Société statistique du Canada
Statistical Society of Canada

47th Annual Meeting of the Statistical Society of Canada

47e Congrès annuel de la Société statistique du Canada

May 26 - May 29, 2019
26 mai au 29 mai 2019

University of Calgary

Table of Contents • Table des matières

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

Welcome to Calgary • Bienvenue à Calgary **2**

Sponsors • Commanditaires **3**

Exhibitors • Exposants **4**

Career Connections • Connexion carrière **4**

Organizers • Organisateurs **5**

General Information • Informations générales **6**

The Conference • Le congrès **9**

Social Events • Activités sociales **11**

Committees and Meetings • Comités et réunions **13**

Social and Information Events • Événements sociaux et informatifs **17**

Program • Programme **18**

Abstracts • Résumés **59**

Author List • Liste des auteurs **266**

Welcome to Calgary • Bienvenue à Calgary

The Departments of Mathematics and Statistics and Community Health Sciences at the University of Calgary welcome you to Calgary! We are thrilled to be hosting our annual meeting, for the very first time. The University of Calgary is a research-intensive university founded in 1966, with 14 faculties offering more than 250 academic programs, and more than 50 research institutes and centres. As one of Canada's top comprehensive research universities, UCalgary combines the best of university tradition with the city of Calgary's vibrant energy and diversity. It is located in the heart of Southern Alberta, the traditional territories of the people of Treaty 7 region. The City of Calgary is also home to the Metis Nation of Alberta, Region III.

The Department of Mathematics and Statistics is home to 15 faculty members in Statistics and Actuarial Science, while the Department of Community Health Sciences has four faculty members in Biostatistics and Health Data Science. The close collaboration between the two departments as well as biostatistics colleagues across campus is evident in the many activities of the University of Calgary Biostatistics Centre (UCBC). Recent educational programs include an Interdisciplinary Specialization in Biostatistics (thesis-based degrees) as well as a proposed Professional Master's in Data Science with a specialization in Health Data Science & Biostatistics.

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

- SSC 2019:
  https://ssc.ca/en/meeting/annual/2019
- Calgary Meetings and Conventions:
  https://choosecalgary.ca/ssc2019/

Les Départements de mathématiques et statistique et de sciences de la santé communautaire de l'Université de Calgary vous souhaitent la bienvenue dans notre ville ! Nous sommes ravis d'accueillir pour la toute première fois notre congrès annuel. Fondée en 1966, l'Université de Calgary est très axée sur la recherche, avec 14 facultés qui offrent plus de 250 programmes universitaires et plus de 50 instituts et centres de recherche. L'une des principales universités polyvalentes de recherche du Canada, UCalgary combine la meilleure tradition universitaire à l'énergie et la diversité dynamiques de la ville de Calgary. Elle est située au cœur du Sud de l'Alberta, sur les territoires traditionnels des peuples de la région du Traité 7 et de la Nation métisse de l'Alberta, Région III.

Le Département de mathématiques et statistique emploie 15 professeurs en statistique et science actuarielle et le Département de sciences de la santé communautaire en compte quatre en biostatistique et science des données de santé. L'étroite collaboration entre les deux départements et avec d'autres collègues actifs en biostatistique ailleurs sur le campus se reflète dans les nombreuses activités du Centre de biostatistique de l'Université de Calgary (CBUC). Les programmes de formation récents comprennent une spécialisation interdisciplinaire en biostatistique (diplômes fondés sur une thèse) ainsi qu'une maîtrise professionnelle proposée en science des données avec une spécialisation en sciences de la santé et biostatistique.

Pour vous aider à planifier votre séjour, veuillez consulter :

- SSC 2019 :
  https://ssc.ca/fr/congres/annuel/2019
- Rencontres et congrès à Calgary :
  https://choosecalgary.ca/ssc2019/

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 :

- Department of Community Health Sciences, University of Calgary
- Department of Mathematics & Statistics, University of Calgary
- Faculty of Science, University of Calgary
- Vice-President (Research), University of Calgary

- Fields Institute
- Pacific Institute for the Mathematical Sciences
- Centre de recherches mathématiques
- Canadian Statistical Sciences Institute / Institut canadien des sciences statistiques

- Goldspot Discoveries

UNIVERSITY OF CALGARY

FIELDS

PIMS

CRM

CANSSI INCASS

GOLDSPOT

Exhibitors • Exposants

The exhibitors, providing information and displays for examination and purchase, will be located in the Science Theatres area, a short distance away from the rooms where all scientific sessions will be held, and near the coffee break/registration location. The exhibitors will be available Monday, May 27 and Tuesday, May 28 from 9:00 am to 4:00 pm.

- Pearson https://www.pearson.com/ca/en.html
- Dovank http://www.dovank.com/

Les kiosques des exposants, où vous trouverez différentes informations et pourrez bouquiner ou acheter, seront situés dans l'édifice Science Theatres, tout juste à côté des locaux où toutes les séances de présentations scientifiques auront lieu et où les pauses-café et les inscriptions se tiendront. Les exposants seront présents les lundi et mardi, 27 et 28 mai, de 9 h à 16 h.

- Pearson https://www.pearsonerpi.com/fr/
- Dovank http://www.dovank.com/

Career Connections • Connexion carrière

Career Connections and Networking Sessions will be held in conjunction with the 2019 SSC Annual Meeting in Calgary.

A panel discussion by 5-6 industry professionals will be followed by a roundtable networking event where each organization can host a table to connect with job seekers in small group conversations. There are multiple rounds of roundtable networking sessions, so each organization can connect with all job seekers. The Career Connections and Networking Sessions are being held on Monday, May 27 from 3-5pm in the Hunter Hub for Entrepreneurial Thinking, MacEwan Student Centre (MSC) 171, University of Calgary. Job Postings and Scheduled Interviews for approved Employers are also being facilitated by Career Services at the University of Calgary.

Se tiendront des sessions de connexion carrière et réseautage en liaison avec le Congrès annuel 2019 de la SSC à Calgary.

Une discussion de table ronde par 5-6 professionnels du secteur sera suivie d'une séance de réseautage en table ronde dont chaque organisation aura sa propre table où échanger en petits groupes avec des chercheurs d'emploi. Plusieurs séries de séances de réseautage sont prévues, afin que chaque organisation puisse se connecter avec tous les demandeurs d'emploi. Les sessions de connexion carrière et réseautage auront lieu le lundi 27 mai de 15h00 à 17h00 dans le Hunter Hub for Entrepreneurial Thinking, MacEwan Student Centre (MSC) 171, Université de Calgary. Les services d'affichages d'offres d'emploi et de planification d'entretiens pour les employeurs approuvés sont également animées par la Centre des services professionnels de l'Université de Calgary.

Organizers • Organisateurs

Local Arrangements Committee • Comité des arrangements locaux

- Karen Kopciuk and / et Alexander R. de Leon (Co-Chairs • Co-Présidents)
- Gemai Chen
- Jim Stallard
- Thuntida Ngamkham
- Thierry Chekouo

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.

Among other responsibilities, our student volunteers have handled AV support and the registration desk. At the SSC Office Miaclaire Woodland and Michelle Benoit have provided crucial support for many aspects, including registration and exhibitors.

We also are grateful for the assistance from Melissa Wrubleski, Administrative Manager in the Department of Mathematics and Statistics.

The SSC Meetings Coordinator Changbao Wu and other SSC executive members, and previous local arrangements chairs all shared their experience, offered useful advice and answered our numerous questions. Finally, Hugh Chipman and 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 l'envergure du congrès annuel de la SSC sans l'aide de nombreux individus et organismes. Le comité des arrangements locaux est très reconnaissant à tous ceux qui ont aidé à mettre sur pied cet événement.

Les étudiants bénévoles ont contribué, entre autres responsabilités, au soutien audiovisuel et au bureau des inscriptions. Au bureau de la SSC, Miaclaire Woodland et Michelle Benoit ont fourni un appui essentiel pour divers aspects, dont l'inscription et les exposants.

Nous sommes également reconnaissants pour l'aide de Melissa Wrubleski, Directrice administrative du Département de mathématiques et statistique.

Le coordonnateur des congrès Changbao Wu et les autres membres de l'exécutif de la SSC, ainsi que les autres anciens présidents des comités des arrangements locaux, ont partagé leurs expériences, offert des conseils utiles et répondu à nos nombreuses questions. Finalement, Hugh Chipman et Angelo Canty ont géré tous les services électroniques reliés au congrès et préparé la version PDF du programme.

Program Committee • Comité du programme

- Lisa Lix (Chair • Président) *University of Manitoba*
- Patrick Brown *University of Toronto*, Biostatistics - Biostatistique
- René Ferland *Université du Québec à Montréal*, Probability - Probabilité
- Chunfang Devon Lin *Queen's University*, Business and Industrial Statistics- Statistique industrielle et de gestion
- Susie Fortier *Statistics Canada • Statistique Canada*, Survey Methods - Méthodes d'enquête
- Asokan Variyath *Memorial University of Newfoundland*, Statistical Education - Éducation en statistique
- Jean-François Renaud *Université du Québec à Montréal*, Actuarial Science - Science actuarielle

General Information • Informations générales

Directions • Emplacement

Campus maps are available on the conference website, in the conference bag and on the back cover of this program. Campus maps and an interactive room finder is also accessible from https://www.ucalgary.ca/map/

The Welcome Reception is being held on the Main floor, Levels 2 and 3 of the Energy Environment Experiential Learning building (EEEL), 750 Campus Dr NW and the banquet at the Red and White Club at the south end of McMahon Stadium (1833 Crowchild Trail NW), about a 15 minute walk from Science Theatres. Head south on campus to the main entrance at University Dr., then south on University Drive NW until you reach the stadium.

Des cartes du campus sont accessibles sur le site web du congrès, dans le sac du congrès et en quatrième de couverture de ce programme. Des cartes du campus et un outil de recherche des salles sont également disponibles à https://www.ucalgary.ca/map/

La réception d'accueil se tiendra à l'étage principal, niveaux 2 et 3 de l'édifice Energy Environment Experiential Learning (EEEL), 750 promenade Campus NO, et le banquet au Red and White Club, à l'extrémité sud du stade McMahon (1833 sentier Crowchild NO), à environ 15 minutes de marche de l'édifice Science Theatres. Prenez au sud sur le campus vers l'entrée principale, promenade University, puis vers le sud sur la promenade University NO jusqu'au stade.

Registration • Inscription

Registered participants can pick up their badges and registration materials at the registration desk at the times and locations indicated below. Walk-ins should register online and present their receipt at the registration desk.

- Sunday May 26, 8:00am-5:00 pm, Science Theatres Lobby
- Sunday, May 26, 6:00pm-8:00pm, EEEL Lobby
- Monday, May 27, 7:45am-5:00pm, Science Theatres Lobby
- Tuesday, May 28, 8:00am-5:00pm, Science Theatres Lobby
- Wednesday, May 29, 8:15am-4:00pm, Science Theatres Lobby

Les participants déjà inscrits peuvent venir chercher leur insigne et leur matériel de congrès à l'endroit et aux heures indiqués ci-dessous. Ceux qui veulent s'inscrire au moment du congrès doivent le faire en ligne et présenter leur reçu au kiosque d'inscription.

- dimanche 26 mai, 8h00-17h00, Hall d'entrée Science Theatres
- dimanche 26 mai, 18h00-20h00, Hall d'entrée EEEL
- lundi 27 mai, 7h45-17h00, Hall d'entrée Science Theatres
- mardi, 28 mai, 8h00-17h00, Hall d'entrée Science Theatres
- mercredi 29 mai, 8h15-16h00, Hall d'entrée Science Theatres

Parking and Transportation • Stationnement sur le campus et transports

We encourage participants to take advantage of Calgary's public transit system. University Station on the CTrain Red Line is located on the east side of campus, quite close to Science Theatres. There are also several bus stops on campus, including a bus loop by the Education Block building. For information and trip planning assistance including a downloadable app, see the Calgary Transit website,
https://www.calgarytransit.com/getting-around/new-calgary

For those who will be driving, please see the annotated campus map on the back cover for suggested parking lots or more extensive information on University of Calgary visitor parking site https://bit.ly/2VeiCNu

Nous encourageons les participants à profiter du réseau de transport public de Calgary. La station University de la ligne rouge du CTrain est située à l'est du campus, tout près de l'édifice Science Theatres. Il y a aussi plusieurs arrêts de bus sur le campus, dont une boucle proche de l'édifice Education Block. Pour plus d'informations sur la planification de vos trajets (y compris une appli à télécharger), consultez le site Web de Calgary Transit,
https://www.calgarytransit.com/getting-around/new-calgary

Si vous devez venir en voiture, veuillez consulter la carte annotée du campus en quatrième de couverture, qui indique des aires de stationnement suggérées, ou la page d'information de l'Université de Calgary sur le stationnement visiteurs,
https://bit.ly/2VeiCNu

Campus Security • Sécurité sur le campus

- Fire/Police/Ambulance: 911
- Campus Security Services:
  - 403-220-5333

- Pompiers/Police/Ambulance : 911
- Force policière du campus :
  - 403-220-5333

Internet Access • Accès internet

The University of Calgary is a member of eduroam (EDUcation ROAMing), an authentication service allowing users (researchers, teachers, students, staff) from participating educational institutions to securely access the wireless network of any eduroam-enabled institution by using the same credentials they would use at their home institution. Connecting through eduroam provides basic network connectivity for web browsing (HTTP), secure shell (SSH) and VPN access. Visitors to the University of Calgary from eduroam participating institutions can access basic wireless services without having to obtain a University of Calgary Guest account. For more information, visit https://bit.ly/2Jf01iz

For those visitors who do not come from eduroam participating institutions, you can access the AirUC Guest Wireless Internet Access network. You will require a valid email address and telephone number (SMS enabled) in order to register; your account is active for 72 hours. More details may be found here: https://bit.ly/2VPNBDZ

L'Université de Calgary est membre de eduroam (EDUcation ROAMing), service d'authentification permettant aux utilisateurs (chercheurs, enseignants, étudiants, personnel) des établissements d'enseignement participants à accéder de manière sécurisée au réseau sans fil de toute institution eduroam à l'aide des mêmes informations d'identification qu'ils utiliseraient à leur institution d'origine. La connexion eduroam permet une connectivité réseau de base pour la navigation sur Internet (HTTP), le secure shell (SSH) et l'accès VPN. Les visiteurs à l'Université de Calgary des établissements d'enseignement participants peuvent accéder aux services sans fil de base sans avoir à demander un compte invité de l'Université de Calgary. Pour plus d'informations, consultez https://bit.ly/2Jf01iz

Les visiteurs qui ne viennent pas d'un établissement d'enseignement participant à eduroam pourront accéder au réseau sans fil d'internet d'AirUC Guest. Il vous faudra une adresse électronique et un numéro de téléphone compatible SMS pour vous inscrire ; votre compte sera actif pendant 72 heures. Vous trouverez plus de détails ici : https://bit.ly/2VPNBDZ

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

Please see the meeting site for some suggestions on or close to campus for lunch or stop by the registration desk for a printed list of nearby options.

Calgary has a vibrant restaurant scene, with options to suit every price-point and palate. For a wide range of suggestions, please see
https://www.visitcalgary.com/things-to-do#/eat--drink

Veuillez consulter le site du congrès pour des suggestions de restaurants sur le campus ou à proximité pour le repas de midi. Le bureau des inscriptions pourra également vous donner une liste écrite de recommandations.

Calgary se vante de nombreux restaurants, avec des options pour tous les budgets et tous les goûts. Pour plus d'informations, consultez
https://www.visitcalgary.com/things-to-do#/eat--drink

Athletics Facilities • Installations sportives

The University of Calgary offers extensive athletics facilities including gym, pool, climbing wall, etc. Guest day passes are available for $ 10.50 per day (plus taxes). See
https://bit.ly/2VvLPZ9 for more information.

L'Université de Calgary offre une gamme complète d'installations sportives : gymnase, piscine, mur d'escalade, etc. Des passes de jour sont disponibles au prix de 10,50 $ (plus taxes). Pour plus de détails, consultez https://bit.ly/2VvLPZ9

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 Géraldine Lo Siou) by the translators Caroline Gras, Catherine Cox, Michelle Blaquière, Caroline Petit-Turcotte 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 Géraldine Lo Siou) par les traducteurs Caroline Gras, Catherine Cox, Michelle Blaquière, Caroline Petit-Turcotte 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.

Rooms • Salles

Poster presentations will take place in the plus-15 walkway between Science Theatres and Biological Sciences on Monday and Tuesday afternoons. Exhibits will take place in the Science Theatres area. Coffee breaks will take place outside of Science Theatres 148, by the Zipper, and close to the scientific sessions held in this and nearby buildings.

Plenary sessions on Monday and Tuesday mornings will take place in the Science Theatre 148.

Les présentations par affichage se tiendront dans le couloir plus-15 entre les édifices Science Theatres et les Sciences biologiques les lundi et mardi après-midis. Les expositions auront lieu dans l'édifice Sciences Theatres. Les pauses-café se tiendront devant la salle Science Theatres 148, près du Zipper, et près des sessions scientifiques qui se tiennent dans cet édifice et d'autres à proximité.

Les séances plénières des lundi et mardi matins auront lieu dans la salle Science Theatres 148.

Poster Sessions • Séances d'affichage

Contributed posters and Case Study posters will be displayed in the plus-15 walkway between Science Theatres and Biological Sciences with Case Study posters being displayed Monday afternoon and contributed research posters appearing on Tuesday afternoon. In both cases, posters will be displayed between noon and 5:00 pm, the authors being with their posters from 1:30 pm until the end of coffee break at 3:30 pm.

Les séances d'affiches libres et d'études de cas se feront dans le couloir plus-15 entre les édifices Science Theatres et les Sciences biologiques, les affiches d'études de cas étant exposées le lundi après-midi et les affiches libres exposées le mardi après-midi. Dans les deux cas, les affiches seront exposées de midi à 17h00, les auteurs devant être présents avec leur affiche de 13h30 jusqu'à la fin de la pause-café à 15h30.

Workshops • Ateliers

Sunday, May 26, 9:00 am - 4:00 pm • Dimanche 26 mai, 9h - 16h ..... Various buildings • divers bâtiments

Workshops organized by the sections will be held Sunday 9:00 am to 4:00 pm unless otherwise noted in rooms in Science B, Science A, Math Sciences and the Taylor Institute for Teaching and Learning. Coffee breaks and lunch will take place in room ST 142 (Collaborative Space) from 10:15 to 10:45 am and 2:15 to 2:45 pm. The lunch break will be between noon and 1:00 pm.

Les ateliers organisés par les groupes auront lieu le dimanche de 9h00 à 16h00, sauf indication contraire, dans les salles des édifices Science B, Science A, Math Sciences et Taylor Institute for Teaching and Learning. Les pauses-café et repas de midi se feront dans la salle ST 142 (espace de collaboration), de 10h15 à 10h45 et de 14h15 à 14h45. La pause pour le repas de midi sera entre midi et 13h.

Other Meetings • Autres réunions

NSERC – Updates and 2018 Competition Results: Sunday 5:00 - 6:00 pm, in Science B 103.

