de données provenant d'études multisites peuvent être nécessaires pour mettre au point une RTI efficace, en raison de la faible puissance des petits ensembles de données pour la détection des interactions traitement-covariables, qui sont souvent minimes. Cependant, le partage des données individuelles peut être interdit pour des raisons de confidentialité. Pour résoudre ce problème, nous adaptons une approche de méta-analyse bayésienne en deux étapes pour estimer les RTI qui optimisent les résultats attendus pour les patients. Cette approche utilise des données multisites sans divulguer les données individuelles au-delà du site de collecte. Nous démontrons l'utilisation de la méthode proposée par le biais de simulations et d'une application sur des données réelles.

[16:15-16:30]

Chengqian Xian (Western University) **Camila P. E. de Souza** (University of Western Ontario) **Wenqing He** (University of Western Ontario) **Felipe F. Rodrigues** (King's University College at Western University) **Renfang Tian** (King's University College at Western University)

Variational Bayes Inference of Survival Data using Log-logistic Accelerated Failure Time Model

Inférence bayésienne variationnelle sur les données de survie à l'aide d'un modèle log-logistique de temps de défaillance accéléré

The log-logistic regression model is one of the most commonly used accelerated failure time (AFT) mod-

Le modèle de régression log-logistique est l'un des modèles de temps de défaillance accéléré (AFT) les plus couramment

els, for which statistical inference methods are mainly established under the frequentist framework. Recently, Bayesian inference for log-logistic AFT models using Markov chain Monte Carlo (MCMC) techniques has also been widely developed. In this work, we develop an alternative approach to MCMC methods and infer the parameters of the log-logistic AFT model via a mean-field variational Bayes (VB) algorithm. A piece-wise approximation technique is embedded in deriving the update equations in the VB algorithm to achieve conjugacy. The proposed VB algorithm is evaluated and compared with typical frequentist inferences using simulated data under various scenarios, and a publicly available dataset is employed for illustration. We demonstrate that the proposed VB algorithm can achieve good estimation accuracy and is not sensitive to sample sizes, censoring rates, and prior information.

utilisés, pour lequel les méthodes d'inférence statistique sont principalement établies dans le cadre fréquentiste. Récemment, l'inférence bayésienne pour les modèles AFT log-logistiques utilisant des techniques de Monte Carlo à chaîne de Markov (MCMC) a également été largement développée. Dans ce travail, nous développons une approche alternative aux méthodes MCMC et déduisons les paramètres du modèle AFT log-logistique via un algorithme bayésien variationnel (VB) en champ moyen. Nous intégrons une technique d'approximation par morceaux dans la dérivation des équations de mise à jour de l'algorithme VB afin de parvenir à la conjugaison. L'algorithme VB proposé est évalué et comparé aux inférences fréquentistes typiques à l'aide de données simulées dans divers scénarios, et un ensemble de données accessibles au public est utilisé à titre d'illustration. Nous démontrons que l'algorithme VB proposé peut atteindre une bonne précision d'estimation et qu'il n'est pas sensible à la taille des échantillons, aux taux de censure et aux informations préalables.

[16:30-16:45]

Leif Erik Lovblom (University of Toronto) **Laurent Briollais** (University of Toronto) **Bruce Perkins** (University of Toronto) **George Tomlinson** (University of Toronto)

Bayesian Estimation of a Joint Model for Multiple Longitudinal Outcomes and a Multistate Process under a Mixed Observation Scheme with Interval-Censoring

Estimation bayésienne d'un modèle conjoint pour des résultats longitudinaux multiples et un processus multi-états dans le cadre d'un schéma d'observation mixte avec censure par intervalles

Joint models can be used to address uncertainties about the co-development of multiple correlated non-fatal disease processes, and recently we formulated a novel joint model for a multistate process under a mixed observation scheme with interval-censoring and one longitudinal process. To improve the model's overall representation of disease, we extended it to allow multiple longitudinal outcomes. Specifically, we formulated a shared random effects joint model with a multivariate linear mixed-effects submodel and a proportional intensities progressive 3-state Markov submodel, with interval-censoring of entry into the second state and exact observation of entry into the absorbing state. A Bayesian approach using Hamiltonian Monte Carlo was used. The model was validated using simulation studies, and an application in diabetic microvascular complications found predictive roles for albuminuria, glomerular filtration, and retinopathy on rates of progression through stages of neuropathy.