Committee meetings: Lunchtime business meetings will occur in the Math Sciences Building in seminar or classrooms that are not used for the scientific sessions. Business meeting participants should pick up their boxed lunches in the Department of Mathematics and Statistics Lounge (MS 461).

NSERC Discovery Grant Application Assistance:
Monday 12:00 - 1:30 pm, in MS 325.
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.

SSC Annual General Meeting: Monday 5:00 - 6:00 pm, in ICT 122.

Section AGM's: Tuesday 5:00 - 6:00 pm, in various conference meeting rooms.

CRSNG – Mise à jour et résultats du concours de 2018 : dimanche 17h - 18h, salle Science B 103.

Réunions de comités : Les réunions d'affaires ayant lieu à l'heure du repas de midi se tiendront dans l'édifice Math Sciences, dans les salles de séminaire ou dans des salles de classes qui ne sont pas utilisées pour les sessions scientifiques. Les participants à ces réunions sont invités à prendre une boîte à lunch dans le salon du Département de mathématiques et statistique (MS 461).

Atelier d'assistance à la demande de subvention à la découverte du CRSNG : lundi 12h00 - 13h30, salle MS 325
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.

Assemblée générale annuelle : lundi 17h - 18h, salle ICT 122.

AGM des groupes : mardi 17h - 18h, dans diverses salles de réunion.

Social Events • Activités sociales

Welcome Reception • Réception de bienvenue

Sunday, May 26, 6:00 - 8:00 pm • Dimanche 26 mai, 18h00 - 20h00 ..... Energy Environment Experiential Learning building (EEEL)

The Welcome Reception will be held on Main floor, Levels 2 and 3 of the Energy Environment Experiential Learning building (EEEL), 750 Campus Dr NW. All conference attendees are welcome to join us to share a few drinks and some appetizers in good company. One drink ticket will be given to all registrants for the reception in their registration materials. For those who will be arriving directly from the workshops it is a five minute walk from the workshop meeting rooms.

La réception de bienvenue aura lieu à l'étage principal, niveaux 2 et 3 de l'édifice Energy Environment Experiential Learning (EEEL), 750 promenade Campus NO. Tous les congressistes sont bienvenus et sont invités à venir partager quelques verres et d'excellentes bouchées en bonne compagnie. Vous trouverez un ticket boisson dans vos documents d'inscription. Pour ceux qui arrivent directement des ateliers, la réception est à cinq minutes de marche des salles de réunion.

Barbeque • Barbecue

Monday, May 27 6:00 - 8:00 pm • Lundi 27 mai, 18h00 - 20h00 ..... Cassio A,B (MH)

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 in their registration materials. The BBQ will take place at Cassio AB in the MacEwan Hall (second level).

Le barbecue étudiant est gratuit pour les étudiants du premier cycle et des cycles supérieurs ; l'inscription est requise. Si vous êtes inscrit au barbecue, vous trouverez un ticket dans vos documents d'inscription. Le barbecue aura lieu à la salle Cassio AB du Centre des étudiants MacEwan (deuxième étage).

Banquet

Tuesday, May 28, 6:15 pm • Mardi 28 mai, 18h15 ..... Red and White Club, McMahon Stadium

The banquet will be held at the Red and White Club in the south end of McMahon Stadium with cocktails and a cash bar at 6:15 pm and dinner service beginning around 7:00 pm. All conference participants who have selected a meal for the banquet will find one banquet ticket with their selected meal in their registration envelope. You are asked to place the banquet ticket on the table in front of you when your table is served. 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 27 - this will help us to better plan for attendance at the banquet.

The Red and White Club is located just south of campus along University Drive NW. and can be reached heading south from the conference meeting rooms to the main entrance to the University of Calgary, then south along University Drive NW. It is located approximately 1.5 km from Science Theatres on the University of Calgary campus (a 15 minute walk).

Le banquet se tiendra au Red and White Club, à l'extrémité sud du stade McMahon, avec cocktails et bar payant à partir de 18h15, suivi du souper servi à table à partir de 19h. Les participants au congrès qui ont sélectionné un repas pour le banquet trouveront un ticket correspondant dans leur enveloppe d'inscription. Nous vous demanderons de placer ce dernier devant vous à la table lorsque votre table sera servie. Si vous avez un ticket mais que vous ne souhaitez plus participer au banquet, veuillez remettre votre ticket au bureau des inscriptions, au plus tard mardi 27 mai à midi – cela nous permettra de mieux planifier les effectifs au banquet.

Le Red and White Club est situé juste au sud du campus, via la promenade University NO. À partir des salles de réunion du congrès, dirigez-vous au sud vers l'entrée principale du campus, puis vers le sud le long de la promenade University NO jusqu'au stade. Il est situé à environ 1,5 km (15 minutes de marche) de l'édifice Science Theatres du campus de l'Université de Calgary.

Other Social Events • Autres évènements sociaux

University of Waterloo Networking Reception in Calgary
Monday, May 27, 5:30 - 7:30 pm, Bianca Room MacEwan Hall 226.

New Investigators Social Gathering
Monday May 27th, 6:00pm onward, at the Last Defence Lounge, University of Calgary.

Réception de réseautage de l'Université de Waterloo à Calgary
Lundi 27 mai, 17h30 - 19h30, salle Bianca du MacEwan Hall 226.

Rencontre sociale des nouveaux chercheurs
Lundi 27 mai à partir de 18h00 au Last Defence Lounge, Université de Calgary.

Committees and Meetings • Comités et réunions

Saturday May 25 — samedi 25 mai

12:00-18:00 — **478 (MS)**
SSC Executive Committee /Comité exécutif de la SSC

14:00-15:00 — **431 (MS)**
CANSSI Meetings/Réunions INCASS

15:00-16:00 — **431 (MS)**
CANSSI Annual General Meeting/Assemblée générale annuelle de l'INCASS

16:00-17:00 — **431 (MS)**
CANSSI Meetings/Réunions INCASS

Sunday May 26 — dimanche 26 mai

09:30-11:00 — **431 (MS)**
Fundraising Committee/Comité de collecte de fonds

11:00-17:00 — **Escalus - 236 (MH)**
SSC Board of Directors/Conseil d'administration de la SSC

14:30-17:00 — **427 (MS)**
Department Heads Meeting/Réunion des chefs de département

16:00-18:00 — **522 (MS)**
CANSSI CRT 11: Spatial Modeling of Infectious Diseases/INCASS PRC Modélisation spatiale des maladies infectieuses

17:00-18:00 — **103 (SB)**
NSERC(2018 Competition) -results/CRSNG(résultats du concours de 2018)

Monday May 27 — lundi 27 mai

12:00-13:30 — **337 (MS)**
Actuarial Science Section Executive Committee/Comité exécutif du Groupe de science actuarielle

12:00-13:30 — **431 (MS)**
Biostatistics Section Executive Committee/Comité exécutif du Groupe de biostatistique

12:00-13:30 — **452 (MS)**
BISS Executive Committee/Comité exécutif du Groupe de statistique industrielle et de gestion

12:00-13:30 **522 (MS)**
Probability Section Executive Committee/Comité exécutif du Groupe de probabilité

12:00-13:30 **478 (MS)**
Statistical Education Section Executive Committee/Comité exécutif du Groupe d'éducation en statistique

12:00-13:30 **569 (MS)**
Committee on Women in Statistics/Comité sur les femmes en statistique

12:00-13:30 **371 (MS)**
Finance Committee/Comité des finances

12:00-13:30 **427 (MS)**
Ad hoc Accreditation Services Committee/Comité ad hoc des services d'accréditation

12:00-18:00 **249 (SA)**
Award for Case Studies in Data Analysis Committee/Prix pour les études de cas en l'analyse de données

12:30-13:30 **325 (MS)**
NSERC Discovery Grant Application Workshop/Atelier subventions du CRSNG

15:00-17:00 **Hunter Hub for Entrepreneurial Thinking, MSC 171 (MSC)**
Career Connections & Networking Sessions/Sessions de connexion carrière et réseautage

17:00-18:00 **122 (ICT)**
SSC AGM/AGA de la SSC

Tuesday May 28 mardi 28 mai

07:30-08:30 **452 (MS)**
Ad hoc Financial Procedures Committee/Comité ad hoc des procédures financières

12:00-18:00 **249 (SA)**
Student Research Presentation Award Committee/Comité du prix pour les présentations de recherche étudiantes

12:00-13:30 **427 (MS)**
Accreditation Committee AGM/AGA du Comité d'accréditation

12:00-13:30 **452 (MS)**
Census @ School Committee/Comité de Recensement à l'école Canada

12:00-13:30 **431 (MS)**
CJS Editorial Board/Conseil de rédaction de la RCS

12:00-13:30 **478 (MS)**
Statistics Education Committee/Comité d'éducation en statistique

12:00-13:30 **569 (MS)**
Office Committee /Comité du bureau

12:00-13:30 **371 (MS)**
CANSSI Health Sciences Collaborating Centres Leaders/INCASS Responsables des centres de collaboration des sciences de la santé

17:00-18:00 **105 (SB)**
Actuarial Science Section AGM /AGA du Groupe de science actuarielle

17:00-18:00 **143 (ST)**
Biostatistics Section AGM/AGA du Groupe de biostatistique

17:00-18:00 **116 (ICT)**
BISS AGM/AGA du GSIG

17:00-18:00 **109 (SS)**
Probability Section AGM/AGA du Groupe de probabilité

17:00-18:00 **146 (SB)**
Statistical Education Section AGM/AGA du Groupe d'éducation en statistique

17:00-18:00 **142 (AD)**
Survey Methods Section AGM/AGA du Groupe des méthodes d'enquête

Wednesday May 29 mercredi 29 mai

12:00-13:30 **371 (MS)**
Accreditation Committee/Comité d'accréditation

12:00-13:30 **325 (MS)**
New Investigators Committee/Comité des nouveaux chercheurs

12:00-13:30 **365 (MS)**
Program Committee/Comité des programmes scientifiques

12:00-13:30 **569 (MS)**
Student and Recent Graduate Committee/Comité des étudiants et diplômés récents

12:00-13:30 **427 (MS)**
Treasurers Committee/Comité des trésoriers

12:00-13:30 **337 (MS)**
Actuarial Science Section Executive Committee/Comité exécutif du Groupe de science actuarielle

12:00-13:30 **431 (MS)**
Biostatistics Section Executive Committee/Comité exécutif du Groupe de biostatistique

12:00-13:30 **452 (MS)**
BISS Executive Committee/Comité exécutif du Groupe de statistique industrielle et de gestion

12:00-13:30 **522 (MS)**
Probability Section Executive Committee/Comité exécutif du Groupe de probabilité

12:00-13:30 **478 (MS)**
Statistical Education Section Executive Committee/Comité exécutif du Groupe d'éducation en statistique

16:45-18:00 **540B (Science Boardroom) (BI)**
SSC Board of Directors/Conseil d'administration de la SSC

18:00-19:00 **478 (MS)**
SSC Executive Committee /Comité exécutif de la SSC

Social and Information Events • Événements sociaux et informatifs

Sunday May 26 — dimanche 26 mai

18:00-20:00 — **Main floor, Levels 2 and 3 (EEEL)**
Welcoming Reception/Réception de bienvenue

Monday May 27 — lundi 27 mai

09:50-10:20 — **148 (ST)**
Coffee break/Pause-café

15:00-15:30 — **148 (ST)**
Coffee break/Pause-café

17:30-19:30 — **Bianca Room – 226 (MH)**
Waterloo Alumni Reception/Réception des anciens de Waterloo

18:00-20:00 — **Cassio AB (MH)**
Student BBQ/BBQ étudiant

18:00-18:00 — **The Last Defence Lounge (MSC)**
New Investigator Social Gathering/Rassemblement social des étudiants et nouveaux chercheurs

Tuesday May 28 — mardi 28 mai

09:50-10:20 — **148 (ST)**
Coffee break/Pause-café

15:00-15:30 — **148 (ST)**
Coffee break/Pause-café

18:15-22:00 — **The Red and White Club**
Banquet/Banquet

Wednesday May 29 — mercredi 29 mai

09:50-10:20 — **148 (ST)**
Coffee break/Pause-café

15:00-15:30 — **148 (ST)**
Coffee break/Pause-café

Program • Programme

Sunday May 26 — dimanche 26 mai

09:00-16:00 **Invited / Sur invitation** (abstract/résumé **??**) **569 (MS)**

Survey Methods Workshop
Atelier du Groupe méthodes d'enquête

Sponsor/Commanditaires: SSC Survey Methods Section / Groupe méthodes d'enquête de la SSC

09:00-16:00 **David Haziza** (Université de Montréal)
When Some Data are Missing: Foundations of Imputation Theory and Recent Developments / Lorsque certaines données sont manquantes : fondements de la théorie de l'imputation et développements récents

---

09:00-16:00 **Invited / Sur invitation** (abstract/résumé **??**) **110 (TI)**

Statistics Education Workshop
Atelier du Groupe d'éducation en statistique

Sponsor/Commanditaires: SSC Statistical Education Section / Groupe d'éducation en statistique de la SSC

09:00-16:00 **Natasha Kenny** (University of Calgary), **Bruce Dunham** (University of British Columbia), **Jim Stallard** (University of Calgary)
Developing a Teaching Portfolio / Développer un portfolio d'enseignement

---

09:00-16:00 **Invited / Sur invitation** (abstract/résumé **??**) **144 (SB)**

Business and Industrial Statistics Workshop
Atelier du Groupe de statistique industrielle et de gestion

Sponsor/Commanditaires: SSC Business and Industrial Statistics Section / Groupe de statistique industrielle et de gestion de la SSC

09:00-16:00 **Jiguo Cao** (Simon Fraser University)
Functional Data Analysis for Big Data / Analyse de données fonctionnelles pour les mégadonnées (Big Data)

09:00-16:00 **Invited / Sur invitation** (abstract/résumé **??**) **146 (SB)**

Biostatistics Workshop
Atelier du Groupe de biostatistique

Sponsor/Commanditaires: SSC Biostatistics Section / Groupe de biostatistique de la SSC

09:00-16:00 **John Martin Bland** (University of York)
Lessons from a life in applied statistics / Leçons d'une vie consacrée à la statistique appliquée

---

09:00-12:00 **Invited / Sur invitation** (abstract/résumé **??**) **103 (SB)**

Accreditation Committee Workshop
Atelier du Comité d'accréditation

Sponsor/Commanditaires: SSC Accreditation Committee / Comité d'accréditation de la SSC

09:00-12:00 **Janet Elizabeth McDougall** (McDougall Scientific Ltd.), **Rhonda Rosychuk** (University of Alberta), **Darcy Pickard** (ESSA Technologies Ltd), **Khalid Lemzouji** (StatVis analystics and DS4G)
Making a Living as an Accredited Statistician: Tips from the Frontlines of Experienced, Successful Accredited Consulting Statisticians / Gagner sa vie en tant que statisticien agréé : Conseils de première ligne de statisticiens-conseil agréés avec expérience et ayant du succès

Monday May 27 lundi 27 mai

08:30-09:50 **Invited / Sur invitation** (abstract/résumé 60) **148 (ST)**

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

Chair/Président: Robert Platt
Organizer/Responsable: Robert Platt

08:30-09:50 **Sylvia Richardson** (University of Cambridge)
Statistical Challenges in the Analysis of Complex Phenotypes in Biomedicine / Défis statistiques relatifs à l'analyse de phénotypes complexes en biomédecine

---

10:20-11:50 **Invited / Sur invitation** (abstract/résumé 61) **144 (SB)**

Capital Allocation
Affectation de capitaux

Chair/Président: Hélène Cossette
Organizer/Responsable: Hélène Cossette
Sponsor/Commanditaires: SSC Actuarial Science Section / Groupe de science actuarielle de la SSC

10:20-10:50 **David Saunders** (University of Waterloo), **Dan Rosen** (d1g1t Inc.)
Conditional Simulation of Risk Factor Distributions for Stress Testing and Capital Allocation / Simulation conditionnelle des distributions des facteurs de risque pour les simulations de crise et l'imputation sur les fonds propres

10:50-11:20 **Edward Furman** (York University), **Alexey Kuznetsov** (York University), **Justin Miles** (York University)
General Risk Aggregation: Is Gamma the New Normal? / Agrégation générale des risques : Gamma représente-t-il la nouvelle norme ?

11:20-11:50 **Mélina Mailhot** (Concordia University)
Capital Allocation under New Canadian Regulations for P&C Insurers / Allocation de capital basée sur la nouvelle règlementation pour les compagnies d'assurance canadiennes

---

10:20-11:50 **Invited / Sur invitation** (abstract/résumé 63) **113 (SS)**

Applications of Nonstandard Analysis to Probability Theory and Statistics
Applications de l'analyse non standard à la théorie des probabilités et à la statistique

Chair/Président: Daniel M. Roy
Organizer/Responsable: Daniel M. Roy
Sponsor/Commanditaires: SSC Probability Section / Groupe de probabilité de la SSC

10:20-10:50 **Haosui Duanmu** (University of Toronto), **Robert Anderson** (University of California, Berkeley), **Aaron Smith** (University of Ottawa)
Mixing Times and Hitting Times of Markov Processes via Nonstandard Analysis / Temps de mélange et temps d'atteinte des processus de Markov au moyen d'une analyse non standard

10:50-11:20 **Peter A Loeb** (University of Illinois, Urbana-Champaign)
Nonstandard Analysis and Boundaries / Analyse non standard et limites

11:20-11:50 **Robert Anderson** (University of California at Berkeley), **Roberto Raimondo** (University of Melbourne)
Hyperfinite Existence Results Imply Convergence Results / Les constructions hyperfinies impliquent des théorèmes de convergence

10:20-11:50 **Invited / Sur invitation** (abstract/résumé 65) **102 (ICT)**

Innovations in Data Science for Undergraduates in Canada
Innovation en science des données pour les étudiants de premier cycle au Canada

Chair/Président: Bruce Dunham
Organizer/Responsable: Bruce Dunham
Sponsor/Commanditaires: SSC Statistical Education Section / Groupe d'éducation en statistique de la SSC

10:20-10:50 **Tiffany A Timbers** (University of British Columbia)
Teaching an "Introduction to Data Science" - A Discussion of Course Design Intent and Lessons Learned / Enseigner une «introduction aux sciences des données» : une discussion relative à l'élaboration de cours et aux leçons apprises

10:50-11:20 **Xu (Sunny) Wang** (Wilfrid Laurier University)
How to Design an Undergraduate Data Science Program to Meet the Needs of Modern Business and Industry? / Comment concevoir un programme de premier cycle en sciences des données pour répondre aux besoins des entreprises et de l'industrie modernes ?

11:20-11:50 **Jim Stallard** (University of Calgary)
Reflections in Building a Data Science Program / Réflexions sur l'élaboration d'un programme en science des données

---

10:20-11:50 **Invited / Sur invitation** (abstract/résumé 67) **201 (ENA)**

Novel Statistical Methods and Applications in Genomics
Nouvelles méthodes statistiques et applications à la génomique

Chair/Président: Mireille E. Schnitzer
Organizer/Responsable: Mireille E. Schnitzer
Sponsor/Commanditaires: Committee on Women in Statistics / Comité sur les femmes en statistique

10:20-10:40 **Jinko Graham** (Simon Fraser University)
Relatedness, Inherited Traits and Trait-Influencing DNA Variants / Lien de parenté, caractères héréditaires et variants d'ADN influençant les caractères

10:40-11:00 **Marie-Pierre Sylvestre** (Université de Montréal), **Angelo Canty** (McMaster University), **Shelley Bull** (University of Toronto), **Paterson Andrew** (University of Toronto), **Laurence Boulanger** (CRCHUM)
Methods for Genomewide Analysis of Complex Phenotypes / Méthodes d'analyse de phénotypes complexes à l'échelle du génome

11:00-11:20 **Jingjing Wu** (University of Calgary), **Tasnima Abedin** (Alberta Health Services)
A Mixture Model under Stochastic Dominance Constraint for Genetic Studies / Modèle de mélange avec contrainte de dominance stochastique pour études génétiques

11:20-11:40 **Ting-Huei Chen** (Laval University), **Hanaa Boughal** (Laval University)
A Powerful Approach to Identify the Genetic Variables Associated with Psychiatric Diseases / Une approche efficace pour l'identification de variables génétiques associées aux maladies psychiatriques

---

10:20-11:50 **Invited / Sur invitation** (abstract/résumé 70) **122 (ICT)**

Recent Developments in Survival Analysis with Complex Data
Récentes évolutions en analyse de survie avec données complexes

Chair/Président: Xuewen Lu
Organizer/Responsable: Xuewen Lu
Sponsor/Commanditaires: SSC Biostatistics Section / Groupe de biostatistique de la SSC

10:20-10:50 **Menggang Yu** (University of Wisconsin)
Cox Regression with Nonignorable Survival-Time-Dependent Missing Covariate Values / Régression de Cox avec des valeurs de covariables manquantes non ignorables dépendantes du temps de survie

10:50-11:20 **Gang Li** (University of California, Los Angeles), **Eric Kawaguchi** (University of California Los Angeles), **Marc Suchard** (University of California Los Angeles), **Zhenqiu Liu** (Penn State University)
Scalable Sparse Cox's Regression for Large-Scale Survival Data via Broken Adaptive Ridge / Régression de Cox creuse et échelonable pour des données de survie à grande échelle au moyen d'un ridge adaptatif brisé

11:20-11:50 **Zhigang Li** (University of Florida), **Janaka Peragaswaththe Liyanage** (University of Florida), **Lihui Zhao** (Northwestern University)
Joint Modeling of Survival and Longitudinal Quality of Life Data with Informative Censoring in Palliative Care Studies / Modélisation conjointe des données de survie et des données longitudinales sur la qualité de vie avec censure informative dans les études sur les soins palliatifs

---

10:20-11:50 **Contributed / Communications libres** (abstract/résumé 72) **142 (AD)**

Statistics, the Environment, and Ecology
Statistique, environnement et écologie

Chair/Président: Quan Long

10:20-10:35 **John R.J. Thompson** (University of Western Ontario)
Estimating Fire Spread Rates from Micro-Fire Experiments / Estimation des taux de propagation du feu à partir d'expériences de micro-feux

10:35-10:50 **Nan Zheng** (Marine Institute of Memorial University of Newfoundland), **Noel Cadigan** (Centre for Fisheries Ecosystems Research, Fisheries and Marine Institute of Memorial University of Newfoundland), **Joanne Morgan** (Fisheries and Oceans Canada)
A Spatiotemporal Von Bertalanffy Growth Model and Its Estimation When Data Are Collected Through Length-Stratified Sampling / Un modèle spatiotemporel de croissance de Von Bertalanffy et son estimation lorsque les données sont recueillies par échantillonnage stratifié dans la longueur

10:50-11:05 **Rolf Turner** (University of Auckland), **Kate Richards** (New Zealand Plant and Food Reseach Institute)
Adventures with Bark Beetles / Aventures avec les scolytes

11:05-11:20 **Matthew R. Parker** (University of Victoria), **Vivian Pattison** (University of Victoria), **Laura L.E. Cowen** (University of Victoria)
Estimating Population Abundance Using Counts from an Auxiliary Population and N-Mixture Models / Estimation de l'abondance d'une population à l'aide de nombres d'une population auxiliaire et de modèles de mélange N

11:20-11:35 **Guowen Huang** (Centre for Global Health Research), **Patrick Brown** (St. Michael's Hospital; University of Toronto)
Daily Mortality and Air Quality: Using Multivariate Time Series with Seasonally-Varying Covariances / Mortalité quotidienne et qualité de l'air : utilisation d'une série temporelle multivariée avec des covariances variables selon les saisons

11:35-11:50 **Inesh Munaweera** (University of Manitoba), **Saman Muthukumarana** (University of Manitoba), **Darren Gillis** (University of Manitoba), **Douglas Watkinson** (Fisheries and Oceans Canada), **Colin Charles** (Fisheries and Oceans Canada)
Understanding Lake Winnipeg Basin Walleye Fish Movement Patterns Using Bayesian State-Space Models / Analyse des schémas de mouvement des dorés jaune du bassin du lac Winnipeg par modèles espace-état bayésiens

---

10:20-11:50 **Contributed / Communications libres** (abstract/résumé 76) **101 (ENA)**

Design and Analysis Approaches for Complex Survey Data
Approches de conception et d'analyse pour des données d'enquête complexes

Chair/Président: Shamsia Sobhan

10:20-10:35 **Tristan Watson** (ICES), **Laura Rosella** (University of Toronto), **Kathy Kornas** (University of Toronto), **Catherine Bornbaum** (University of Toronto)
Age-Standardizing Prevalence Estimates from Combined Cycles of the Canadian Community Health Survey / Estimations de la prévalence normalisées selon l'âge à partir d'une combinaison de cycles de l'Enquête sur la santé dans les collectivités canadiennes Ⓔ E

10:35-10:50 **Joshua Gutoskie** (Statistics Canada)
Simulating Survey Design Weights to Account for Sample Coordination Between Surveys / Pondérations simulées de conception d'enquête pour prendre en compte la coordination échantillonnale entre les questionnaires Ⓔ E

10:50-11:05 **Vanessa McNealis** (Université de Montréal), **Christian Léger** (Université de Montréal)
Smoothed Bootstrap Estimator of the Variance of a Quantile Estimator in a Finite Population Context / Estimation de bootstrap lissé de la variance d'un estimateur de quantile dans le contexte d'une population finie Ⓔ E

11:05-11:20 **Yuxiang Gao** (University of Toronto), **Lauren Kennedy** (Columbia University), **Daniel Simpson** (University of Toronto)
Incorporating Structured Priors into Multilevel Regression and Poststratification / Intégrer des lois a priori structurées dans la régression multi-niveaux et la post-stratification Ⓔ E

11:20-11:35 **Yilin Chen** (University of Waterloo), **Changbao Wu** (University of Waterloo), **Pengfei Li** (University of Waterloo)
Estimation of Population Proportions with Non-Probability Survey Samples / Estimation des proportions de la population à l'aide d'échantillons d'enquêtes non probabilistes Ⓔ E

11:35-11:50 **Jules J. S. de Tibeiro** (Université de Moncton), **Filipe Afonso** (Symbad, Symbolic Data Lab), **Edwin Diday** (Paris Dauphine University)
Analyzing Aggregated Household Consumption with Symbolic Data Analysis / Analyse de la consommation des ménages agrégée par analyse de données symboliques Ⓕ E

10:20-11:50 **Contributed / Communications libres** (abstract/résumé 80) **109 (SS)**

Developments in Statistical Theory
Développements en théorie statistique

Chair/Président: John Joseph Koval

10:20-10:35 **François A Marshall** (Queen's University)
A Statistical Characterization of Multitaper Statistical Detectors using Monte Carlo / Une caractérisation statistique des détecteurs statistiques ¡¡multitaper¿¿ à l'aide de Monte Carlo Ⓔ E

10:35-10:50 **Feiyu Zhu** (University of Waterloo), **Martin Lysy** (University of Waterloo)
Particle Physics Representation of a Continuous Stationary Gaussian Process / Représentation en physique des particules d'un processus gaussien stationnaire continu Ⓔ E

10:50-11:05 **Armin Hatefi** (Memorial University of Newfoundland), **Mohammad Jafari Jozani** (University of Manitoba), **Omer Ozturk** (Ohio State University)
Finite Mixture Modeling Based on Multi-Observer Sampling Design / Modèles de mélange finis fondés sur un plan d'échantillonnage d'ensembles ordonnés multi-observateurs Ⓔ E

11:05-11:20 **Jeffrey D. Picka** (University of New Brunswick)
Ontological Aspects of Statistical Modelling / Aspects ontologiques de la modélisation statistique Ⓔ E

11:20-11:35 **Anthony Coache** (Université du Québec à Montréal), **François Watier** (Université du Québec à Montréal)
Stochastic Algorithms for Solving a Multi-Period Quantile-Based Portfolio Optimization Problem / Algorithmes stochastiques pour résoudre un problème d'optimisation multi-périodique de portefeuille basé sur un quantile

11:35-11:50 **Nirodha Mihirani Epasinghege Dona** (University of Manitoba), **Brad Johnson** (University of Manitoba)
Estimating Random Walk Centrality / Estimation de la centralité par marche aléatoire

---

12:00-17:00 **Poster / Poster** (abstract/résumé 83) **103Z (ST)**

Case Study 1: Counting Cells from Microscopic Images
Étude de cas 1 : Comptage de cellules dans des images microscopiques

Chair/Président: Pingzhao Hu

12:00-17:00 **Jiani Heng** (Queen's University), **Xinyi Ge** (Queen's University), **Na Li** (Queen's University), **Qianhui Yu** (Queen's University)
Queen's University / Queen's University

12:00-17:00 **Colin Weaver** (University of Calgary), **Ali Naqvi** (University of Calgary), **Mark Lowerison** (University of Calgary), **David Schulz** (University of Calgary)
University of Calgary / University of Calgary

12:00-17:00 **Xin Ding** (University of British Columbia), **Qiong Zhang** (University of British Columbia)
University of British Columbia / University of British Columbia

12:00-17:00 **Dylan Loader** (University of Calgary), **Daniel Yang** (University of Calgary), **Mingkuan Wu** (University of Calgary), **Andrew Pohl** (University of Calgary)
University of Calgary / University of Calgary

12:00-17:00 **Scott White** (The University of Manitoba), **Adeola Adegoke** (University of Manitoba), **Margaret Pecku** (University of Manitoba), **Bowei Yang** (University of Manitoba), **Zimo Zhu** (University of Manitoba)
University of Manitoba / University of Manitoba

12:00-17:00 **Jingyu Wang** (University of Manitoba), **Isuru Dharmasena** (University of Manitoba), **Shanika Basnayake** (University of Manitoba), **Sachithra Opathalage** (University of Manitoba), **Azizur Rahman** (University of Manitoba)
University of Manitoba / University of Manitoba

12:00-17:00 **Yunjing Li** (University of Toronto), **Leif Erik Lovblom** (University of Toronto), **Hyejung Jung** (University of Toronto), **Faizan Mohsin** (University of Toronto), **Kai Zhang** (University of Toronto), **Ling Lin** (University of Toronto)
University of Toronto / University of Toronto

12:00-17:00 **Hua Lu** (University of Toronto), **Xiande Yang** (University of Toronto), **Fangming Liao** (University of Toronto), **Lisu Zhang** (University of Toronto), **Yunfang Liu** (University of Toronto)
University of Toronto / University of Toronto

12:00-17:00 **Jingyi Yan** (University of Alberta), **Matthew Pietrosanu** (University of Alberta), **Wei Tu** (University of Alberta), **Jiaxin Zhang** (University of Alberta), **Yue Wang** (University of Alberta)
University of Alberta / University of Alberta

12:00-17:00 **Larry Dong** (McGill University), **Peter Park** (McGill University), **Zhiyue Zhang** (McGill University)
McGill University / McGill University

12:00-17:00 **Poster / Poster** (abstract/résumé 84) **103Z (ST)**

Case Study 2: Risk of Cardiovascular Disease among Osteoarthritis Patients: Exploring the Relationship in a National Health Survey
Étude de cas 2 : Risque de malade cardiovasculaire chez les patients souffrant d'arthrose : étude de la relation dans une enquête nationale sur la santé

Chair/Président: Pingzhao Hu

12:00-17:00 **Mohammed Mujaab Kamso** (University of Calgary), **Mili Roy** (University of Calgary), **Mubasiru Lamidi** (University of Calgary), **Meng Wang** (University of Calgary)
University of Calgary / University of Calgary

12:00-17:00 **Dominik Zhongda Yang** (McGill University), **Haoyu Wu** (McGill University)
McGill University / McGill University

12:00-17:00 **Matthew Berkowitz** (Simon Fraser University), **Coco Liu** (Simon Fraser University), **Barinder Thind** (Simon Fraser University), **Jiahao Tian** (Simon Fraser University)
Simon Fraser University / Simon Fraser University

12:00-17:00 **Christina Zihan Zhou** (University of Toronto), **Song Pham** (University of Toronto), **Dongyang Yang** (University of Toronto), **Hana Dampf** (University of Toronto)
University of Toronto / University of Toronto

12:00-17:00 **Shamsia Sobhan** (University of Manitoba), **Md Erfanul Hoque** (University of Manitoba), **Olawale Ayilara** (University of Manitoba), **Naomi Hamm** (University of Manitoba)
University of Manitoba / University of Manitoba

12:00-17:00 **Qiongbin Wang** (University of Toronto), **Yanni Zeng** (University of Toronto), **Xiayi Ma** (University of Toronto), **Yidi Jiang** (University of Toronto), **Yan Chen** (University of Toronto), **Yang Zhu** (University of Toronto)
University of Toronto / University of Toronto

13:30-15:00 **Invited / Sur invitation** (abstract/résumé 85) **144 (SB)**

Stochastic Processes and Applications
Processus et applications stochastiques

Chair/Président: Mary E. Thompson
Organizer/Responsable: Mary E. Thompson
Sponsor/Commanditaires: SSC Probability Section / Groupe de probabilité de la SSC

13:30-14:00 **Lam Ho** (Dalhousie University), **Vu Dinh** (University of Delaware), **Frederick A. Matsen** (Fred Hutchinson Cancer Research Center), **Marc Suchard** (University of California, Los Angeles)
On the Consistency of the MLE for the Transition Rate of a Two-State Symmetric Markov Process on a Tree / La convergence des estimateurs du maximum de vraisemblance pour le taux de transition d'un processus de Markov symétrique à deux états sur un arbre

14:00-14:30 **Hélène Guérin** (Université du Québec à Montréal), **Ninon Fétique** (Université de Tours), **Florent Malrieu** (Université de Tours)
Long Time Behavior of Interacting Zig-Zag Particles / Le comportement en temps long des particules Zig Zag en interaction

14:30-15:00 **Yaozhong Hu** (University of Alberta), **Khoa Le** (Imperial College London)
Density of Parabolic Anderson Random Variable / Densité de la variable aléatoire parabolique d'Anderson

13:30-15:00 **Invited / Sur invitation** (abstract/résumé 87) **201 (ENA)**

Measuring the Quality of Multisource Statistics
Évaluation de la qualité des statistiques multisources

Chair/Président: Katherine Jenny Thompson
Organizer/Responsable: Wesley Yung
Sponsor/Commanditaires: SSC Survey Methods Section / Groupe des méthodes d'enquête de la SSC

13:30-14:00 **John Eltinge** (US Census Bureau)
Assessment of Inferential Quality in the Integration of Multiple Data Sources / Évaluation de la qualité inférentielle dans l'intégration de sources de données multiples E E

14:00-14:30 **Susie Fortier** (Statistique Canada), **Martin Beaulieu** (Statistics Canada), **Ryan Chepita** (Statistics Canada)
Defining, Measuring and Communicating Quality in a Multi-Source Environment / Définir, mesurer et transmettre la qualité dans un environnement multisources E F E

14:30-15:00 **Rachel Skentelbery** (Office for National Statistics United Kingdom), **Hannah Finselbach** (Office for National Statistics UK)
Measuring the Quality of Multisource Statistics / Mesurer la qualité des statistiques multisources E E

13:30-15:00 **Invited / Sur invitation** (abstract/résumé 89) **101 (ENA)**

Measurement Error Models and Its Impacts in Health Sciences
Modèles d'erreur de mesure et impacts sur les sciences de la santé

Chair/Président: Mahmoud Torabi
Organizer/Responsable: Mahmoud Torabi
Sponsor/Commanditaires: SSC Biostatistics Section / Groupe de biostatistique de la SSC

13:30-14:00 **Grace Yi** (University of Waterloo)
Analysis of Multi-State Models with Misclassified States / Analyse des modèles multi-états avec erreurs de classification des états E E

14:00-14:30 **Liqun Wang** (University of Manitoba), **Lin Xue** (University of Manitoba), **Hengjian Cui** (Capital Normal University)
Variable Selection and Estimation in Generalized Linear Models with Measurement Error / Sélection et estimation de variables dans les modèles linéaires généralisés avec erreur de mesure E E

14:30-15:00 **Juxin Liu** (University of Saskatchewan), **Annshirley Afful** (University of Saskatchewan)
Joint Misclassification Errors in Both Response and Explanatory Variables / Erreurs de classification conjointes dans les variables de réponse et explicatives E E

13:30-15:00 **Invited / Sur invitation** (abstract/résumé 91) **122 (ICT)**

Rocky and Atlantic Collaborations in the Health Sciences
Collaborations en sciences de la santé dans les Rocheuses et l'Atlantique

Chair/Président: John Braun
Organizer/Responsable: John Braun
Sponsor/Commanditaires: CANSSI / INCASS

13:30-15:00 **John Braun** (The University of British Columbia), **Rasika Rajapakshe** (BC Cancer Agency), **Yingqi Wang** (University of Calgary), **Andrew Jirasek** (University of British Columbia), **Renjun Ma** (University of New Brunswick), **Henrik Stryhn** (University of Prince Edward Island)
Rocky and Atlantic Collaborations in the Health Sciences / Collaborations en sciences de la santé dans les Rocheuses et l'Atlantique E E

13:30-15:00 **Invited / Sur invitation** (abstract/résumé 92) **102 (ICT)**

Computational challenges in statistical learning for complex data
Défis computationnels en apprentissage statistique pour données complexes

Chair/Président: Teng Zhang

13:30-14:00 **Marianna Pensky** (University of Central Florida), **Rasika Rajapakshage** (University of Central Florida)
Clustering in Statistical Ill-Posed Linear Inverse Problems / Regroupement dans les problèmes inverses linéaires statistiques mal posés

14:00-14:30 **Yuwen Gu** (University of Connecticut), **Simon Fontaine** (University of Montreal), **Yi Yang** (McGill University), **Wei Qian** (University of Delaware), **Bo Fan** (University of Oxford)
A Unified Approach to Sparse Tweedie Modeling of Multi-Source Insurance Claims Data / Une approche unifiée pour modélisation éparse Tweedie de données de réclamations d'assurance à sources multiples

14:30-15:00 **Asad Haris** (McGill University), **Robert Platt** (McGill University)
A Targeted Approach to Confounder Selection for High-Dimensional Data / Approche ciblée dans le choix de variables de confusion pour des données de grande dimension