Les modèles conjoints peuvent être utilisés pour traiter les incertitudes concernant le développement conjoint de plusieurs processus pathologiques non mortels corrélés. Nous avons récemment formulé un nouveau modèle conjoint pour un processus à plusieurs états dans le cadre d'un schéma d'observation mixte avec une censure d'intervalle et un processus longitudinal. Afin d'améliorer la représentation globale de la maladie dans le modèle, nous l'avons étendu pour permettre des résultats longitudinaux multiples. Plus précisément, nous avons formulé un modèle conjoint à effets aléatoires partagés avec un sous-modèle linéaire multivarié à effets mixtes et un sous-modèle de Markov progressif à trois états à intensités proportionnelles, avec une censure par intervalles de l'entrée dans le deuxième état et une observation exacte de l'entrée dans l'état d'absorption. Nous utilisons une approche bayésienne utilisant l'algorithme de Monte Carlo hamiltonien. Nous validons le modèle à l'aide d'études de simulation, et une application aux complications microvasculaires diabétiques permet d'établir des rôles prédictifs pour l'albuminurie, la filtration glomérulaire et la rétinopathie sur les taux de progression à travers les stades de la neuropathie.

[16:45-17:00]

Thierry Chekouo (University of Minnesota) **Xiaotian Dai** (University of Calgary) **Lu Xuewen** (University of Calgary)

A Bayesian variable selection approach incorporating prior feature ordering and population structures

Une approache bayésienne de sélection de variables qui incorpore a priori l'ordre des variables et la structure de la population

In this talk, we propose a novel Bayesian variable selection framework for the identification of important genetic variants associated with Coronary artery disease status. Instead of treating each feature independently as in conventional Bayesian variable selection methods, we propose an innovative prior for the inclusion probabilities of genetic variants that account for their ordering structure. We assume that neighboring variants are more likely to be selected together as they tend to be highly correlated and have similar biological functions. Additionally, we propose to group participating subjects based on underlying population structure and fit separate regressions, so that the regression coefficients can better reflect different disease risks in different population groups. Our approach borrows strength across regression models through an innovative prior inspired by the Markov random fields. The proposed framework can improve variable selection and prediction performances as demonstrated in the simulation studies. We also apply the proposed framework to the CATHGEN data with binary CAD disease status.

Dans cet exposé, nous proposons un nouveau cadre bayésien de sélection de variables pour l'identification d'importants variants génétiques associés au statut de maladie coronarienne. Au lieu de traiter chaque caractéristique indépendamment comme dans les méthodes conventionnelles de sélection de variables, nous proposons une nouvelle distribution a priori pour les probabilités d'inclusion des variants génétiques qui prend en compte leur structure ordonnée. Nous supposons que les variants voisins sont plus susceptibles d'être sélectionnés ensemble car ils ont tendance à être fortement corrélés et ont des fonctions biologiques similaires. De plus, nous proposons de regrouper les sujets participants en fonction de la structure de la population sous-jacente et d'ajuster des régressions distinctes, afin que les coefficients de régression puissent mieux refléter les différents risques de maladie dans différents groupes de population. Notre approche emprunte de la force aux modèles de régression grâce à un a priori innovant inspiré des champs aléatoires de Markov. Le cadre proposé peut améliorer les performances de sélection et de prédiction des variables comme démontré dans les études de simulation. Nous appliquons également le cadre proposé aux données CATHGEN avec un statut de maladie CAD binaire.

Author List • Liste des auteurs

Abbas, Eman Sreeni, 221
Abbas, EmanSreeni, 61
Abdi, Hervé, 57, 197
Abed, Amin, 46
Abrahamowicz, Michal, 71, 271
Acar, Elif, 42, 127
Adamowicz, Sarah J., 43, 131
Adams, Billy, 64, 239
Aghababaei Jazi, Omidali, 22, 29
Ahmadi Nadi, Adel, 38, 99
Ahmed, S. Ejaz, 61, 66, 219, 246
Ahmed, Suborna Shekhor, 67, 252
Alamer, Eman, 73, 287
Alenazi, Abdulaziz, 57, 200
Alexander, Monica, 48, 145
Alison, John, 71, 275
Andrews, Jeffrey L., 53, 177
Arango-Castillo, Lenin, 63, 230
Asghari, Shabnam, 43, 129
ASSANI, Ismael Afolabi, 124
Assani, IsmaelAfolabi, 42
Auger-Méthé, Marie, 239
Auger-Méthé, Marie, 64
Augustyniak, Maciej, 42, 66, 124, 248, 249
Awosoga, David, 47

Bégin, Jean-François, 66, 72
Ba, Ismaila, 54, 182
Badescu, Alexandru, 66, 248
Badescu, Andrei L., 51, 165
Bakker, Dmitri, 53, 176
Bandari, Ela, 45
Bandyopadhyay, Dipankar, 58, 205
Banik, Arjun, 50, 161
Baribeau, Béatrice, 69, 261
Barigou, Karim, 37, 96
Basu, Samopriya, 62, 223
Batakci, Leyla, 35, 81
Battle, Alexis, 63, 235
Baxter, Matthew, 49, 155
Beaulac, Cédric, 44, 49, 136, 154
Beaumont, Jean-François, 55, 118
Beaumont, Jean-François, 41, 55, 184
Becker, Devan G., 63, 235
Bédard, Mylène, 68, 257
Bégin, Jean-François, 49, 152, 248, 279
Belalia, Mohamed, 47, 73, 144, 284
Béliveau, Audrey, 63, 234
Ben-Ameur, Hatem, 54, 181
Benedetti, Andrea, 43, 128
Benjamin, Bloem-Reddy, 45, 141
Benoit, Hugues, 64, 240
Berger, David, 52, 170
Bhagwat, Pankaj Uttam, 55, 188
Bhardwaj, Shivani, 37, 95
Bian, Mengjie, 50, 159
Bickis, Mikelis, 56, 193
Bilodeau, Blair, 56, 194
Bingham, Derek, 54, 72, 183, 277
Binte Farooq, Fabiha, 39, 105
Blanchet, F. Guillaume, 37, 97
Bocci, Cynthia, 55, 184
Boire, François-Michel, 54, 181
Boisvert-Beaudry, Gabriel, 68, 257
Bond, Marjorie E., 35, 81
Bonner, Ron, 70, 267
Bonner, Simon J., 38, 70, 104, 267
Boonen, Tim, 59, 210
Bosa, Keven, 41, 55, 118, 184
Boudreault, Jérémie, 45, 140
Bowala Mudiyanselage, Sulalitha, 42, 126
Braun, John, 60, 223

Brick, J. Michael, 74, 290
Briollais, Laurent, 74, 75, 291, 297
Brown, Doug, 37, 97
Brown, Patrick, 41, 55, 62, 75, 120, 186, 227, 293
Browne, Ryan P., 53, 178
Bryant, Patrick, 71, 275
Buckeridge, David, 41, 121
Bui, Trang, 47, 50, 156
Burkett, Kelly, 61, 215

Cadenillas, Abel, 36, 88
Cadigan, Noel, 38, 101
Caetano, Samantha-Jo, 37, 99
Cai, Song, 70, 267
Calcetero, Sebastian F., 51, 165
Camirand Lemyre, Félix, 73, 75, 284, 295
Campagna, Celine, 45, 140
Campbell, Dave, 70, 268
Campbell, Kieran R, 68, 257
Canty, Angelo J., 50, 159
Cao, Jiguo, 44, 136
Cao, Jingyi, 47, 143
Cabrera, Manuel Ordóñez, 58, 206
Carone, Marco, 65, 243
Chan, Christian, 43, 46, 132
Chaouch, Mohamed, 55, 187
Chapman, Judy-Anne W., 40, 114
Charpentier, Arthur, 22, 26
Chavez-Martinez, Gilberto, 66, 247
Chebana, Fateh, 45, 73, 140, 285
Chekouo, Thierry, 43, 75, 132, 297
Chelsea Uggenti, Chelsea, 22
Chen, Chen, 58, 205
Chen, Guanyu, 61, 221
Chen, Jiachen, 38, 101
Chen, Jiahua, 55, 187
Chen, Jianchu, 39, 105
Chen, Li-Pang, 64, 237
Chen, Meixi, 38, 104
Chen, Ruohui, 72, 281
Chen, Sixia, 37, 92
Chen, Yiting, 46, 61, 217
Chen, Yuxin, 35, 84
Chen, Ze, 40, 112
Chen, Zhiyin, 45
Chen, Ziming, 72, 280
Chenouri, Shojaeddin, 56, 194
Chien, Yu-chun, 46
Chinchure, Aditya, 47
Choi, Yun-hee, 74, 291
Chong, Michael Y.C., 48, 145
Chowdhury, Abhiroop, 47
Chung, Young Jo, 46
Colijn, Caroline, 62, 227
Cook, Richard J., 39, 105
Corona, María José Orraca, 63, 230
Cotton, Cecilia A., 75, 293
Coulombe, Janie, 34, 79
Cowen, Laura, 50, 62, 161, 227
Cox, Miranda, 53, 176
Craiu, Radu, 54, 68, 183, 257
Crandles, Bryn, 46

da Cruz, Ana Carolina, 75, 295
Dagdoug, Mehdi, 55, 190
Dai, Wenshu, 62, 228
Dai, Xiaotian, 43, 75, 132, 297
Daignault, Katherine, 22, 29
Dang, Sanjeena, 53, 56, 62, 63, 177, 191, 222, 228, 231
Dang, Utkarsh J., 62, 63, 222, 231
Danielson, Rick E., 72, 279
Dasylva, Abel C., 41, 118
Datta, Gauri, 55, 185
de Haan-Ward, Johanna, 38, 104
de Mast, Jeroen, 49, 154
De Silva, Balage Don Harshani Hiranthika, 45, 50, 161
de Souza, Camila P. E., 69, 75, 263, 296
Dean, Nema, 51, 167
Deardon, Rob, 41, 74, 120, 292
Deeth, Lorna, 41, 120
Del Moral, Pierre, 34, 77
DeMatteis, Jill M, 290
DeMatteis, Jill M., 74
Deng, Dianliang, 71, 274
Desrochers, Rachel, 38, 67, 103, 254
Dey, Ankana, 37, 97
Diao, Liqun, 48, 146
Dias, Ronaldo, 69, 263
Díaz, Iván, 43, 133
Diaz-Rodriguez, Jairo, 38, 100
Dioni, Abdoulaye, 42, 127
Doig, Christopher, 56, 193
Dong, Gracia, 54, 181
Dong, Larry, 36, 89
Douwes-Schultz, Dirk, 41, 121
Dubin, Joel A., 44, 50, 63, 137, 159, 232