13:30-15:00 **Contributed / Communications libres** (abstract/résumé 94) **142 (AD)**

Diagnostic Tests and Prediction Models
Tests diagnostiques et modèles prédictifs

Chair/Président: Farouk Nathoo

13:30-13:45 **Alexander de Leon** (University of Calgary), **Joyce Raymond Punzalan** (University of the Philippines Diliman), **Hua Shen** (University of Calgary)
Estimation of Diagnostic Accuracy Measures of Correlated Diagnostic Tests for Paired Organs in the Absence of a Gold Standard / Estimation des mesures d'exactitude de diagnostic de tests diagnostiques correlés d'organes pairs en l'absence d'un étalon de référence

13:45-14:00 **Meng Yuan** (University of Waterloo), **Changbao Wu** (University of Waterloo), **Pengfei Li** (University of Waterloo)
Semiparameteric Inference of the Youden Index and Optimal Cut-Off Point / Inférence semiparamétrique de l'index de Youden et de la valeur-seuil optimale

14:00-14:15 **Wanhua Su** (MacEwan University)
Determining the Optimal Break-Point(s) Based on Precision-Recall Curves / Détermination du point de rupture optimal par courbes de précision-rappel

14:15-14:30 **Junhan Fang** (University of Waterloo), **Grace Yi** (University of Waterloo)
Regularized Matrix-Variate Regression with Misclassification in Binary Responses / Régression matrice-variables régularisée avec erreurs de classification dans des réponses binaires

14:30-14:45 **Ali Karimnezhad** (University of Ottawa), **Pearl Campbell** (Ottawa Hospital Research Institute), **Bryan Lo** (University of Ottawa), **David J. Stewart** (University of Ottawa), **Theodore J. Perkins** (University of Ottawa)
An Empirical Bayes Variant Calling Algorithm Designed for Next-Generation Sequencing Data Analysis / Un algorithme bayésien empirique d'identification de variantes pour analyse de données de séquençage de nouvelle génération

13:30-15:00 **Contributed / Communications libres** (abstract/résumé 97) **119 (SA)**

Recent Advances in Clinical Trials and Experimental Design and Inference
Percées récentes dans les essais cliniques et en conception et inférence expérimentales

Chair/Président: Judy-Anne W. Chapman

13:30-13:45 **Su Hwan Kim** (University of Alberta), **Keumhee Chough Carriere** (University of Alberta)
Optimal Crossover Designs with Baselines and Proportional Carryover Effects / Plans d'étude croisés optimaux avec effets de report proportionnels et de base

13:45-14:00 **Xinyi Ge** (Queen's University), **Yingwei Peng** (Queen's University), **Dongsheng Tu** (NCIC Clinical Trials; Queen's University)
A single-Index Threshold Linear Mixed Model for Identification of Treatment-Sensitive Subsets in a Clinical Trial Based on Longitudinal Outcomes and a Continuous Covariate / Un modèle linéaire mixte de seuil à indice unique pour l'identification des sous-ensembles sensibles à un traitement dans un essai clinique fondé sur des variables longitudinales et une covariable continue

14:00-14:15 **Keyue Ding** (Queen's University)
Group Sequential Test for Nested Multiple Populations According to Biomarker Expression Levels / Test séquentiel groupé pour populations multiples emboîtées selon les niveaux d'expression des biomarqueurs

14:15-14:30 **Eric Sanders** (University of British Columbia), **Paul Gustafson** (University of British Columbia), **Mohammad Ehsanul Karim** (University of British Columbia)
Incorporating Partial Adherence into the Principal Stratification Analysis Framework / Intégration de l'adhésion partielle dans le cadre de l'analyse de stratification principale

14:30-14:45 **Gurbakhshash Singh** (Central Connecticut State University), **Mark Lowerison** (University of Calgary), **Ayoola Ademola** (University of Calgary), **Bijoy K. Menon** (University of Calgary), **Michael D. Hill** (University of Calgary), **Tolulope Sajobi** (University of Calgary)
On Covariate Adaptive Randomization in Clinical Trials / De la randomisation adaptée aux covariables dans les essais cliniques

14:45-15:00 **Anthony Greco** (Brock University), **Xiaojian Xu** (Brock University)
Active Learning and Optimal Experimental Design / Apprentissage actif et conception expérimentale optimale

---

13:30-15:00 **Contributed / Communications libres** (abstract/résumé 101) **109 (SS)**

Missing Data Methods and Applications
Données manquantes : méthodes et applications

Chair/Président: Shakhawat Hossain

13:30-13:45 **Yang Zhao** (University of Regina), **Meng Liu** (University of Regina)
Consistent Estimation in Multiple Imputation for Regression Models with Missing Data / Estimation convergente sous imputation multiple pour les modèles de régression avec données manquantes

13:45-14:00 **David Luke Thiessen** (University of Regina), **Yang Zhao** (University of Regina)
Non-Monotone Missing Covariates in Cox Regression / Covariables manquantes non-monotones dans la régression Cox

14:00-14:15 **Shixiao Zhang** (University of Waterloo), **Peisong Han** (University of Michigan), **Changbao Wu** (University of Waterloo)
A Multiply Robust Mann-Whitney Test for Non-Randomized Pretest-Posttest Studies with Missing Data / Un test Mann-Whitney multi-robuste pour études non randomisées menées avant et après l'essai avec des données manquantes

14:15-14:30 **Sophie Castel** (Trent University), **Melissa Van Bussel** (Trent University), **Wesley Burr** (Trent University)
Imputation of Missing Values in Time Series Data / Imputation de valeurs manquantes dans des séries chronologiques

14:30-14:45 **Md. Shaddam Hossain Bagmar** (University of Calgary), **Hua Shen** (University of Calgary)
Causal Inference with Missingness in Confounders / Inférence causale accompagnée de facteurs de confusion comportant des lacunes

14:45-15:00 **Menglu Che** (University of Waterloo), **Jerry Lawless** (University of Waterloo), **Peisong Han** (University of Michigan)
Empirical and Conditional Likelihoods for Two-Phase Studies with Response-Dependent Samples / Vraisemblances empiriques et conditionnelles pour les études biphasées avec échantillonnage dépendant de la réponse Ⓔ E

13:30-15:00 **Invited / Sur invitation** (abstract/résumé 105) **113 (SS)**

Implementation, Advances and Precision in Mixture Model-Based Classification
Mise en œuvre, progrès et précision en classification fondée sur les modèles de mélange

Chair/Président: Jeffrey L. Andrews
Organizer/Responsable: Brian Franczak
Sponsor/Commanditaires: SSC Business & Industrial Statistics Section / Groupe de statistique industrielle et de gestion de la SSC

13:30-14:00 **Antonio Punzo** (University of Catania)
On the Use of the Contaminated Normal Distribution in Model-Based Clustering / L'utilisation de la loi normale contaminée dans le regroupement modélisé Ⓔ E

14:00-14:30 **Cristina Tortora** (San Jose State University), **Antonio Punzo** (University of Catania)
Advances in Model-Based Clustering and Outlier Detection / Progrès en matière de regroupement fondé sur un modèle et de détection des valeurs aberrantes Ⓔ E

14:30-15:00 **Hua Shen** (University of Calgary)
Mixture Model with Inaccurate Measurements / Modèle de mélange accompagné de mesures imprécises Ⓔ E

15:30-17:00 **Invited / Sur invitation** (abstract/résumé 107) **144 (SB)**

New Directions in Mathematical Finance
Nouvelles orientations en mathématiques financières

Chair/Président: Jean-François Bégin
Organizer/Responsable: Alexandru Badescu
Sponsor/Commanditaires: SSC Probability Section / Groupe de probabilité de la SSC

15:30-16:00 **Cody Hyndman** (Concordia University), **Anastasis Kratsios** (ETH Zurich)
The NEU Meta-Algorithm for Geometric Learning / Le méta-algorithme NEU pour l'apprentissage géométrique Ⓔ E

16:00-16:30 **Mark Reesor** (Wilfrid Laurier), **Xinghua Zhou** (Morgan Stanley)
Calibration of Capital Structure Models / Calibration de modèles de structure du capital Ⓔ E

16:30-17:00 **Jinniao Qiu** (University of Calgary)
Viscosity Solutions of Stochastic Hamilton-Jacobi-Bellman Equations and their Applications in Mathematical Finance / Solutions de viscosité des équations stochastiques de Hamilton-Jacobi-Bellman et applications en mathématiques financières Ⓔ E

15:30-17:00 **Invited / Sur invitation** (abstract/résumé 109) **102 (ICT)**

Making Sense of Complex Featured Data with Statistical Methods
Exploitation de données à caractéristiques complexes par méthodes statistiques

Chair/Président: Grace Yi
Organizer/Responsable: Grace Yi
Sponsor/Commanditaires: SSC Biostatistics Section / Groupe de biostatistique de la SSC

15:30-15:50 **Mireille E. Schnitzer** (Université de Montréal), **Lucie Blais** (Université de Montréal), **Robert Platt** (McGill University), **Madeleine Durand** (Centre Hospitalier de l'Université de Montréal and the Research Center of Centre Hospitalier de l'Université de Montréal)
Dealing with Time-Varying Eligibility for Exposure Using the Target Trials Approach to Causal Inference / Traitement de l'admissibilité variant dans le temps de l'exposition à l'aide de l'approche des essais ciblés pour l'inférence causale

15:50-16:10 **Wenqing He** (University of Western Ontario), **Grace Yi** (University of Waterloo), **Junhan Fang** (University of Waterloo)
Prediction for Error-Contaminated Image Data with an Application of the Prostate Cancer Imaging Study / Prédiction des données d'images contaminées par des erreurs à l'aide d'une application de l'étude d'imagerie du cancer de la prostate

16:10-16:30 **Trevor Thomson** (Simon Fraser University), **John Braun** (University of British Columbia - Okanagan), **Joan Hu** (Simon Fraser University)
On Time to First Spot Fire / Délai avant premier feu disséminé

16:30-16:50 **Gabrielle Simoneau** (McGill University), **Erica Moodie** (McGill University), **Laurent Azoulay** (McGill University), **Robert Platt** (McGill University)
Estimating Optimal Dynamic Treatment Regimes with Survival Outcomes : An Application to the Treatment of Type 2 Diabetes / Estimation optimale des plans dynamiques de traitements avec les issues de survie : une application sur le diabète de type 2

15:30-17:00 **Invited / Sur invitation** (abstract/résumé 112) **119 (SA)**

Recent Developments in Statistical Analysis of Nutrition Data
Récentes évolutions en analyse statistique de données nutritionnelles

Chair/Président: Dominique Ibañez
Organizer/Responsable: Dominique Ibañez

15:30-16:00 **Hassan Vatanparast**, **Rashmi Prakash Patil** (University of Saskatchewan), **Naorin Islam** (University of Saskatchewan), **Seyed Hamzeh Hosseini** (University of Saskatchewan), **Zeinab Hosseini** (University of Saskatchewan), **Arash Shamloo** (University of Saskatchewan), **Pardis Keshavarz** (University of Saskatchewan)
National Nutrition and Health Survey Data Analyses, Challenges and Opportunities / Analyses de données d'enquêtes nationales sur la nutrition et la santé, défis et opportunités

16:00-16:30 **Dominique Ibañez** (Health Canada), **Karelyn Davis** (Health Canada), **Alejandro Gonzalez** (Health Canada), **Lidia Loukine** (Health Canada), **Cunye Qiao** (Health Canada), **Alireza Sadeghpour** (Health Canada), **Michel Vigneault** (Health Canada), **Kuan Chiao Wang** (Health Canada)
Early Experience with the National Cancer Institute (NCI) Method for Estimating Usual Intakes Using the Canadian Community Health Survey / Premières expériences avec la méthode d'estimation des apports usuels du National Cancer Institute (NCI) sur l'Enquête sur la santé dans les collectivités canadiennes

16:30-17:00 **Alireza Sadeghpour** (Health Canada), **Karelyn Davis** (Health Canada), **Nadine Kebbe** (Health Canada), **Isabelle Rondeau** (Health Canada), **Michel Vigneault** (Health Canada), **Dominique Ibañez** (Health Canada)
The Effect of the Food Model Booklet on Reported Foods in the Canadian Community Health Survey (CCHS) – Nutrition / L'effet du livret de modèles de portions (LMP) sur les aliments déclarés dans l'Enquête sur la santé dans les collectivités canadiennes (ESCC) – Nutrition

---

15:30-17:00 **Invited / Sur invitation** (abstract/résumé 114) **146 (SB)**

Recent Statistics Research of New Investigators Across Canada
Récentes recherches des nouveaux chercheurs en statistique au Canada

Chair/Président: Reza Ramezan
Organizer/Responsable: Reza Ramezan
Sponsor/Commanditaires: SSC New Investigator Committee / Comité des nouveaux chercheurs de la SSC

15:30-16:00 **Félix Camirand Lemyre** (Université de Sherbrooke), **Aurore Delaigle** (University of Melbourne), **Raymond J. Carroll** (Texas A&M University)
Non-Parametric Estimation of the Distribution of Episodically Consumed Food Measured with Error / Estimation non paramétrique de la distribution de l'apport habituel d'aliments épisodiquement consommés Ⓔ E

16:00-16:30 **Jeffrey L. Andrews** (University of British Columbia — Okanagan)
On Overfitting in Cluster Analysis / Du surajustement en analyse typologique Ⓔ E

16:30-17:00 **Jonathan Jalbert** (Polytechnique Montreal), **Luc Perreault** (Institut de recherche d'Hydro-Québec)
Interpolation of Extreme Precipitation of Multiple Durations in Eastern Canada / Interpolation d'extrêmes de précipitations de multiples durées dans l'Est du Canada Ⓔ E

15:30-17:00 **Contributed / Communications libres** (abstract/résumé 116) **142 (AD)**

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

Chair/Président: Alison L. Gibbs

15:30-15:45 **Steven Wu** (Shopify)
People Analytics: How Shopify Uses Statistical Methods to Make Better Decisions for Its Employees / People analytics : comment Shopify utilise-t-il des méthodes statistiques pour prendre de meilleures décisions pour ses employés Ⓔ E

15:45-16:00 **Shan Shi** (University of Victoria), **Farouk Nathoo** (University of Victoria)
Feature Learning and Classification in Neuroimaging: Predicting Cognitive Impairment from Magnetic Resonance Imaging / Apprentissage et classification des caractéristiques en neuroimagerie : prédiction des troubles cognitifs causés par l'imagerie par résonance magnétique Ⓔ E

16:00-16:15 **Christopher Salahub** (University of Waterloo), **Wayne Oldford** (University of Waterloo)
About 'Her Emails' / À propos des « courriels d'Hillary Clinton » Ⓔ E

16:15-16:30 **Usama Zafar Ansari** (The University of British Columbia), **Chengkai Zhang** (University of British Columbia, Okanagan)
Statistical Analysis of Vessel Motion Patterns in the Ports and Waterways Using Automatic Identification System (AIS) / Analyse statistique des mouvements des navires dans les ports et sur les voies navigable en utilisant le système d'identification automatique (SIA) Ⓔ E

16:30-16:45 **Gabriel C. Phelan** (Simon Fraser University), **David A. Campbell** (Simon Fraser University)
Geographically Aware Latent Dirichlet Allocation via Random Effects / Allocation de Dirichlet latente géographiquement consciente par l'entremise d'effets aléatoires Ⓔ E

16:45-17:00 **Bo Chen** (DLSPH, University of Toronto), **Keith A. Lawson** (University Health Network), **Antonio Finelli** (University Health Network), **Olli Saarela** (University of Toronto)
Four-Way Causal Variance Decompositions for Evaluating Hospital and Surgeon Performance / Décompositions de variance causale à quatre sens pour évaluer la performance des hôpitaux et des chirurgiens Ⓔ E

15:30-17:00 **Contributed / Communications libres** (abstract/résumé 120) **201 (ENA)**

Software Development and Computationally-Intensive Methods
Mise au point de logiciels et méthodes à forte intensité de calculs

Chair/Président: David Saunders

15:30-15:45 **Song Cai** (Carleton University), **Laura Dumitrescu** (Victoria University of Wellington), **JNK Rao** (Carleton University), **Golshid Chatrchi** (Carleton University)
A Simple and Effective Variable Selection Method for Two-Fold Sub-Area Models in Small Area Estimation / Méthode de sélection de variables simple et efficace pour modèles doubles de sous-domaine en estimation pour petits domaines

15:45-16:00 **Martin Lysy** (University of Waterloo)
The 'msde' Package: Fast Inference for Stochastic Differential Equations in R / La bibliotèque R «msde» : une inférence rapide appliquée aux équations différentielles stochastiques dans R

16:00-16:15 **Dan Richard** (MacEwan University), **Karen Buro** (MacEwan University), **Wanhua Su** (MacEwan University)
Zero Order vs (Semi) Partial Correlation Test and Confidence Interval / Test de corrélation d'ordre zéro vs corrélation (semi) partielle et intervalle de confiance

16:15-16:30 **Avinash Prasad** (University of Waterloo), **Marius Hofert** (University of Waterloo), **Mu Zhu** (University of Waterloo)
Quasi-Random Number Generators for Multivariate Distributions Based on Generative Neural Networks / Générateurs de nombres quasi-aléatoires pour distributions multivariées fondés sur des réseau de neurones génératifs

16:30-16:45 **Peter D.M. Macdonald** (McMaster University)
Bootstrapping Finite Mixture Distributions / Bootstrap de distributions de mélange finies

16:45-17:00 **Jun Yang** (University of Toronto), **Zhou Zhou** (University of Toronto)
Spectral Inference under Complex Temporal Dynamics / Inférence spectrale en dynamique temporelle complexe

15:30-17:00 **Contributed / Communications libres** (abstract/résumé 123) **101 (ENA)**

Patient-Focused Statistical Methods
Méthodes statistiques axées sur le patient

Chair/Président: Ali Karimnezhad

15:30-15:45 **Zhihui (Amy) Liu** (Princess Margaret Cancer Centre), **Olli Saarela** (University of Toronto), **Wei Xu** (Princess Margaret Cancer Centre, University Health Network)
Swimmers and Sinkers: Statistical Basis of the Swimmer Plot / Nageurs et pesées : base statistique du graphique du nageur (swimmer plot)

15:45-16:00 **Cong Jiang** (University of Waterloo), **Michael Wallace** (University of Waterloo), **Mary Thompson** (University of Waterloo)
Dynamic Treatment Regimes with Interference / Régimes de traitement dynamiques avec interférence

16:00-16:15 **Alomgir Hossain** (University of Ottawa/University of Ottawa Heart Institute), **Benjamin Chow** (University of Ottawa Heart institute)
Long-Term Prognostic Value of Coronary CT Angiography / Valeur prognostique à long terme d'une coronarographie par tomodensitométrie (TDM)

16:15-16:30 **Dylan Spicker** (University of Waterloo), **Michael Wallace** (University of Waterloo)
Measurement Error in Precision Medicine and Dynamic Treatment Regimes / Erreur de mesure dans la médicine de précision et régimes thérapeutiques dynamiques