Dubois, Janique, 67, 252
Dunham, Bruce, 34, 81
Dyck, Justin, 46

Eadie, Gwendolyn, 54, 183
El Emam, Khaled, 70, 270
El Kababji, Samer, 70, 270
Emelko, Monica, 73, 282
Epasinghege Dona, Nirodha M., 46, 50, 158
Ertefaie, Ashkan, 52, 170
Eslami, Aida, 57, 197
Espin-Garcia, Osvaldo, 65, 241
Esteavo, Victor, 55, 184
Ethier, Jean-François, 73, 284
Eustache, Esther, 37, 93
Evans, Michael J., 68, 259

Fakiolas, Georgia, 53, 176
Fan, Eddy, 72, 280
Fan, Jun, 41, 117
Fan, Zhaozhi, 74, 291
Fang, Yuan, 38, 62, 101, 228
Fazelpour, Sina, 54, 179
Feng, Ben, 22, 26
Feng, Ben Mingbin, 72, 280
Feng, Cindy, 71, 273
Feng, Shui, 48, 149
Feng, Zeny, 43, 131
Ferreira Guerra, Steve, 71, 271
Fick, Gordon, 56, 195
Field, Chris, 64, 240
Fletcher, Corey, 45
Fonseca, Thaís C. O., 70, 269
Forkutza, Greg, 46
Fournier, Patrick, 46
Franczak, Brian, 73, 288
Frantz, Brower, 64, 239
French, Megan, 46
Fu, Haoda, 65, 243
Fu, Sze, 75, 293
Fung, Daryl L.X., 64, 236

Gaensler, Bryan, 54, 183
Gai, Chunyi, 35, 84
Gallagher, Colin, 50, 161
Gallanosa, Jervis, 37, 95
Gallaugher, Michael, 73, 287
Gao, Lucy, 63, 235
Gao, Wanye, 60, 225
Gao, Xin, 56, 191
Garrett, Rose, 45, 139
Gatten, Justin, 64, 239
Ge, Haoxuan, 60, 215
Gellband, Helen, 75, 293
Ghaffartehrani, Alireza, 47
Ghannam, Mai, 63, 232
Ghoudi, Kilani, 55, 187
Gibbs, Alison L., 42, 123
Gillis, Darren, 64, 240
Giuliano, Rita, 58, 206
Gobeil, Gabriel, 70, 268
Godin, Frédéric, 66, 249
Goffard, Pierre-Olivier, 37, 96
Goga, Camelia, 55, 190
Golchi, Shirin, 39, 75, 109, 296
Goligher, Ewan, 72, 280
Gordillo, Luis F., 35, 83
Graham, Jinko, 50, 158
Granholm, Anders, 72, 280
Greenberg, Morris, 68, 257
Greenwood, Celia M. T., 51, 68, 167, 256
Greenwood, Priscilla E., 35, 83
Gryba, Rowenna, 64, 239
Gu, Yuwen, 41, 117
Guan, Tianyu, 61, 219
Guérin, Hélène, 69, 263
Guo, Jinqi, 46
Gweon, Hyukjun (Jay), 66, 248

Hagar, Luke, 63, 230
Han, Lengyi, 60, 223
Han, Peisong, 59, 212
Han, Yutong, 60, 216
Harcharek, Qaiyaan, 64, 239
Harhay, Michael, 72, 281
Hasan, Md., 46
Hasler, Caren, 37, 93
Hastie, Trevor, 65, 242
Hatefi, Armin, 37, 95
Haziza, David, 37, 38, 55, 67, 70, 92, 103, 190, 254, 267
He, Hao, 70, 267
He, Jiayang, 46
He, Jiayin, 46
He, Jingni, 71, 272
He, Shiyu, 47, 63, 231
He, Wenqing, 75, 296
Heath, Anna, 36, 39, 72, 89, 108, 281

Henry, Greg, 64, 239
Herman, Jonathan, 49, 151
Herrera Martin, Antonio, 54, 183
Herrera, Ronald, 43, 133
Herrmann, Klaus Peter, 69, 264
Higgins, Julian P. T., 44, 135
Hlady, Marcus, 43, 131
Hofert, Marius, 69, 264
Hoque, Md. Erfanul, 42, 127
Hou, Zhaoran, 51, 166
Hou-Liu, Jason, 22, 27
Hu, Boyi, 56, 190
Hu, Pingzhao, 45, 64, 141, 236
Huang, Beili, 37, 93
Huang, Jingyue, 39, 106
Huang, Mei Ling, 60, 62, 222, 224
Huang, Shiheng, 47, 143
Huang, Yuying, 47
Huntington, Henry, 64, 239
Hurley, Oliver, 43, 129

Ibañez, Dominique, 61, 215
Ilagan, Michael John, 73, 283
Ivancevic, Dana, 61, 215

Jaffer, Shaista, 48, 147
Jaimungal, Sebastian, 59, 66, 210, 250
Jalbert, Jonathan, 70, 268
Jaman, Ajmery, 51, 163
Jana, Sayantee, 75, 293
Janeczko, Bailey Trenton, 70, 265
Jayaraman, Sarath Kumar, 66, 248
Jeganathan, Pratheepa, 47, 143
Jha, Prabhat, 75, 293
Jian, Jie, 51, 166
Jiang, Cong, 34, 79
Jiang, Depeng, 43, 57, 128, 197
Jiang, Wenjun, 59, 210
Jiang, Wenyu, 36, 90
Jimoh, Hameed A., 43, 132
Jing, Kaili, 48, 148
Johnson, Brad C., 61, 221
Johnson, David W., 36, 89
Joshipura, Kashish Hemal, 47

Kalaria, Shreena Nisha, 46
Kalugama Gardige, Hashan Savinda Peiris, 42, 46, 125
Kamath, Ravish, 45
Kang, Sohee, 67, 251
Kansal, Abhinav, 47
Karim, Mohammad Ehsanul, 62, 228
Karimnezhad, Ali, 61, 215
Karunanayake, Nayanthi, 45
Karunarathna, Charith, 50, 158
Kazi, Ayman Arik, 45
Ke, Yuwei, 45
Keith, David, 36, 90
Kerby-Helm, April, 35, 81
Khalili, Abbas, 66, 247
Khan, Shahedul, 43, 50, 132, 162
Kim, Jae-Kwang, 23, 32, 60, 212
Klassen, Terry P., 36, 89
Ko, Myoungjin, 44, 135
Kohar, Herteg, 37, 97
Kolhatkar, Varada, 22, 31
Kolokolnikov, Theodore, 35, 84
Kong, Linglong, 66, 246
Konstantinidis, Menelaos, 44, 135
Kopciuk, Karen A., 61, 218
Krause, Heather, 54, 179
Krikella, Tatiana, 63, 232
Kroell, Emma, 66, 250
Kulperger, Reg, 39, 110
Kuusela, Mikael, 71, 275
Kwan, Karina, 51, 167

Labbe, Aurélie, 59, 208, 209
Labiba, Zayeeda Shahreen, 45
Lac, Leann, 64, 236
Laib, Naamane, 55, 187
Lawler, Ethan, 64, 240
Ledesma, Luis, 44, 136
Lee, Chel Hee, 56, 193
Lee, Hongzhe, 57, 201
Lee, Juhyung, 55, 185
Lee, Seungwoo, 74, 291
Lee, Tae Yoon, 46
Lei, Mengying, 59, 209
Leung, Carson K., 64, 236
Lévesque, Simon, 73
Lévesque, Simon, 284
Li, Bin, 36, 87
Li, Chendong, 43, 128
Li, Deli, 55, 186
Li, Dongchen, 47, 143
Li, Hangjian, 40, 116
Li, Hong, 40, 112, 113