16:30-16:45 **Katherine Daignault** (University of Toronto), **Keith A. Lawson** (Princess Margaret Cancer Centre, University Health Network), **Antonio Finelli** (Princess Margaret Cancer Centre, University Health Network), **Olli Saarela** (University of Toronto)
Implementation of Causal Mediation Analysis in Hospital Profiling / Exécution d'une analyse de médiation causale pour le profilage hospitalier

16:45-17:00 **Zayd Omar** (McGill University), **David Stephens** (McGill University), **Alexandra M. Schmidt** (McGill University)
Estimating ICU Heart Rate Data Using a Bayesian State-Space Model with GARCH(1,1) Errors / Estimation des données relatives à la fréquence cardiaque dans une unité de soins intensifs à l'aide d'un modèle d'espace-état bayésien avec erreurs GARCH(1,1)

15:30-17:00 **Invited / Sur invitation** (abstract/résumé 127) **113 (SS)**

Graduate students in actuarial science
Étudiants de troisième cycle en science actuarielle

Chair/Président: Anne Mackay
Organizer/Responsable: Anne Mackay
Sponsor/Commanditaires: SSC Actuarial Science Section / Groupe de science actuarielle de la SSC

15:30-15:45 **Tsz Chai Fung** (University of Toronto), **Andrei Badescu** (University of Toronto), **Sheldon Lin** (University of Toronto)
A Class of Mixture of Experts Models for General Insurance / Une classe de mélange de modèles experts pour l'assurance dommages

15:45-16:00 **Francis Duval** (Université du Québec à Montréal), **Mathieu Pigeon** (Université du Québec à Montréal)
Gradient Boosting Techniques for Individual Loss Reserving in Non-Life Insurance / Techniques de gradient boosting pour la modélisation des réserves individuelles en assurance non-vie

16:00-16:15 **Jessica Ou Dang** (University of Waterloo), **Mingbin Feng** (University of Waterloo), **Mary Hardy** (University of Waterloo)
Efficient Nested Simulation of Tail Risk Measures / Simulation emboitée efficace de mesures de risques extrêmes

16:15-16:30 **Carlos Andres Araiza Iturria** (Concordia University), **Mélina Mailhot** (Concordia University), **Frédéric Godin** (Concordia University)
Modeling and Measuring Insurance Risks within the IFRS 17 Framework: A Hierarchical Copula Approach / Modélisation et mesure des risques d'assurance conformes à l'IFRS 17 : approche de copules hiérarchiques

16:30-16:45 **Yunran Wei** (University of Waterloo), **Ruodu Wang** (University of Waterloo)
Risk Functionals with Convex Level Sets / Fonctions de risque et ensembles de niveaux convexes

16:45-17:00 **Ihsan Chaoubi** (Laval University), **Hélène Cossette** (Laval University), **Étienne Marceau** (Laval University)
On Sums of Two Counter-Monotonic Risks / Les sommes de deux risques monotones contraires

---

15:30-17:00 **Invited / Sur invitation** (abstract/résumé 131) **122 (ICT)**

Designed experiments for complex engineered systems
Plans d'expériences pour systèmes techniques complexes

Chair/Président: Ryan Lekivetz
Organizer/Responsable: Ryan Lekivetz
Sponsor/Commanditaires: SSC Business & Industrial Statistics Section / Groupe de statistique industrielle et de gestion de la SSC

15:30-16:00 **Joseph Morgan** (SAS Institute)
Covering Arrays: A Tool for Testing Complex Engineered Systems / Tableaux de couverture : un outil pour tester des systèmes complexes en ingénierie

16:00-16:30 **Ryan Lekivetz** (SAS Institute), **Joseph Morgan** (SAS Institute)
Design and Analysis of Covering Arrays Using Prior Information / Conception et analyse de tableaux de couverture en utilisant de l'information préalable

16:30-17:00 **Karen Meagher** (University of Regina)
Covering Arrays and Pure Math / Tableaux de couverture et les mathématiques pures

Tuesday May 28 — mardi 28 mai

08:30-09:50 — **Invited / Sur invitation** (abstract/résumé 133) — **148 (ST)**

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

Chair/Président: Brian Allen
Organizer/Responsable: Jack Gambino

08:30-09:50 **Douglas Wiens** (University of Alberta)
Robustness of Design: A Survey / Robustesse du plan : une étude

---

10:20-11:50 — **Invited / Sur invitation** (abstract/résumé 134) — **102 (ICT)**

Advances in spatial epidemiology and ecology
Avancées en épidémiologie et écologie spatiales

Chair/Président: Rob Deardon
Organizer/Responsable: Rob Deardon

10:20-10:50 **Andrew Lawson** (Medical University of South Carolina)
Bayesian Spatial Modeling for Prospective ID Surveillance: With Application to Seasonal Influenza / Modélisation spatiale bayésienne pour la surveillance prospective des maladies infectieuses, avec application à la grippe saisonnière

10:50-11:20 **Joanna Elizabeth Mills Flemming** (Dalhousie University)
New Approaches for Estimating Population Size for Marine Species / Nouvelles approches pour l'estimation de la taille des populations d'espèces marines

11:20-11:50 **Md Mahsin** (University of Calgary), **Rob Deardon** (University of Calgary), **Patrick Brown** (University of Toronto)
A New Class of Spatiotemporal Individual-Level Models for Infectious Diseases Transmission / Nouvelle catégorie de modèles spatio-temporels au niveau de l'individu en matière de transmission de maladies infectieuses

---

10:20-11:50 — **Invited / Sur invitation** (abstract/résumé 136) — **116 (ICT)**

Modern methods in functional data analysis
Méthodes modernes pour l'analyse de données fonctionnelles

Chair/Président: Joel A. Dubin
Organizer/Responsable: Joel A. Dubin

10:20-10:50 **Hans-Georg Müller** (University of California, Davis), **Xiongtao Dai** (Iowa State University)
Nonparametric Modeling of Longitudinal Compositional and Functional Data on Riemannian Manifolds / Modélisation non paramétrique des données fonctionnelles et compositionnelles longitudinales sur les variétés riemanniennes

10:50-11:20 **Peijun Sang** (University of Waterloo), **Jiguo Cao** (Simon Fraser University)
Distance-Weighted Discrimination for Functional Data / Discrimination pondérée par la distance pour données fonctionnelles

11:20-11:50 **Jane-Ling Wang** (University of California, Davis), **Xiaoke Zhang** (George Washington University)
Varying-Coefficient Additive Models: Two Birds with One Stone? / Modèles additifs à coefficients variables : d'une pierre deux coups ?

10:20-11:50 **Invited / Sur invitation** (abstract/résumé 138) **146 (SB)**

A Showcase of Student Research from the CANSSI CRT 'Joint Analysis of Neuroimaging Data: High-Dimensional Problems, Spatiotemporal Models and Computation'
Recherches d'étudiants du PRC de l'INCASS "Analyse conjointe de données de la neuroimagerie : problèmes en grande dimension, modèles spatiotemporels et calculs"

Chair/Président: Farouk Nathoo
Organizer/Responsable: Farouk Nathoo

10:20-10:50 **Yin Song** (University of Victoria), **Farouk Nathoo** (University of Victoria), **Arif Babul** (University of Victoria)
A Potts-Mixture Spatiotemporal Joint Model for Combined MEG and EEG Data / Un modèle de mélange Potts spatio-temporel conjoint pour données combinées MEG et EEG Ⓔ E

10:50-11:20 **Farouk Nathoo** (University of Victoria), **Yunlong Nie** (Simon Fraser University)
Spectral Dynamic Causal Modelling of Resting-State fMRI: Relating Effective Brain Connectivity in the Default Mode Network to Genetics / Modélisation causale à dynamique spectrale de l'IRMf au repos : faire le pont de la génétique à la connectivité cérébrale efficace dans le réseau du mode par défaut Ⓔ E

11:20-11:50 **Eugene Opoku** (University of Victoria), **Farouk Nathoo** (University of Victoria), **Ejaz Syed Ahmed** (Brock University)
Ant Colony System Optimization for Estimation in Spatial Hidden Markov Models / Optimisation du système de colonie de fourmis pour l'estimation dans un modèle de Markov caché spatial Ⓔ E

10:20-11:50 **Invited / Sur invitation** (abstract/résumé 140) **109 (SS)**

Effective Implementation of Statistics Capstone Courses
Mise en place effective de cours finaux en statistique

Chair/Président: Asokan Mulayath Variyath
Organizer/Responsable: Asokan Mulayath Variyath
Sponsor/Commanditaires: SSC Statistical Education Section / Groupe d'éducation en statistique de la SSC

10:20-10:50 **Gemai Chen** (University of Calgary)
Capstone Course: What For? / Cours de synthèse : pour quoi ? Ⓔ E

10:50-11:20 **Karen Buro** (MacEwan University)
A Course Engaging Undergraduates in Statistical Consultation / Un cours qui incite les étudiants de premier cycle à la consultation statistique Ⓔ E

11:20-11:50 **Gabriela Cohen Freue** (University of British Columbia)
Case Studies and Consulting in Statistics / Études de cas et consultation en statistique Ⓔ E

10:20-11:50 **Invited / Sur invitation** (abstract/résumé 142) **142 (AD)**

Recent advances in statistical inference for complex data structures
Dernières avancées en inférence statistique pour les structures de données complexes (commandité par le chapitre canadien de l'ICSA)

Chair/Président: Liqun Wang
Organizer/Responsable: Liqun Wang
Sponsor/Commanditaires: ICSA-Canada Chapter / ICSA-Canada Chapter

10:20-10:50 **Christopher M. Manuel** (Texas A&M University)
Matched Case-Control Data with a Misclassified Exposure: What Can Be Done with Instrumental Variables? / Données de comparaison avec les témoins appariés accompagnées d'une classification erronée de l'exposition : que faire avec des variables instrumentales ? Ⓔ E

10:50-11:20 **Mahmoud Torabi** (University of Manitoba), **Vahid Tadayon** (Higher Education Center of Eghlid, Iran)
Measurement Error in Spatial Models with Fat Tails and Skewed Errors / Erreur de mesure dans les modèles spatiaux avec queues épaisses et erreurs asymétriques

11:20-11:50 **Zhou Zhou** (University of Toronto), **Weichi Wu** (Tsinghua University)
MACE: Multiscale Abrupt Change Estimation under Complex Temporal Dynamics / EMCA : Estimation multiéchelle de changement abrupt en fonction d'une dynamique temporelle complexe

---

10:20-11:50 **Contributed / Communications libres** (abstract/résumé 144) **143 (ST)**

Decisions in Finance and Economics
Décisions en matière de finance et d'économie

Chair/Président: François Bellavance

10:20-10:35 **Jean-François Bégin** (Simon Fraser University)
Economic Scenario Generator and Parameter Uncertainty: A Bayesian Approach / Générateur de scénarios économiques et incertitude des paramètres : une approache bayésienne

10:35-10:50 **Yifan Li** (Western University), **Reg Kulperger** (Western University), **Hao Yu** (Western University)
A Test of Knightian Uncertainty under the G-Expectation Framework / Un test sur l'incertitude de Knight dans le cadre de la G-espérance

10:50-11:05 **Zhenxian Gong** (Western University), **Marcos Escobar-Anel** (Western University)
The Mean-Reverting 4/2 Stochastic Volatility Model: Properties and Financial Applications. / Modèle de volatilité stochastique 4/2 avec retour à la moyenne : propriétés et applications financières

11:05-11:20 **Wei-Hsiang Lin** (Simon Fraser University), **Shih-Kuei Lin** (National Chengchi University), **Cary Chi-Liang Tsai** (Simon Fraser University)
Impact of Interest Rate, Surrender, and Liquidity Risks on the Surplus of a Portfolio of Endowment Policies Using Optimal Portfolio Selection Techniques / Analyse de l'impact du taux d'intérêt, du rachat et des risques d'illiquidité sur l'excédent d'un portefeuille de polices de dotation de fonds à l'aide de techniques de sélection de portefeuille optimal

11:20-11:35 **Xing Gu** (University of Western Ontario), **Rogemar Mamon** (Western University), **Matt Davison** (Western University), **Hao Yu** (Western University)
An Analysis and Forecasting of Financial Market Liquidity Regimes / Analyse et prévision des régimes de liquidité des marchés financiers

11:35-11:50 **Javad Rastegari** (Western University), **Lars Stentoft** (Western University), **Marcos Escobar-Anel** (Western University)
Option Pricing with Conditional GARCH Models / Établir le prix des options avec les modèles GARCH conditionnels

---

10:20-11:50 **Contributed / Communications libres** (abstract/résumé 148) **105 (SB)**

Methods for High-Dimensional and Large Data I
Méthodes pour traiter les données volumineuses et de grande dimension I

Chair/Président: Whitney K. Huang

10:20-10:35 **Zheng Gao** (University of Michigan), **Stilian Stoev** (University of Michigan)
Fundamental Limits of Exact Support Recovery in High Dimensions / Limites fondamentales du support de redressement exact en haute dimension

10:35-10:50 **Grace Guan Hsu** (Simon Fraser University), **Derek Bingham** (Simon Fraser University)
Super Fast Emulation and Calibration of Large Computer Experiments / Émulation et étalonnage ultra-rapides de grandes expériences informatiques

10:50-11:05 **Shubhadeep Chakraborty** (Texas A&M University, USA), **Xianyang Zhang** (Texas A&M University)
A New Framework for Distance and Kernel-Based Metrics in High Dimensions / Un nouveau cadre pour les mesures fondées sur des distances et des noyaux en haute dimension

11:05-11:20 **Carolyn Augusta** (University of Guelph), **Graham W. Taylor** (University of Guelph), **Rob Deardon** (University of Calgary)
Conditional Variational Recurrent Graph Autoencoders / Auto-encodeurs de graphes conditionnels de variations récurrentes Ⓔ E

11:20-11:35 **Anh Nam Tran** (University of Manitoba), **Saumen Mandal** (University of Manitoba)
Construction of Bayesian Optimal Designs for Nonlinear Models / Construction de plans bayésiens optimaux pour modèles non linéaires Ⓔ E

11:35-11:50 **Klaus Herrmann** (University of Waterloo), **Maximilian Coblenz** (Karlsruhe Institute of Technology), **Oliver Grothe** (Karlsruhe Institute of Technology), **Marius Hofert** (University of Waterloo)
Smooth Bootstrapping of Copula Functionals / Procédure de bootstrap lissé pour les fonctions de copule Ⓔ E

10:20-11:50 **Contributed / Communications libres** (abstract/résumé 152) **201 (ENA)**

Classification and Learning
Classification et apprentissage

Chair/Président: Wei Liu

10:20-10:35 **Pramoda Sachinthana Jayasinghe** (University of Manitoba), **Mohammad Jafari Jozani** (University of Manitoba), **Behzad Kordi** (University of Manitoba)
Developing New Statistical Pattern Recognition and System Identification Techniques for Partial Discharge Analysis / Élaboration de nouvelles techniques de reconnaissance des formes statistique et d'identification de systèmes pour l'analyse des décharges partielles Ⓔ E

10:35-10:50 **Pingzhao Hu** (University of Manitoba), **Jiaying You** (University of Manitoba), **Bob McLeod** (University of Manitoba)
New Machine Learning Approaches for Drug-Target Interaction Network Prediction and Drug Repurposing / Nouvelles approches dans l'apprentissage machine pour la prédiction du réseau d'interaction médicament-cible et la reconversion de médicaments Ⓔ E

10:50-11:05 **Hanning Chen** (University of Calgary), **Jingjing Wu** (University of Calgary)
Two-Class Classification Problem of Rare and Weak Signal on Variances / Problème de classification en deux classes d'un signal rare et faible des variances Ⓔ E

11:05-11:20 **Jiaxin Zhang** (University of Alberta), **Adam Kashlak** (University of Alberta)
High Dimensional Classification Using Sparse Covariance Matrices / Classification de grande dimension à l'aide de matrices de covariance éparses Ⓔ E

11:20-11:35 **Rachid Kharoubi** (UQAM), **Karim Oualkacha** (UQAM)
The Cluster Correlation-Network Support Vector Machine for High-Dimensional Binary Classification / La machine à vecteurs de support avec réseau de corrélation par regroupement pour la classification binaire de haute dimension Ⓔ E

11:35-11:50 **Zihang Lu** (DLSPH, University of Toronto), **Wendy Lou** (University of Toronto)
Bayesian Growth Mixture Model with Variable Selection for Clustering Longitudinal Trajectories / Modèle de mélange bayésien de croissance avec sélection de variables pour le regroupement des trajectoires longitudinales Ⓔ E

12:00-17:00 **Contributed / Communications libres** (abstract/résumé 156) **103Z (ST)**

Poster Session
Session d'affiches

12:00-17:00 **Yunjing Li** (University of Toronto), **Nicholas Mitsakakis** (University of Toronto)
Categorical State Sequence Analysis to Describe and Analyze Pathways of Health State Transitions / Analyse de séquence d'états catégorique pour décrire et analyser les chemins de transitions d'états de santé Ⓔ E

12:00-17:00 **Xiaohua Liu** (University of Regina), **Taehan Bae** (University of Regina)
Predictive Modelling of Extreme Values in Stock Return Data / Modélisation prédictive des valeurs extrêmes dans les données sur le rendement des actions Ⓔ E

12:00-17:00 **Xuewen Lu** (University of Calgary), **Rutong Cai** (University of Calgary), **Beijia Hu** (University of Calgary), **Alexander Liu** (University of Calgary), **Liping Luo** (Hangyang Normal University), **Connie Sze** (University of Calgary), **Yao Yao** (University of Calgary)
Group Selection for Accelerated Failure Time Models with Random Effects / Sélection de groupes pour les modèles de temps de défaillance accéléré avec effets aléatoires Ⓔ E

12:00-17:00 **Dongmeng Liu** (Simon Fraser University), **Jinko Graham** (Simon Fraser University)
Sampling Partial Genealogies Using Sequential Importance Sampling / Échantillonnage de généalogies partielles à l'aide de l'échantillonnage séquentiel par importance Ⓔ E

12:00-17:00 **Mehdi Rostamiforooshani** (University of Toronto), **Nancy Reid** (University of Toronto), **Olli Saarela** (University of Toronto)
Non-Parametric Feature Selection with False Discovery Rate Control / Sélection de caractéristiques non paramétriques avec contrôle du taux de fausses découvertes Ⓔ E

12:00-17:00 **Menglin Zhou** (University of British Columbia), **Natalia Nolde** (University of British Columbia), **Chen Zhou** (Erasmus University)
An Extreme Value Approach to CoVaR Estimation / Approche des valeurs extrêmes de l'estimation de la CoVaR Ⓔ E

12:00-17:00 **Li-Pang Chen** (University of Waterloo), **Grace Yi** (University of Waterloo)
Analysis of Graphical Models with Error-Prone Variables / Analyse de modèles graphiques avec variables sujettes à erreur Ⓔ E

12:00-17:00 **Anqi Chen** (Simon Fraser University), **Boxin Tang** (Simon Fraser University)
Selecting Baseline Two-Level Designs Using Optimality and K-Aberration Criteria When Some Two-Factor Interactions Are Important / Sélection de plans de référence à deux niveaux à l'aide de critères d'optimalité et d'aberration K lorsque certaines interactions entre deux facteurs sont importantes Ⓔ E