Li, Jiacheng, 55, 185
Li, Jiaxiu, 71, 273
Li, Linke, 46, 61, 219
Li, Longhai, 41, 121
Li, Mufan, 37, 96
Li, Na, 71, 271
Li, Peizhi, 56, 193
Li, Qing, 71, 272
Li, Shu, 66, 248
Li, Xintong, 72, 280
Li, Yifan, 39, 110
Lian, Yi, 41, 117
Liang, Haodi, 55, 187
Liang, Wei, 36, 92
Liang, You, 58, 203
Lim, Kyuson, 62, 222
Lin, Becky Wei, 49, 151
Lin, Chunfang Devon, 71, 73, 271, 285
Lin, Eric, 64, 236
Lin, Shili, 48, 147
Lin, X. Sheldon, 51, 165
Lin, Xiyou, 46
Lipman, Danika M., 60, 225
Liu, Boyuan, 64, 236
Liu, Haiyan, 59, 211
Liu, Kuan, 58, 204
Liu, Lin, 72, 281
Liu, Pengqi, 71, 275
Liu, Wenyue, 36, 88
Liu, Yan, 43, 61, 133, 221
Liu, Yang, 37, 97
Liu, Yixiu, 57, 197
Lix, Lisa M., 52, 172
Lockhart, Richard, 43, 129
Loisel, Stéphane, 37, 96
Long, Quan, 43, 132
Loukine, Lidia, 61, 215
Lovblom, Leif Erik, 75, 297
Lu, Xuewen, 36, 43, 89, 132
Lu, Zihang, 58, 204
Lunetta, Kathryn, 38, 101
Luong, The Minh, 61, 215
Luu, My, 45
Lysy, Martin, 38, 63, 104, 234
Lyu, Guanjie, 47, 73, 144, 284
Lyu, Yangsai, 45
Lyu, Yunhong, 55, 61, 188, 218
Ma, Renjun, 71, 273
Mackay, Anne, 66, 249
Madrid Padilla, Oscar Hernan, 40, 116
Mahmoudi, Fatemeh, 36, 89
Manole, Tudor A., 71, 275
Mao, Yu, 40, 112
Maranda, Guillaume, 41, 118
Marchand, Éric P., 69, 260
Mariathas, Hensley Hubert, 129
Mariathas, HensleyHubert, 43
Marin, Mike, 35, 82
Marshall, François A, 198
Marshall, FrançoisA, 57
Marshall, William, 60, 62, 222, 224
Martínez-Ramírez, Francisco J., 63, 230
Martsynyuk, Yuliya V., 37, 43, 70, 95, 131, 265
Mavridis, Dimitris, 44, 135
May, Jacqueline A., 43, 131
Mcauley, Danny, 72, 281
McDonald, Raphael Robert, 36, 90
McDonald, Shaun, 70, 268
McDougald, Warren, 53, 176
McGee, Glen, 48, 145
McGregor, Kevin, 65, 241
McNichol, Jennifer, 69, 261
McNicholas, Paul David, 53, 56, 73, 74, 177, 191, 287, 288
McNicol, Jamie, 47, 143
McVittie, James H., 56, 195
Mei, Ying, 46
Miao, Ying, 60, 224
Miao, Yu, 55, 186
Michal, Victoire, 56, 192
Mien, Laura Lim Huey, 46
Mills Flemming, Joanna Elizabeth, 36, 52, 64, 90, 173, 240
Mills, Kathryn, 69, 261
Min, Shijie, 47
Min, Sonny, 46
Mitani, Aya A., 58, 205
Mitsakakis, Nicholas, 51, 70, 165, 270
Mondol, Momenul Haque, 62, 228
Mongeon, Kevin, 67, 252
Moodie, Erica E. M., 34, 51, 54, 69, 75, 79, 167, 182, 261, 296
Morenz, Eric, 65, 243
Morgan, Kevonn, 40, 114
Morrissette, Samuel, 45, 51, 163

Mossman, Alexandra, 46, 61, 220
Mukherjee, Manan, 46
Mulayath Variyath, Ashokan, 74, 291
Munaweera Arachchilage, Inesh Prabuddha, 64, 240
Murabito, Joanne, 38, 101
Muthukumarana, Saman, 39, 50, 51, 64, 106, 161, 163, 240

Nasari, Masoud, 39, 111
Nasri, Bouchra, 53, 174
Natarajan, Loki, 72, 281
Nauenberg, Shimon, 51, 165
Nayaka Bandaralage, Ayesha Madhushani Rathnayake, 74, 291
Neal, Mackenzie, 73, 287
Negeri, Zelalem Firisa, 38, 100
Nešlehová, Johanna G., 69, 264
Neufeld, Anna, 63, 235
Neupane, Saurav, 47
Ng, Wai Yu Amanda, 37, 61, 99, 216, 220
Ngi-Song, Adele, 38, 67, 103, 254
Nguyen, Minh Chau, 47
Nicholls, Gradon, 46
Nickchi, Payman, 50, 158
Nieuwoudt, Christina, 39, 105
Nirmalkanna, Kunasekaran, 38, 101
Niu, Ziang, 65, 244
Nkurunziza, Sévérien, 55, 61, 63, 188, 218, 232
Nunkoo, Aadesh, 47

O'Loughlin, Jennifer, 43, 52, 133, 170
Olobatuyi, Kehinde I., 62, 227
Ostblom, Joel, 22, 31
Ould Haye, Mohamedou, 58, 206
Oyet, Alwell, 73, 283

Pahwa, Punam, 50, 162
Panahi, Mahsa, 49, 154
Panasiuk, Sofia, 61, 220
Parsons, Marc Angelo, 43, 128
Payne, Andrea, 56, 191
Pechlivanoglou, Petros, 74, 289
Pekos, Jason, 47
Peng, Yingwei, 56, 193
Perkins, Bruce, 75, 297
Perry, Brittany Alana, 60, 224
Pesenti, Silvana Manuela, 59, 66, 210, 250
Petrakos, Niki Z., 46
Philippe, Anne, 58, 206
Picka, Jeffrey D., 70, 269
Pigeon, Mathieu, 42, 124
Platt, Robert, 71, 271
Plint, Amy, 36, 89
Poon, Tak Shing (Dustin), 46
Popp, Joshua, 63, 235
Posner, Michael A., 35, 81
Provost, Serge B., 69, 264
Pullenayegum, Eleanor M., 44, 45, 74, 136, 139, 289

Qi, Yongcheng, 55, 186
Qin, Xiaoke, 53, 57, 177, 198
Qin, Yingli, 44, 70, 137, 270
Quessy, Jean-François, 75, 295

Raad, May, 40, 114
Rahul, Chinmoy Roy, 46, 74, 292
Ramezan, Reza, 38, 104
Ramsay, Kelly, 56, 194
Rana, Masud, 50, 162
Rao, Jonnagadda N., 55, 184
Rao, R. Prabhakar, 73, 283
Ratz, Philipp, 42, 125
Ravikumar, Preethi, 45, 139
Raymond, Mélanie, 46
Reesor, Mark, 54, 181
Rémillard, Bruno N., 48, 149
Renoux, Christel, 34, 79
Rezaee, Hassan, 59, 208
Rice, Gregory, 44, 137
Robet, Caroline, 58, 206
Roch, Alexandre, 36, 87
Rodrigues, Felipe F., 75, 296
Rodriques, Myanca, 44, 135
Romaniuk, Raymond, 62, 222
Roopnarine, Cameron, 46
Rosenberg, Morgan, 45
Rothstein, Steven, 53, 56, 177, 191
Roy, Daniel M., 68, 258
Roy, Paritosh Kumar, 70, 269
Roy, Philippe, 70, 268
Ruth, William, 43, 129

Sajobi, Tolu, 53, 173
Salhi, Yahia, 37, 96
Salim, Christopher, 61, 219
Salisbury, Thomas, 49, 152
Sama, Fousseni, 63, 234
Sameoto, Jessica, 36, 90

Sanders, Barbara, 49, 152
Sang, Peijun, 41, 44, 116, 137
Saporito, Yuri, 59, 210
Sarren, Robert, 64, 239
Sarvet, Aaron, 34, 78
Saunders, David, 49, 153
Scealy, Janice Lea, 57, 200
Schanzer, Dena L., 50, 156
Schaubroeck, Matt, 39, 106
Schmidt, Alexandra, 41, 56, 59, 70, 121, 192, 208, 269
Schmidt, Philip J., 73, 282
Schnitzer, Mireille, 43, 52, 133, 170
Schwarzer, Guido, 44, 135
Scotti, Simone, 36, 87
Segal, Mark, 65, 242
Shahrokhian, Anita, 73, 285
Sharma, Divyansh, 47
Sharp, Alexander, 53, 178
Shen, Junwei, 46, 75, 296
Shen, Yannan, 41, 121
Shi, Xiaofei, 22, 26
Shi, Xiaoping, 61, 217
Shi, Yuliang, 50, 159
Shi, Yuxin, 46
Shih, Ching-Lin, 61, 221
Shortreed, Susan M., 34, 79
Sidrow, Evan, 69, 262
Silva, Anjali, 53, 56, 177, 191
Singh, Gurbakhshash, 56, 195
Sjarova , Nathalie , 176
Sjarova, Nathalie, 53
Skrzydlo, Diana Katherine, 53, 175
Slater, Justin, 67, 251
Smith, Chrystal E., 53, 175
Snyman, Simon, 60, 223
Soave, David, 51, 167
Sobhan, Shamsia, 46
Sombo, Samuel, 55, 184
Son, Donghui, 46
Song, Yanglei, 36, 71, 90, 271
Souza, Pedro H. T. O., 69, 263
Spicker, Dylan Z., 54, 182
St-Pierre, Julien, 46
St. John, Philip, 43, 128
Stafford, James, 55, 186
Steele, Russell, 43, 128
Steiner, Stefan, 38, 49, 50, 99, 154, 156
Stenning, David, 54, 183
Stentoft, Lars, 54, 181
Stepanova, Natalia, 58, 207
Stephens, David A., 52, 169
Stevens, Nathaniel T., 38, 50, 63, 99, 156, 230
Stewart, Connie, 57, 200
Stipancic, Joshua, 59, 208
Strawderman, William, 55, 188
Stringer, Alex, 55, 56, 186, 194
Stukel, Thérèse A., 69, 261
Su, Jianxi, 40, 112, 113
Su, Wanhua, 62, 228
Sun, Lijun, 59, 208, 209
Sutradhar, Brajendra, 73, 283
Swishchuk, Anatoliy, 23, 33
Syed, Saifuddin, 72, 278
Sylvestre, Marie-Pierre, 43, 52, 133, 170
Szturm, Tony, 50, 161
Szyszkowicz, Barbara, 39, 110

Talbot, Denis, 52, 170, 171
Tang, Yanbo, 56, 194
Targino, Rodrigo, 59, 210
Tate, Robert, 43, 128
Thabane, Lehana, 22, 25, 52, 60, 172, 214
Thompson, Mary, 34, 70, 79, 265
Thorpe, Kevin, 22, 25
Tian, Renfang, 75, 296
Tomlinson, George, 75, 297
Torabi, Mahmoud, 42, 127
Toupin, Marie-Hélène, 41, 119
Tran, Nam-Anh, 62, 229
Tricco, Andrea, 44, 135
Tsagris, Michail, 57, 200
Tsokani, Sofia, 44, 135
Tu, Dongsheng, 36, 71, 90, 271
Tu, Wangshu, 63, 231
Turcicova, Marie, 58, 207
Turgeon, Maxime, 51, 163
Tuzhilina, Elena, 65, 242
Tzou, Justin, 35, 84