12:00-17:00 **Yunqi Ji** (Alberta Health Services), **Farrell Cahill** (Memorial University of Newfoundland), **Yanqing Yi** (Memorial University of Newfoundland), **Edward Randell** (Memorial University of Newfoundland), **Guang Sun** (Memorial University of Newfoundland)
New Sex Specific Predictive Equation for Evaluating Body Fat Percentage / Une nouvelle équation prédictive spécifique au sexe pour déterminer le pourcentage de masse grasse Ⓔ E

12:00-17:00 **Jeffrey Negrea** (University of Toronto)
Optimal Scaling, Optimal Shaping and Speed Limits of Random Walk Metropolis via Diffusion Limits of Block-I.I.D. Targets / Échelle optimale, formation optimale et limites de vitesse d'une marche aléatoire Metropolis par des limites de diffusion des cibles blocs I.I.D Ⓔ E

12:00-17:00 **Zhiyang Zhou** (Simon Fraser University), **Peijun Sang** (University of Waterloo)
Continuum Centroid Classifier for Functional Data / Classificateur centroïde continu pour les données fonctionnelles Ⓔ E

12:00-17:00 **Myriam Brossard** (Sinai Health System), **Andrew Paterson** (Hospital for Sick Children & University of Toronto), **Osvaldo Espin-Garcia** (Sinai Health System & University of Toronto), **Radu Craiu** (University of Toronto), **Shelley Bull** (Sinai Health System & University of Toronto)
Joint Modelling of Multivariate Longitudinal and Time-To-Event Phenotypes in Genetic Association Studies of Complex Traits / Modélisation conjointe de phénotypes longitudinaux et de durées de vie multivariés dans les études sur l'association génétique de caractéristiques complexes Ⓔ E

12:00-17:00 **Michela Panarella** (University of Toronto DLSPH), **Razvan Romanescu** (Lunenfeld-Tanenbaum Research Institute), **Gessica Gos** (Lunenfeld-Tanenbaum Research Institute), **Irene Andrulis** (Lunenfeld-Tanenbaum Research Institute), **Shelley Bull** (Lunenfeld-Tanenbaum Research Institute)
Extending Rare Variant Association Tests in Affected Siblings to Account for Background Genetic Risk / Prolongement des tests d'association de variantes rares chez des frères-sœurs atteints pour prendre en compte le risque génétique Ⓔ E

12:00-17:00 **Faezeh Yazdi** (Simon Fraser University), **Derek Bingham** (Simon Fraser University), **Daniel Williamson** (University of Exeter)
Emulation of Computer Models Using Deep Gaussian Processes / Émulation de modèles informatisés à l'aide de processus gaussiens profonds Ⓔ E

12:00-17:00 **Amirhossein Alvandi** (Memorial University of Newfoundland), **Armin Hatefi** (Memorial University of Newfoundland)
Population Proportion Estimation Using Rank-Based Sampling / Estimation proportionnelle d'une population à l'aide d'un échantillonnage basé sur les rangs Ⓔ E

13:30-15:00 **Invited / Sur invitation** (abstract/résumé 164) **146 (SB)**

Recent Advances in Actuarial and Quantitative Finance
Dernières avancées en finance actuarielle et quantitative

Chair/Président: Jean-François Bégin
Organizer/Responsable: Jean-François Bégin
Sponsor/Commanditaires: SSC Actuarial Science Section / Groupe de science actuarielle de la SSC

13:30-14:00 **Jean-François Renaud** (Université du Québec à Montréal)
How Are Dividend Payments Affected by Parisian Ruin? / Comment la ruine parisienne influence-t-elle les paiements de dividende ? Ⓔ E

14:00-14:30 **Patrice Gaillardetz** (Concordia University), **Saeb Hachem** (Concordia University)
Risk-Control Strategies / Stratégies de contrôle des risques Ⓔ E

14:30-15:00 **Adam Metzler** (Wilfrid Laurier University), **Mark Reesor** (Wilfrid Laurier University), **Wisdom S. Avusuglo** (Western University)
A General Framework for Modelling PD-LGD Correlation in Loan Portfolios: Some Interesting Observations / Un cadre général pour la modélisation de corrélations PD-LGD dans des portefeuilles de prêts, avec quelques observations intéressantes Ⓔ E

13:30-15:00 **Invited / Sur invitation** (abstract/résumé 166) **201 (ENA)**

Modeling, Imputation and non response
Modélisation, imputation et non-réponse

Chair/Président: David Haziza
Organizer/Responsable: Susie Fortier
Sponsor/Commanditaires: SSC Survey Methods Section / Groupe des méthodes d'enquête de la SSC

13:30-14:00 **Katherine Jenny Thompson** (U.S. Census Bureau), **Nicole Czaplicki** (U.S. Census Bureau), **Brian Dumbacher** (U.S. Census Bureau), **Stephen Kaputa** (U.S. Census Bureau)
Developing Imputation Models for the Advanced Monthly Retail Trade Survey: A Subsampled Survey with Seasonal Data, Low Unit Response, and High Profile / Élaboration de modèles d'imputation pour l'Advance Monthly Retail Trade Survey : étude par sous-échantillonnage avec données saisonnières, faible réponse par unité et haut profil Ⓔ E

14:00-14:30 **Geneviève Vézina** (Statistics Canada), **Andrew Brennan** (Statistics Canada), **Catherine Deshaies-Moreault** (Statistics Canada)
Measuring Cannabis Prevalence, Consumption and Price: The Statistics Canada Experience / Mesure de la prévalence, de la consommation et du prix du cannabis : l'expérience de Statistique Canada Ⓔ E

14:30-15:00 **Valéry Dongmo Jiongo** (Canada Mortgage and Housing Corporation)
Predicting Rental Prices for Canadian Rural Centres / Prévision des prix de location des centres ruraux canadiens ⒺⒻ E

13:30-15:00 **Invited / Sur invitation** (abstract/résumé 168) **109 (SS)**

Models and Applications for Functional Data Analysis
Modèles et applications pour l'analyse de données fonctionnelles

Chair/Président: Haocheng Li
Organizer/Responsable: Haocheng Li

13:30-14:00 **David A. Campbell** (Simon Fraser University), **Subhash Lele** (University of Alberta), **Peter Solymos** (University of Alberta)
Functional Data Analysis for Assessing Convergence of Sampled Densities / Analyse de données fonctionnelles pour évaluer la convergence de densités échantillonnées Ⓔ E

14:00-14:30 **Greg Rice** (University of Waterloo), **Piotr Kokoszka** (Colorado State University), **Han Lin Shang** (Australian National University), **Yuqian Zhao** (University of Waterloo), **Tony Wirjanto** (University of Waterloo)
Inference for the Autocovariance of a Functional Time Series and Goodness-Of-Fit Tests for fGARCH Models / Inférence pour l'autocovariance d'une série chronologique fonctionnelle et tests de qualité de l'ajustement de modèles fGARCH Ⓔ E

14:30-15:00 **Jiguo Cao** (Simon Fraser University), **Fei Jiang** (University of Hong Kong), **Seungchul Baek** (University of South Carolina), **Yanyuan Ma** (Penn State University)
Functional Single Index Model / Modèle à indice fonctionnel simple Ⓔ E

13:30-15:00 **Invited / Sur invitation** (abstract/résumé 170) **102 (ICT)**

New statistical techniques in exploring modern data with complex structure
Nouvelles techniques statistiques pour l'exploration de données modernes à structure complexe

Chair/Président: Linglong Kong
Organizer/Responsable: Linglong Kong

13:30-14:00 **Kai Zhang** (University of North Carolina at Chapel Hill)
Binary Expansion Testing (BET) on Independence / Tests d'expansion binaire sur l'indépendance Ⓔ E

14:00-14:30 **Ruoqing Zhu** (University of Illinois, Urbana-Champaign), **Wenzhuo Zhou** (University of Illinois Urbana-Champaign)
Semiparametric Models for Personalized Dose Finding / Modèles semi-paramétriques pour la détermination de la posologie personnalisée Ⓔ E

14:30-15:00 **Zhengwu Zhang** (University of Rochester), **Xiao Wang** (Purdue University), **Hongtu Zhu** (University of North Carolina), **Linglong Kong** (University of Alberta)
High-Dimensional Spatial Quantile Function-On-Scalar Regression / Régression quantile spatiale fonctions-scalaires en haute dimension Ⓔ E

13:30-15:00 **Invited / Sur invitation** (abstract/résumé 172) **116 (ICT)**

Recent developments in quantitative psychology/psychometrics
Récentes évolutions en psychologie quantitative / psychométrie

Chair/Président: Heungsun Hwang
Organizer/Responsable: Heungsun Hwang

13:30-14:00 **Heungsun Hwang** (McGill University)
Imaging Genetics Structural Equation Modeling for Examining Gene-Brain-Behavioural/Cognitive Relationships / Modélisation d'équations structurelles d'imagerie génétique pour l'étude des relations gène-cerveau-comportement/cognition Ⓔ E

14:00-14:30 **James O. Ramsay** (McGill University), **Marie Wiberg** (Umea University), **Juan Li** (Ottawa Hospital Research Institute)
Efficient Scoring of Test Data / Notation efficace des données d'essai Ⓔ E

14:30-15:00 **Carl F. Falk** (McGill University), **Leah M. Feuerstahler** (Fordham University)
On the Performance of Semi- and Non-Parametric Item Response Functions in Computer Adaptive Tests / De la performance des fonctions semiparamétriques et non paramétriques de réponse aux items dans les tests adaptatifs informatisés Ⓔ E

13:30-15:00 **Invited / Sur invitation** (abstract/résumé 174) **201 (ENG)**

Advanced statistical methods for the integration of omic data
Méthodes statistiques avancées pour l'intégration de données -omiques

Chair/Président: Thierry Chekouo
Organizer/Responsable: Thierry Chekouo
Sponsor/Commanditaires: SSC Biostatistics Section / Groupe de biostatistique de la SSC

13:30-14:00 **Sandra Safo** (University of Minnesota), **Thierry Chekouo** (University of Calgary)
Bayesian Integrative Analysis Method with Incorporation of Grouping Information / Méthode bayésienne intégrative d'analyse avec incorporation d'information de regroupement Ⓔ E

14:00-14:30 **Francesco Claudio Stingo** (University of Florence)
Bayesian Data Integration in Cancer Genomics / Intégration de données bayésiennes dans le domaine de la génomique du cancer Ⓔ E

14:30-15:00
Discussion / Discussion Ⓔ E

13:30-15:00 **Contributed / Communications libres** (abstract/résumé 176) **105 (SB)**

Modeling Risk
Risque de la modélisation

Chair/Président: Shu Li

13:30-13:45 **Étienne Marceau** (Université de Laval), **Hélène Cossette** (Université Laval), **Julien Trufin** (Université Libre de Bruxelles), **Pierre Zuyderhoff** (Université Libre de Bruxelles)
Ruin-Based Risk Measures: Properties and Capital Allocation / Mesures de risque basées sur la ruine : propriétés et allocation de capital Ⓔ E

13:45-14:00 **Shanoja Naik** (Laurentian University), **Peter Adamic** (Laurentian University)
Stochastic Cause-Deleted Life Expectancy for Multiple Risks / Espérance de vie stochastique avec cause de mortalité éliminée pour risques multiples Ⓔ E

14:00-14:15 **Daniel Hadley** (University of British Columbia), **Natalia Nolde** (University of British Columbia), **Harry Joe** (University of British Columbia)
The Selection of Loss Severity Distributions to Model Operational Risk / Choix des distributions de la gravité des pertes pour modéliser le risque opérationnel Ⓔ E

14:15-14:30 **Anne Mackay** (Université du Québec à Montréal), **Michael Kouritzin** (University of Alberta)
Simulating the Heston Model Using Explicit Weak Solutions / Simuler le modèle de Heston en utilisant des solutions faibles explicites Ⓔ E

14:30-14:45 **Jose Garrido** (Concordia University), **Deive Ciro de Oliveira** (ICSA - UNIFAL-MG: Federal University of Alfenas)
Hidden Markov Over-Dispersed Poisson Models Applied to Highways Accident Counts / Modèles de Poisson surdispersés par Markov caché appliqués au dénombrement d'accidents routiers Ⓔ E

13:30-15:00 **Contributed / Communications libres** (abstract/résumé 179) **142 (AD)**

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

Chair/Président: Alomgir Hossain

13:30-13:45 **Yunqi Ji** (Alberta Health Services), **Jerry Ren** (Alberta Health Services), **Geoff Schultz** (Alberta Health Services)
Using Administrative Data for Health Services Planning: Healthcare Utilization Projection / L'emploi de données administratives pour planifier les services de santé : une projection de l'utilisation des soins de santé

13:45-14:00 **Leif Erik Lovblom** (University of Toronto), **Nicholas Mitsakakis** (University of Toronto)
Exponential Dispersion Models for Healthcare Cost Data / Modèles de dispersion exponentielle pour les données relatives au coût des soins de santé

14:00-14:15 **Madeline Ward** (University of Guelph), **Anu Stanley** (University of Guelph), **Lorna Deeth** (University of Guelph), **Rob Deardon** (University of Calgary), **Zeny Feng** (University of Guelph), **Lise Trotz-Williams** (Wellington-Dufferin-Guelph Public Health)
Evaluation of School Absenteeism Surveillance Systems for Influenza Outbreaks in Wellington-Dufferin-Guelph, Ontario / Évaluation des systèmes de surveillance de l'absentéisme scolaire lors d'éclosions de grippes à Wellington-Dufferin-Guelph en Ontario

14:15-14:30 **Justin Wayne Dyck** (University of Manitoba), **Mahmoud Torabi** (University of Manitoba)
Statistical Models for Spatially Misaligned Data: An Application to Ischemic Heart Disease in Manitoba / Modèles statistiques pour données spatiales désalignées : une application aux maladies cardiaques ischémiques au Manitoba

14:30-14:45 **Jinhui Ma** (McMaster University), **Hon Yiu So** (University of Waterloo), **Lauren Griffith** (McMaster University), **Cynthia Balion** (McMaster University), **Mylinh Doung** (McMaster University), **Carol Bassim** (McMaster University), **Chris Verschoor** (McMaster University), **Edwin van den Heuvel** (Eindhoven University of Technology), **Parminder Raina** (McMaster University)
Imputation Strategies for Handling Missing Spirometry Data in Population-Based Studies / Stratégies d'imputation pour le traitement des données spirométriques manquantes dans les études basées sur une population

14:45-15:00 **Roya Gavanji** (University of Saskatchewan), **Cindy Xin Feng** (University of Saskatchewan), **Catherine Trask** (University of Saskatchewan)
Identifying Risk Factors Associated with High Risk of Occupational Injury in Saskatchewan Using Machine Learning Methods / Identifier les facteurs de risque associés au risque élevé d'accidents de travail en Saskatchewan en utilisant des méthodes d'apprentissage machine

---

13:30-15:00 **Contributed / Communications libres** (abstract/résumé 183) **143 (ST)**

Advances in Estimation Methods
Progrès en matière de méthodes d'estimation

Chair/Président: Mélina Mailhot

13:30-13:45 **Lahiru R. Wickramasinghe** (University of Manitoba), **Alexandre Leblanc** (University of Manitoba), **Saman Muthukumarana** (University of Manitoba)
Model-Based Estimation of Baseball Batting Metrics / Estimation fondée sur le modèle des mesures de la performance au bâton au baseball

13:45-14:00 **Shakhawat Hossain** (University of Winnipeg), **Le An Lac** (University of Winnipeg)
Efficient Estimation in Partially Linear Single-Index Models for Binary Longitudinal Data / Estimation efficace dans des modèles à un seul indice partiellement linéaires pour des données longitudinales binaires

14:00-14:15 **Victoire Michal** (Université de Montréal), **David Haziza** (Université de Montréal), **Sixia Chen** (University of Oklahoma)
Efficient Multi-Robust Estimation in the Presence of Influential Units / Estimation efficace multi-robuste en présence d'unités influentes

14:15-14:30 **Julien Miron** (Université de Genève), **Benjamin Poilane** (Université de Genève), **Eva Cantoni** (Université de Genève)
Robust Estimation of Polytomous Logistic Regression Models / Estimation robuste de modèles de régression logistique polytomique

14:30-14:45 **Luc Villandré** (HEC Montreal), **Patrick Brown** (University of Toronto), **Thierry Duchesne** (Université Laval), **Nancy Reid** (University of Toronto), **Jean-François Plante** (HEC Montréal)
Integrated Nested Laplace Approximation (INLA) Estimation for a Spatio-Temporal Regression Model Applicable to Large Datasets / Estimation de l'approximation de Laplace imbriquée (ALI) pour un modèle de régression spatiotemporel applicable à de grands jeux de données

14:45-15:00 **Xiufang Liu** (University of Regina), **Dianliang Deng** (University of Regina), **Dehui Wang** (Jilin University)
Estimating the Quantile Function for History Process with Time-Dependent Covariates and Censoring Mechanism / Estimation de la fonction quantile pour processus historique avec covariables à dépendance chronologique et mécanisme de censure

---

15:30-16:35 **Invited / Sur invitation** (abstract/résumé 186) **201 (ENA)**

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

Chair/Président: Susie Fortier
Organizer/Responsable: Susie Fortier
Sponsor/Commanditaires: SSC Survey Methods Section / Groupe des méthodes d'enquête de la SSC

15:30-16:35 **Jack Gambino** (Statistics Canada)
The Evolving Role of Non-Survey Data in Official Statistics / L'évolution du rôle des données non issues d'enquêtes dans les statistiques officielles

---

15:30-17:00 **Invited / Sur invitation** (abstract/résumé 187) **116 (ICT)**

Analytics in Sports
Analyse sportive

Chair/Président: Shirley Mills
Organizer/Responsable: Shirley Mills
Sponsor/Commanditaires: SSC Business & Industrial Statistics Section / Groupe de statistique industrielle et de gestion de la SSC

15:30-16:00 **Tim B. Swartz** (Simon Fraser University), **Rajitha Silva** (University of Sri Jayewardenepura), **Lucas Wu** (Simon Fraser University), **Joan Hu** (Simon Fraser University)
Soccer Insights / Regards sur le soccer

16:00-16:30 **Michael E. Schuckers** (St. Lawrence University)
Statistical Analysis of the National Hockey League Entry Draft / Analyse statistique du repêchage dans la Ligue nationale de hockey

16:30-17:00 **Nathan Sandholtz** (Simon Fraser University), **Jacob Mortensen** (Simon Fraser University), **Luke Bornn** (Sacramento Kings; Simon Fraser University)
Measuring Spatial Allocative Efficiency in Professional Basketball / Mesure de l'allocation spatiale optimale des ressources dans le domaine du basketball professionnel

---

15:30-17:00 **Invited / Sur invitation** (abstract/résumé 189) **142 (AD)**

Tenure and Promotion: Insightful Tips from the Applicants and Reviewers
Permanence et promotion : conseils de candidats et d'évaluateurs

Chair/Président: Hua Shen
Organizer/Responsable: Hua Shen
Sponsor/Commanditaires: SSC New Investigator Committee / Comité des nouveaux chercheurs de la SSC