Uggenti, Chelsea, 29
Unfried, Alana, 35, 81
Urner, Martin, 72, 281

Vachon, Marie-Claude, 66, 249
Vaillancourt, Jean, 48, 149
Valdez, Emiliano, 22, 26
Vanessa, McNealis, 51, 167

Ventz, Steffen, 39, 108
Veroniki, Areti-Angeliki, 44, 135
Volodin, Andrei, 58, 206
VonDuyke, Andrew, 64, 239
Vu, Quynh Phuong, 37, 61, 99, 217

Waechter, Jason, 56, 193
Wakefield, Jon, 56, 192
Walker, Braedan, 70, 267
Wallace, Michael, 34, 79
Wanasundara, Surajith N., 39, 106
Wang, Guanbo, 52, 65, 170, 244
Wang, Jue, 60, 215
Wang, Lan, 65, 243
Wang, Lu, 34, 78
Wang, Rina, 47
Wang, Xinrui, 45
Wang, Xu (Sunny), 37, 97
Wang, Yiran, 63, 234
Wang, Zehui, 45
Wang, Zhonglei, 60, 212
Wang, Zilin, 70, 265
Ward, Caitlin, 41, 120
Ward, Madeline, 41, 46, 120
Wasserman, Larry, 71, 275
Weerasekera, S. J. W. W. M. M. P, 38, 101
Wen, Lan, 52, 170
Whitaker, Douglas, 35, 81
White, Bethany J. G., 67, 251
White, Ian R., 44, 135
Whoriskey, Kim, 64, 240
Wickramasinghe, Ashani N., 39, 45, 106
Willard, James, 39, 109
Witten, Daniela, 63, 235
Wolock, Charles, 65, 243
Wong, Samuel W.K., 63, 231
Woolford, Douglas G., 38, 50, 104, 156
Wu, Changbao, 36, 39, 60, 92, 106, 213
Wu, Haoyu, 46
Wu, Jou-Chin, 64, 237
Wu, Sidi, 44, 47, 136
Wu, Xiao, 37, 61, 99, 216
Wu, Yilei, 70, 270
Wu, Yunan, 65, 243

Xi, Quanhan, 45, 141
Xia, Yuan, 46
Xian, Chengqian, 75, 296
Xie, Aoqi, 47
Xie, Minge, 69, 259
Xing, Li, 58, 202
Xu, Deen, 46
Xu, Jingyi, 46
Xu, Mingwei, 51, 162
Xu, Wei, 45, 56, 140, 141, 191
Xu, Zhenhuan, 46
Xuewen, Lu, 75, 297

Yan, Guohua, 71, 273
Yan, Tianxing, 47
Yanez, Juan-Sebastian, 42, 124
Yang, Archer, 41, 117
Yang, Mu, 47
Yang, Po, 60, 224
Yarnell, Christopher, 72, 281
Yazdi, Faezeh, 62, 223
Yeh, Chi-Kuang, 44, 137
Yi, Grace Y., 48, 64, 146, 237, 238
Yi, Yanqing, 62, 227
Young, Virginia R., 47, 143
Yu, Hao, 39, 110
Yu, Tingzhou, 35, 85
Yuan, Jinxin, 50, 155
Yuan, Yan, 60, 216

Zeng, Leilei, 39, 106
Zhang, Chi, 44, 137
Zhang, Erin, 47
Zhang, Jasper Zhongyuan, 45, 46, 141
Zhang, Jia Ning, 37, 92
Zhang, Qihuang, 64, 238
Zhang, Qingrun, 71, 272
Zhang, Qiyue, 46
Zhang, Wenling, 47, 75, 293
Zhang, Xiaolei, 71, 273
Zhang, Xiaoqing, 71, 274
Zhang, Xuekui, 57, 202
Zhang, Ziang, 55, 186
Zhao, Xianglin, 46
Zhao, Yue, 41, 117
Zhao, Zixuan, 36, 90
Zheng, Chuyi, 67, 252
Zheng, Jianing, 56, 193
Zheng, Liyuan, 70, 270
Zheng, Sunqiaohe, 46
Zhong, Yuan, 56, 191

Zhou, Kenneth Q., 40, 112
Zhou, Qing, 40, 116
Zhou, Wenyuan, 49, 152
Zhou, Zhou, 68, 255
Zhu, Jennifer, 35, 85
Zhu, Mu, 70, 270
Zhu, Yeying, 50, 159
Zhuang, Haoxin, 48, 146
Zhuang, Ziqian, 45, 46, 140
Zidek, Jim V., 68, 256
Zou, Bin, 47, 143
Zou, Yushu, 47