15:30-17:00 **Hua Shen** (University of Calgary), **Richard Lockhart** (Simon Fraser University), **Xikui Wang** (University of Manitoba), **Wendy Lou** (University of Toronto), **Louis-Paul Rivest** (Université Laval), **Hugh Chipman** (Acadia University), **Linglong Kong** (University of Alberta)
Tenure and Promotion: Insightful Tips from the Applicants and Reviewers / Titularisation et promotion : des conseils judicieux de la part de candidats et d'évaluateurs Ⓔ E

15:30-17:00 **Invited / Sur invitation** (abstract/résumé 190) **102 (ICT)**

New perspectives and challenges in analysis of linked genomic and phenomic data
Nouvelles perspectives et nouveaux défis en analyse de données génomiques et phénomiques couplées

Chair/Président: Jinko Graham
Organizer/Responsable: Jinko Graham

15:30-16:00 **Quan Long** (University of Calgary), **Shengjie Lu** (University of Calgary), **Chen Cao** (University of Calgary), **Zhi Xiong** (Shantou University), **Xuewen Lu** (University of Calgary)
Less-Is-More and More-Is-Less in Integrating Multi-Scale Omics with Polygenic Phenotype Predictors / « Moins donne plus » et « plus donne moins » dans l'intégration de l'omique multiéchelle dans les prédicteurs de phénotype polygéniques Ⓔ E

16:00-16:30 **Lloyd T Elliott** (Simon Fraser University)
Towards Modern Machine Learning for Genome-Wide Association / Vers un apprentissage machine moderne pour l'association pangénomique Ⓔ E

16:30-17:00 **Kun Liang** (University of Waterloo), **Yu Gao** (University of Waterloo)
Controlling the False Discovery Rate of GWAS / Contrôle du taux de fausses découvertes dans les études d'association pangénomiques (GWAS) Ⓔ E

15:30-17:00 **Invited / Sur invitation** (abstract/résumé 192) **146 (SB)**

Advances in model-based clustering of complex data
Progrès des méthodes de groupage par modèle pour données complexes

Chair/Président: Linglong Kong
Organizer/Responsable: Bei Jiang

15:30-16:00 **Thomas Brendan Murphy** (University College Dublin), **Michael Fop** (University College Dublin), **Luca Scrucca** (Università degli Studi di Perugia)
Model-Based Clustering with Sparse Covariance Matrices / Regroupement selon un modèle avec des matrices de covariance éparses Ⓔ E

16:00-16:30 **Bei Jiang** (University of Alberta), **Adrian E. Raftery** (University of Washington), **Russell J. Steele** (McGill University), **Naisyin Wang** (University of Michigan)
Balancing Inferential Integrity and Disclosure Risk: A Mixture Modeling Approach / Équilibrer l'intégrité de l'inférence et le risque de divulgation : une approche de modélisation par mélange Ⓔ E

16:30-17:00 **Gongjun Xu** (University of Michigan), **Yuqi Gu** (University of Michigan)
Learning Attribute Patterns in High-Dimensional Structured Latent Attribute Models / Apprentissage des structures d'attributs dans les modèles à attributs latents structurés en haute dimension Ⓔ E

15:30-17:00 **Invited / Sur invitation** (abstract/résumé 194) **109 (SS)**

CANSSI Postdoctoral Showcase
Présentations des stagiaires postdoctoraux de l'INCASS

Chair/Président: W. John Braun
Organizer/Responsable: W. John Braun
Sponsor/Commanditaires: CANSSI / INCASS

15:30-16:00 **Whitney K. Huang** (University of Victoria), **Francis Zwiers** (University of Victoria), **Adam Monahan** (University of Victoria)
Modeling Compound Wind and Precipitation Extremes Using a Large Climate Model Ensemble / Modélisation de vents et de précipitations extrêmes en utilisant un ensemble de grands modèles climatiques Ⓔ E

16:00-16:30 **David Soave** (Ontario Institute for Cancer Research), **Jerry Lawless** (University of Waterloo)
Regularized Regression Methods for Two Phase Studies / Méthodes de régression régularisée pour les études en deux phases Ⓔ E

16:30-17:00 **Luc Villandré** (McGill University), **Aurélie Labbe** (HEC Montréal), **Ilinca-Ruxandra Ibanescu** (Jewish General Hospital) **Isabelle Hardy** (Centre Hospitalier de l'Université de Montréal), **Bluma Brenner** (Jewish General Hospital), **Michel Roger** (Centre Hospitalier de l'Université de Montréal), **David Stephens** (McGill University)
HIV transmission cluster inference using Bayesian phylogenetics / Inférence des grappes de transmission du VIH par l'intermédiaire de la phylogénétique bayésienne Ⓔ E

15:30-17:00 **Contributed / Communications libres** (abstract/résumé 196) **201 (ENG)**

New Approaches for Functional and Longitudinal Data
Nouvelles approches des données fonctionnelles et longitudinales

Chair/Président: Cindy Xin Feng

15:30-15:45 **Janie Coulombe** (McGill University), **Erica Moodie** (McGill University), **Robert Platt** (McGill University)
Two Weighted Estimators for the Treatment Effect that Account for Covariate-Dependent Monitoring Times and Confounding in Longitudinal Studies / Deux estimateurs pondérés pour l'effet de traitement qui considèrent les temps de mesure informatifs et les effets confondants dans les études longitudinales Ⓔ E

15:45-16:00 **Md Erfanul Hoque** (University of Manitoba), **Elif Acar** (University of Manitoba), **Mahmoud Torabi** (University of Manitoba)
A D-Vine Copula Model for Unbalanced and Unequally Spaced Longitudinal Data / Un modèle de copule en vigne D pour les données longitudinales déséquilibrées et inégalement espacées Ⓔ E

16:00-16:15 **Adam Kashlak** (University of Alberta)
Symmetrization for Exact Nonparametric Functional ANOVA / Symétrisation pour analyse de variance fonctionnelle non paramétrique exacte Ⓔ E

16:15-16:30 **Marie-Hélène Descary** (Université du Québec à Montréal)
Recovering Covariance from Functional Fragments / Reconstituer la fonction de covariance à partir de fragments de données fonctionnelles Ⓔ E

16:30-16:45 **Yijun Xie** (University of Waterloo), **Adam Kolkiewicz** (University of Waterloo), **Greg Rice** (University of Waterloo)
Functional Normality Test / Test de normalité fonctionnel Ⓔ E

15:30-17:00 **Contributed / Communications libres** (abstract/résumé 199) **105 (SB)**

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

Chair/Président: Nicholas Mitsakakis

15:30-15:45 **Qihuang Zhang** (University of Waterloo), **Grace Yi** (University of Waterloo)
Analysis of Bivariate Responses in Genetic Association Studies with Measurement Error and Misclassification / Analyse de réponses bivariées dans des études d'association génétique avec erreurs de mesure et classification erronée Ⓔ E

15:45-16:00 **Joycelyne E Ewusie** (University of Ottawa), **Kelly Burkett** (University of Ottawa), **Marie-Hélène Roy-Gagnon** (University of Ottawa)
Using External Controls to Account for Mating Asymmetry in Maternal Genetic Association / Prendre en compte l'asymétrie d'accouplement dans les études d'effets génétiques maternels en utilisant des donnée externes de parents témoins Ⓔ E

16:00-16:15 **Mei Dong** (University of Saskatchewan), **Longhai Li** (University of Saskatchewan), **Lloyd Balbuena** (University of Saskatchewan)
Using External Cross Validation for Measuring Predictivity of Selected Features with Application to Genome Wide Predictive Analysis for Alzheimer's Disease / Validation croisée externe pour mesurer la prédictivité de certaines caractéristiques et application à l'analyse prédictive pangénomique de la maladie d'Alzheimer Ⓔ E

16:15-16:30 **Changjiang Xu** (University of Toronto), **Gary Bader** (University of Toronto), **Veronique Voisin** (University of Toronto), **Ruth Isserlin** (University of Toronto), **Jeff Liu** (University of Toronto)
Gene Set Analysis Using GSEA and GSVA: Performance Comparison and Improvement / Analyse d'ensembles de gènes à l'aide des méthodes d'enrichissement des ensembles de gènes (GSEA) et de la variation des ensembles de gènes (GSVA) : comparaison et amélioration de la performance Ⓔ E

16:30-16:45 **Osvaldo Espin-Garcia** (Dalla Lana School of Public Health, University of Toronto / Lunenfeld-Tanenbaum Research Institute), **Radu Craiu** (University of Toronto), **Shelley Bull** (University of Toronto & Lunenfeld-Tanenbaum Research Institute)
Optimal Two-Phase Designs in Post-Genome-Wide Association Studies (GWAS) / Plan optimal en deux phases pour les études d'association post-pangénomiques (GWAS) Ⓔ E

15:30-17:00 **Contributed / Communications libres** (abstract/résumé 202) **143 (ST)**

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

Chair/Président: Anita Brobbey

15:30-15:45 **Shabnam Fani** (University of Calgary), **Hua Shen** (University of Calgary), **Xuewen Lu** (University of Calgary), **Jingjing Wu** (University of Calgary)
Semiparametric Regression with the U-Shaped Baseline Hazard Function in the Additive Hazards Model / Régression semiparamétrique avec fonction de risque de base en forme de U dans le modèle de risque additif Ⓔ E

15:45-16:00 **Longlong Huang** (University of the Fraser Valley), **Karen Kopciuk** (Cancer Control Alberta; University of Calgary), **Xuewen Lu** (University of Calgary)
Adaptive Group Bridge Selection in the Semiparametric Accelerated Failure Time Model / Sélection bridge groupé adaptive dans le modèle semiparamétrique du temps de défaillance accéléré Ⓔ E

16:00-16:15 **Jingyu Cui** (University of New Brunswick), **Renjun Ma** (University of New Brunswick), **M. Tariq Hasan** (University of New Brunswick)
Generalized Linear Mixed Model with Crossed Random Effects / Modèles linéaires mixtes généralisés avec effets croisés aléatoires Ⓔ E

16:15-16:30 **Kaida Cai** (University of Calgary), **Xuewen Lu** (University of Calgary), **Hua Shen** (University of Calgary)
Bi-Level Variable Selection for Multivariate Failure Time Data with Observed Heterogeneity / Sélection de variables à deux niveaux pour les données multivariées de temps de défaillance avec hétérogénéité observée Ⓔ E

Wednesday May 29 mercredi 29 mai

08:40-09:50 **Invited / Sur invitation** (abstract/résumé 205) **102 (ICT)**

SSC 2018 Impact Award Address
2018 Prix pour impact de la SSC

Chair/Président: Peter X Song
Organizer/Responsable: Carl James Schwarz

08:40-09:50 **Geneviève Gauthier** (HEC Montreal)
The Use of Filters and Large Databases in Financial Engineering / L'utilisation des filtres et des grandes bases de données en ingénierie financière Ⓔ E

08:40-09:50 **Invited / Sur invitation** (abstract/résumé 206) **148 (ST)**

CRM-SSC Prize in Statistics invited Address
Allocution de la récipiendaire du Prix CRM-SSC en statistique

Chair/Président: Alejandro Murua

08:40-09:50 **Johanna G. Neslehova** (McGill University)
Tales of tails, tiles and ties in dependence modeling / La queue, la tuile, le bris d'égalité et leur rôle dans les modèles de dépendance Ⓔ E

10:20-11:50 **Invited / Sur invitation** (abstract/résumé 207) **201 (ENG)**

Recent Advances in Risk Theory
Dernières avancées en théorie des risques

Chair/Président: Bin Li
Organizer/Responsable: Bin Li
Sponsor/Commanditaires: SSC Actuarial Science Section / Groupe de science actuarielle de la SSC

10:20-10:50 **Alexey Kuznetsov** (York University), **Runhuan Feng** (University of Illinois at Urbana-Champaign), **Fenghao Yang** (Royal Bank of Canada)
Exponential Functionals of Levy Processes and Variable Annuity Guaranteed Benefits / Fonctionnelles exponentielles de processus Levy et annuités variables avec prestations garanties Ⓔ E

10:50-11:20 **Jiandong Ren** (Western University), **Wenjun Jiang** (Western University), **Hanping Hong** (Western University)
Reinsurance Policies with the Maximal Synergy Potential / Polices de réassurance ayant le potentiel maximal de synergie Ⓔ E

11:20-11:50 **Yi Lu** (Simon Fraser University), **Shuanming Li** (University of Melbourne), **Kristina Sendova** (University of Western Ontario)
The Expected Discounted Penalty Functions: From Infinite Time to Finite Time / Les fonctions de pénalité actualisées attendues : d'un temps infini à fini Ⓔ E

10:20-11:50 **Invited / Sur invitation** (abstract/résumé 209) **146 (SB)**

Extreme values
Valeurs extrêmes

Chair/Président: Gail B. Ivanoff
Organizer/Responsable: Gail B. Ivanoff
Sponsor/Commanditaires: SSC Probability Section / Groupe de probabilité de la SSC

10:20-10:40 **Stilian Stoev** (University of Michigan, Ann Arbor), **Zheng Gao** (University of Michigan, Ann Arbor)
Concentration of Maxima and the Fundamental Limits of Exact Support Recovery in High Dimensions / Concentration des maxima et les limites fondamentales du support de redressement exact en haute dimension Ⓔ E

10:40-11:00 **Clemonell Lord Baronat Bilayi-Biakana** (University of Ottawa), **Gail Ivanoff** (University of Ottawa), **Rafal Kulik** (University of Ottawa)
Heavy-Tailed Long Memory Stochastic Volatility Model with Leverage / Modèle de volatilité stochastique à mémoire longue et à queue lourde avec effet de levier

11:00-11:20 **Natalia Nolde** (University of British Columbia), **Jinyuan Zhang** (INSEAD)
Conditional Extremes in Asymmetric Financial Markets / Les extrêmes conditionnels dans les marchés financiers asymétriques

11:20-11:40 **Rafal Kulik** (University of Ottawa), **Philippe Soulier** (Paris-Nanterre)
Limit Theorems for Empirical Cluster Functionals / Théorèmes limites pour les fonctionnelles de grappe empiriques

---

10:20-11:50 **Invited / Sur invitation** (abstract/résumé 211) **102 (ICT)**

Building the Pipeline: the International Data Science in Schools Project
Construire l'avenir : le projet "International Data Science in Schools Project"

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

10:20-10:50 **Alison L. Gibbs** (University of Toronto)
Introduction to the International Data Science in Schools Project / Introduction au Projet International de la Science des Données dans les Écoles

10:50-11:20 **Wesley Burr** (Trent University)
Case Studies in Data Science Education: Limits and Scope / Études de cas dans l'enseignement de la science des données : limites et champ d'application

11:20-11:50 **Robert Gould** (UCLA)
Implementing a Data Science Course in Secondary Schools / Instaurer un cours sur la science des données dans les écoles secondaires

---

10:20-11:25 **Invited / Sur invitation** (abstract/résumé 213) **122 (ICT)**

CJS Award Address
Allocution du récipiendaire du Prix de la RCS

Chair/Président: Louis-Paul Rivest
Organizer/Responsable: Louis-Paul Rivest

10:20-11:25 **Radu Craiu** (University of Toronto)
LISA for BART / Algorithme d'échantillonnage de vraisemblance gonflée pour les arbres additifs de régression bayésienne

---

10:20-11:50 **Invited / Sur invitation** (abstract/résumé 214) **142 (AD)**

Statistical challenges and methods for clinical trial design
Défis et méthodes statistiques pour la conception d'essais cliniques

Chair/Président: Depeng Jiang
Organizer/Responsable: Depeng Jiang

10:20-10:50 **Ying Yuan** (University of Texas MD Anderson Cancer Center), **Ruitao Lin** (University of Texas MD Anderson Cancer Center), **Daniel Li** (Juno Therapeutics), **Lei Nie** (Food and Drug Administration), **Katherine Warren** (National Cancer Institute)
Time-To-Event Bayesian Optimal Interval Design to Accelerate Phase I Trials / Plan d'intervalle optimal bayésien de temps avant l'évènement pour accélérer les essais de phase I

10:50-11:20 **Bo Huang** (Pfizer), **Xiaodong Luo** (Sanofi), **Hui Quan** (Sanofi)
Design and Monitoring of Survival Trials in the Presence of Non-Proportional Hazard / Conception et suivi d'essais de survie en présence de risque non-proportionnel Ⓔ E

11:20-11:50 **Suyu Liu** (MD Anderson Cancer Center), **Beibei Guo** (Louisiana State University), **Ying Yuan** (MD Anderson Cancer Center)
A Bayesian Phase I/II Trial Design for Immunotherapy / Un plan d'essai bayésien de phases I et II pour l'immunothérapie Ⓔ E

10:20-11:50 **Contributed / Communications libres** (abstract/résumé 216) **109 (SS)**

Advances in Distribution Theory
Progrès en matière de théorie de la distribution

Chair/Président: Alberto Nettel-Aguirre

10:20-10:35 **Liu Yi** (University of Alberta), **Peng Liu** (University of Alberta), **Rui Zhu** (University of Alberta), **Linglong Kong** (University of Alberta), **Bei Jiang** (University of Alberta), **Di Niu** (University of Alberta)
Optimal Smooth Approximation for Quantile Matrix Factorization / Approximation optimale lisse pour la factorisation quantile de matrice Ⓔ E

10:35-10:50 **Yuan Sun** (University of Michigan, Ann Arbor), **Xuming He** (University of Michigan, Ann Arbor)
A Model-Based Bootstrap Method for Regional Quantiles Treatment Effects Detection with a Quantile Regression Rank Test / Méthode bootstrap fondée sur un modèle afin de détecter des effets de traitement sur les quantiles régionaux avec un test de rang par régression quantile Ⓔ E

10:50-11:05 **Matthew Pietrosanu** (University of Alberta), **Dengdeng Yu** (University of Toronto), **Linglong Kong** (University of Alberta)
Extending Partial Quantile Regression to Multidimensional Functional Linear Models via Tensor Decomposition / Extension de la régression quantile partielle aux modèles linéaires fonctionnels multidimensionnels par la décomposition tensorielle Ⓔ E

11:05-11:20 **Yi Lian** (McGill University), **Yi Yang** (McGill University), **Robert Platt** (McGill University)
Tweedie Compound Poisson Model in the Reproducing Kernel Hilbert Space / Modèle de Poisson composé Tweedie appliqué à l'espace de Hilbert à noyau reproduisant Ⓔ E

11:20-11:35 **René Ferland** (Université du Québec à Montréal), **François Watier** (Université du Québec à Montréal)
Goal Achieving Probabilities of Regime-Switching Mean-Variance Portfolios / Probabilités de réussite des objectifs de portefeuilles moyenne-variance à changement de régime Ⓕ E

11:35-11:50 **Salma Saad** (University of Regina), **Andrei Volodin** (University of Regina)
Asymptotic Analysis of Method of Moments / Analyse asymptotique de la méthode des moments Ⓔ E

10:20-11:50 **Contributed / Communications libres** (abstract/résumé 219) **201 (ENA)**

Novel Biostatistical Methods
Nouvelles méthodes biostatistiques

Chair/Président: Angelo J. Canty
Sponsor/Commanditaires: /

10:20-10:35 **Thierry Chekouo** (University of Calgary), **Himadri Mukherjee** (University of Minnesota Duluth)
Model-Based Clustering and Gene Selection via Bayesian Hierarchical Hidden Markov Models / Regroupement fondé sur un modèle et sélection des gènes via des modèles hiérarchiques bayésiens de Markov cachés Ⓔ E

10:35-10:50 **Maryam Yetunde Onifade** (University of Ottawa), **Kelly Burkett** (University of Ottawa)
Comparison of Mixed Model-Based Approaches for Correcting for Population Substructure with Application to Extreme Phenotype Sampling / Comparaison d'approches à modèles mixtes pour corriger l'effet de la sous-structure de la population et application à l'échantillonnage de phénotypes extrêmes

10:50-11:05 **Wendimagegn Alemayehu** (University of Alberta), **Cynthia Westerhout** (University of Alberta)
On Statistical Modeling of the Relationship of Temporal Change of Biomarkers with Clinical Outcomes / La modélisation statistique de la relation du changement temporel des biomarqueurs au moyen de résultats cliniques.

11:05-11:20 **Changchang Xu** (University of Toronto), **Shelley Bull** (University of Toronto), **Shelley Bull** (University of Toronto)
Improving Mixture Cure Modelling of Molecular Genetic Biomarkers in Cancer Prognosis by Penalized Maximum Likelihood / Amélioration de la modélisation de traitement par mélange des biomarqueurs génétiques moléculaires dans le prognostic du cancer par le maximum de vraisemblance pénalisé

11:20-11:35 **Rajib Dey** (McGill University), **Paramita Saha Chaudhuri** (McGill University)
Estimation of Time-Dependent Predictive Accuracy in the Presence of Competing Risks / Estimation de l'exactitude prédictive dépendante du temps en présence de risques concurrents

---

10:20-11:50 **Contributed / Communications libres** (abstract/résumé 222) **119 (SA)**

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

Chair/Président: Daniel Zi Yang

10:20-10:35 **Mili Roy** (University of Calgary)
Joint Analysis of Correlated Non-Gaussian Continuous Outcomes: Impact of Conditional Dependence / Analyse conjointe de résultats continus non gaussiens corrélés : impact de la dépendance conditionnelle

10:35-10:50 **James G. MacKinnon** (Queen's University), **Morten O. Nielsen** (Queen's University), **Matthew D. Webb** (Carleton University)
Wild Bootstrap Inference with Multiway Clustering / Inférence bootstrap sauvage avec regroupements multivoies

10:50-11:05 **Ismaila Ba** (Université du Québec à Montréal), **Jean François Coeurjolly** (Université du Québec à Montréal)
Regularization Techniques for Inhomogeneous Gibbs Point Process Models with a Diverging Number of Covariates / Méthodes de régularisation pour des processus ponctuels de Gibbs inhomogènes avec un nombre divergent de covariables

11:05-11:20 **Saumen Mandal** (University of Manitoba)
Optimal Designs Subject to Achieving Equality of Variances of the Estimators of Linear Functions of Parameters / Conceptions optimales soumises à l'atteinte de l'égalité des variances des estimateurs de fonctions linéaires des paramètres

11:20-11:35 **Harlan Campbell** (University of British Columbia), **Daniel Lakens** (Eindhoven University of Technology)
Can We Disregard the Whole Model? Non-Inferiority Testing for Omnibus Effects in Linear Models / Peut-on ne pas prendre en compte le modèle entier ? Tests de non infériorité des effets omnibus dans les modèles linéaires

13:30-15:00 **Invited / Sur invitation** (abstract/résumé 225) **119 (SA)**

Best Practices in Experiential Learning
Pratiques exemplaires en apprentissage expérientiel

Chair/Président: Sohee Kang
Organizer/Responsable: Sohee Kang
Sponsor/Commanditaires: SSC Statistical Education Section / Groupe d'éducation en statistique de la SSC

13:30-14:00 **Shirley Mills** (Carleton University)
Experiential Learning via Co-ops and Internships / L'apprentissage expérientiel par les coopératives et les stages

14:00-14:30 **Albert Y. Kim** (Smith College)
Moderndive: Statistical Inference Using the Tidyverse / Moderndive : l'inférence statistique avec le tidyverse

14:30-15:00 **Nathan A. Taback** (University of Toronto)
ASA DataFest@UofT and Beyond / ASA DataFest@UofT, et plus encore

---

13:30-14:35 **Invited / Sur invitation** (abstract/résumé 227) **102 (ICT)**

Isobel Loutit Lecture
Allocution Isobel Loutit

Chair/Président: Chunfang Lin
Organizer/Responsable: Chunfang Lin

13:30-14:35 **Max D. Morris** (Iowa State University)
A Brief History of Statistical Computer Experiments / Bref historique des expériences informatiques en statistique

---

13:30-15:00 **Invited / Sur invitation** (abstract/résumé 228) **122 (ICT)**

Data Integration and Distributed Inference
Intégration de données et inférence distribuée

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

13:30-14:00 **Peter X Song** (University of Michigan)
Integrative Data Analytics via Distributed Inference Functions / Analyse intégrative des données au moyen des fonctions d'inférence distribuée

14:00-14:30 **Su Chen** (University of Memphis), **Wilfried Karmaus** (University of Memphis)
A Nonparametric Test of Variance Heterogeneity in DNA Methylation Influenced by Genetic Variants / Un test non paramétrique de l'hétérogénéité de la variance dans une méthylation de l'ADN influencée par des variants génétiques

14:30-15:00 **You Liang** (Ryerson University), **Xikui Wang** (University of Manitoba), **Lysa Porth** (University of Manitoba)
Risk Management for Heavy Tailed and Tail Dependent Claims / Gestion du risque pour les réclamations à queue lourde et à dépendance de queue

---

13:30-15:00 **Invited / Sur invitation** (abstract/résumé 230) **109 (SS)**

Recent progress for quantile regression analysis
Récents progrès en analyse par régression quantile

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

13:30-14:00 **Dianliang Deng** (University of Regina), **Mashfiqul Chowdhury** (University of Regina), **Mashfiqul Chowdhury** (University of Regina)
Quantile Regression Analysis for Gene Expression Data / Analyse de régression quantile pour données d'expression génique

14:00-14:30 **Mei Ling Huang** (Brock University), **Jenny Tieu** (Brock University)
A Nonparametric Quantile Regression Method / Méthode de régression quantile non paramétrique

14:30-15:00 **Mohammad Jafari Jozani** (University of Manitoba)
More Efficient Quantile Regression Analysis Using Rank Information / Analyse plus efficace de la régression quantile en utilisant l'information sur le rang

---

13:30-14:35 **Invited / Sur invitation** (abstract/résumé 232) **144 (SB)**

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

Chair/Président: Gordon H Fick

13:30-14:35 **Peijun Sang** (University of Waterloo)
Sparse Estimation for Functional Semiparametric Additive Models / Estimation éparse pour des modèles additifs semi-paramétriques fonctionnels

---

13:30-15:00 **Contributed / Communications libres** (abstract/résumé 233) **201 (ENA)**

Methods for High-Dimensional and Large Data II
Méthodes pour traiter les données volumineuses et de grande dimension II

Chair/Président: Mohammad Ehsanul Karim

13:30-13:45 **Patrick Fournier** (Université du Québec à Montréal), **Fabrice Larribe** (Université du Québec à Montréal)
Modelling Horizontal Gene Transfer in Bacteria via Conditional Ancestral Recombination Graphs / Modélisation du transfert horizontal de gènes dans les bactéries via tableau de recombinaison ancestral conditionnel

13:45-14:00 **Yuming Zhang** (University of Geneva), **Stéphane Guerrier** (University of Geneva), **Maria-Pia Victoria-Feser** (University of Geneva), **Mucyo Karemera** (Pennsylvania State University), **Samuel Orso** (University of Geneva)
A Study of Simulation Based Estimators for High Dimensional Generalized Linear Models / Une étude d'estimateurs basés sur la simulation pour des modèles linéaires généralisés de grande dimension

14:00-14:15 **Wei Tu** (University of Alberta), **Linglong Kong** (University of Alberta), **Zhihua Su** (University of Florida), **Rohana Karunamuni** (University of Alberta)
Envelope-Based High-Dimensional Gaussian Copula Regression / Régression de copule gaussienne de grande dimension avec une méthode d'enveloppes

14:15-14:30 **Sonja Surjanovic** (University of British Columbia), **William J. Welch** (University of British Columbia)
Gaussian Process Regression with Large Datasets / Régression par processus gaussien avec de grands ensembles de données

14:30-14:45 **Gyanendra Pokharel** (University of Calgary), **Paula Robson** (Alberta Health Services), **Lorriane Shack** (University of Calgary), **John Spinelli** (BC Cancer Agency), **Karen Kopciuk** (Alberta Health Services)
Dimensionality Reduction and Stage Shifting by Modifying Determinants of Cancer at Diagnosis / Réduction de la dimensionnalité et dépistage par étapes en modifiant les déterminants du cancer au moment du diagnostic

14:45-15:00 **Min Zhang** (University of South China), **Xuewen Lu** (University of Calgary)
Intelligent Search of Radiation Source and Its Optimization / Recherche intelligente des sources de rayonnement et son optimisation

---

13:30-15:00 **Contributed / Communications libres** (abstract/résumé 237) **146 (SB)**

Statistical Issues for Longitudinal and Time Series Analyses
Problèmes statistiques liés aux analyses longitudinales et de séries temporelles

Chair/Président: Lei Sun

13:30-13:45 **Melody Ghahramani** (The University of Winnipeg), **Scott White** (University of Manitoba)
Time Series Regression for Zero-Inflated and Overdispersed Count Data: A Functional Response Model Approach / Régression chronologique pour les données de dénombrement à surreprésentation de zéros et surdispersées : approche modélisée de réponse fonctionnelle

13:45-14:00 **Olawale Ayilara** (University of Manitoba), **Tolulope Sajobi** (University of Calgary), **Lisa Lix** (University of Manitoba)
Goodness-Of-Fit Indices for Testing Longitudinal Measurement Invariance in Ordinal Data / Indices d'adéquation pour tester l'invariance de la mesure longitudinale dans les données ordinales

14:00-14:15 **Jia Li** (University of Calgary), **Alexander de Leon** (University of Calgary), **Haocheng Li** (University of Calgary and Roche Canada)
Likelihood Analysis of Gaussian Copula Mixed Models for Multiple Correlated Disparate Longitudinal Non-Gaussian Continuous Outcomes / Analyse de vraisemblance des modèles mixtes à copules gaussiennes pour de multiples résultats disparates et corrélés, longitudinaux et non gaussiens continus

14:15-14:30 **Amadou Diogo Barry** (Université du Québec à Montréal), **Karim Oualkacha** (Université du Québec à Montréal), **Arthur Charpentier** (Université du Québec à Montréal)
Penalized Weighted Asymmetric Least Squares Regression for Longitudinal Data with Fixed-effects / Régression au moindre carré asymétrique pondérée et pénalisée pour les données longitudinales avec effets fixes

14:30-14:45 **Julan Al-Yassin** (University of Windsor), **Richard Caron** (University of Windsor), **Robin Gras** (University of Windsor)
Time Series: Stochastic or Chaotic? / Séries chronologiques : stochastiques ou chaotiques ?

14:45-15:00 **Mohsen Soltanifar** (University of Toronto)
A Time Series Based Point Estimation of Stop Signal Reaction Times (SSRT) / Estimation ponctuelle chronologique des temps de réaction du signal d'arrêt

---

13:30-15:00 **Contributed / Communications libres** (abstract/résumé 241) **142 (AD)**

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

Chair/Président: Lisa M. Lix

13:30-13:45 **Selvakkadunko Selvaratnam** (University of Alberta), **Linglong Kong** (University of Alberta), **Douglas Wiens** (University of Alberta)
The Impact of Scale Functions on the Construction of Robust Designs for Nonlinear Quantile Regression / L'incidence des fonctions d'échelle sur la construction de plans robustes de régression quantile non linéaire

13:45-14:00 **Gun Ho Jang** (Ontario Institute for Cancer Research)
Tail Probability of Maximal Chi-Squared Statistic for Association Studies / La probabilité de queue d'une statistique au chi carré maximale en études d'association

14:00-14:15 **Cindy Xin Feng** (University of Saskatchewan)
Modelling Count Data with Excessive Zeros: Does the Choice Between Zero-Inflated Model and Hurdle Model Matter? / Modéliser des données de dénombrement avec surreprésentation de zéros : le choix entre un modèle à surreprésentation de zéros et un modèle ≪hurdle≫ est-il important ?

14:15-14:30 **Yidan Shi** (University of Waterloo), **Leilei Zeng** (University of Waterloo), **Mary Thompson** (Univeristy of Waterloo), **Suzanne Tyas** (Univeristy of Waterloo)
Mixture Hidden Markov Model with Partially Observed Component Memberships / Modèle de Markov caché par mélange avec composantes d'appartenance partiellement observées

14:30-14:45 **Zheng Fan** (University of Calgary), **Hua Shen** (University of Calgary), **Haocheng Li** (University of Calgary)
Causal Inference with Misclassification in Confounding Variables / Inférence causale avec classification erronée des variables de confusion

---

13:30-15:00 **Contributed / Communications libres** (abstract/résumé 244) **113 (SS)**

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

Chair/Président: Johanna G. Neslehova

13:30-13:45 **Lenin Arango-Castillo** (Queen's University), **Glen Takahara** (Queen's University)
Long-Range Dependence Parameter Estimation for Mixed Spectra Gaussian Processes / Estimation du paramètre de dépendance de longue portée pour processus gaussien à spectres mixtes

13:45-14:00 **Katherine Burak** (University of Calgary), **Alexander de Leon** (University of Calgary)
Cluster analysis of non-Gaussian data via mixtures of Gaussian copula distributions / Analyse de groupement de données mixtes au moyen de mélanges de distributions de copules gaussiennes

14:00-14:15 **Marie-Pier Côté** (Université Laval), **Christian Genest** (McGill University)
Dependence in a Background Risk Model / La dépendance dans le modèle de risque contextuel

14:15-14:30 **Devan G Becker** (University of Western Ontario), **Douglas G. Woolford** (Western University), **Charmaine B. Dean** (University of Waterloo)
A Joint-Modelling Framework for Inducing Dependence in Compound Poisson Models for Aggregate Losses with Application to Wildland Fire / Cadre de modélisation conjointe pour l'induction d'une dépendance dans les modèles de processus de Poisson composé pour les pertes globales, avec application à un feu de végétation

14:30-14:45 **Haoxin Zhuang** (University of Waterloo), **Liqun Diao** (University of Waterloo), **Grace Yi** (University of Waterloo)
Composite Likelihood Methods for Analyzing Longitudinal Data with Periodic Patterns under Vine Copula Models / Méthodes de vraisemblance composée pour l'analyse de données longitudinales à structure périodique dans les modèles de copules en vignes

14:45-15:00 **Ce Zhang** (University of Calgary), **Xuewen Lu** (University of Calgary)
Efficient Estimation of the Additive Hazards Model with Bivariate Current Status Data / Estimation efficace du modèle à risques additifs avec données d'état actuel bivariées

---

15:30-17:00 **Invited / Sur invitation** (abstract/résumé 248) **144 (SB)**

Integration of probability and non-probability samples
Intégration d'échantillons probabilistes et non probabilistes

Chair/Président: Susie Fortier
Organizer/Responsable: Jean-François Beaumont
Sponsor/Commanditaires: SSC Survey Methods Section / Groupe des méthodes d'enquête de la SSC

15:30-16:00 **Changbao Wu** (University of Waterloo), **Yilin Chen** (University of Waterloo), **Pengfei Li** (University of Waterloo)
Sample Matching and Double Robust Estimation with Non-Probability Samples / Comparaison d'échantillons et estimation doublement robuste avec échantillonnages non probabilistes

16:00-16:30 **Kenneth C.K. Chu** (Statistics Canada), **Jean-François Beaumont** (Statistics Canada)
Formation of Homogeneous Self-Selection Propensity Classes for Non-Probability Samples via Probability Samples / La formation de classes homogènes de la propension à l'autosélection pour des échantillons non probabilistes via des échantillons probabilistes

16:30-17:00 **Marie-Hélène Felt** (Bank of Canada), **Heng Chen** (Bank of Canada), **Christopher Henry** (Bank of Canada)
Calibration and Variance Estimation for Non-Probability Samples: An Application to the 2017 Bank of Canada Methods-Of-Payment Survey / Calibrage et estimation de la variance des échantillons non-probabilistes : le cas de l'enquête 2017 de la Banque du Canada sur les modes de paiement

---

15:30-17:00 **Invited / Sur invitation** (abstract/résumé 250) **201 (ENG)**

Statistical Mining with Complex and Noisy Data
Exploitation statistique avec données complexes et bruitées

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

15:30-16:00 **Zhao Chen** (Fudan University), **Zhanxiong Xu** (Penn State University), **Zhibiao Zhao** (Penn State University)
Efficient Estimation for Nonlinear Heteroscedastic Models Through Quantile Regression / Estimation efficace pour modèles hétéroscédastiques non-linéaires par la régression quantile

16:00-16:30 **Yi Yang** (McGill University)
Insurance Premium Prediction via Gradient Tree-Boosted Tweedie Compound Poisson Models / Prédiction d'une prime d'assurance au moyen de modèles de Poisson composés Tweedie avec boosting par arbre et par descente du gradient

16:30-17:00 **Christina Dan Wang** (New York University Shanghai), **Zhao Chen** (Fudan University), **Yimin Lian** (University of Science and Technology of China), **Min Chen** (Academy of Mathematics and Systems Science)
Asset Selection Based on High Frequency Sharpe Ratio / Choix d'actifs en fonction d'un ratio de Sharpe à fréquence élevée

---

15:30-17:00 **Invited / Sur invitation** (abstract/résumé 252) **122 (ICT)**

Recent developments in high-dimensional statistics
Progrès récents en statistique de grande dimension

Chair/Président: Kun Liang
Organizer/Responsable: Yingli Qin

15:30-16:00 **Jun Li** (Kent State University)
Change-Point Detection in High-Dimensional Time Series / Détection de points de changement dans des séries chronologiques de haute dimension

16:00-16:30 **Pingshou Zhong** (University of Illinois At Chicago), **Shawn Santo** (Michigan State University)
Covariance Change Point Detection and Identification with Applications in the Brain's Dynamic Functional Connectivity / Détection et repérage de point de changement de covariance avec applications dans la connectivité cérébrale fonctionnelle et dynamique

16:30-17:00 **Yingli Qin** (University of Waterloo), **Yilei Wu** (University of Waterloo), **Mu Zhu** (University of Waterloo)
Joint Estimation of Multiple High-Dimensional Covariance Matrices / Estimation conjointe de multiples matrices de covariance de haute dimension

15:30-16:35 **Invited / Sur invitation** (abstract/résumé 254) **102 (ICT)**

Biostatistics Section Presidential Address
Allocution de l'invité du Président du Groupe de biostatistique

Chair/Président: Patrick E. Brown
Organizer/Responsable: Patrick E. Brown
Sponsor/Commanditaires: SSC Biostatistics Section / Groupe de biostatistique de la SSC

15:30-16:35 **John Martin Bland** (University of York)
Improving Statistical Quality in Published Research : The Clinical Experience / Améliorer la qualité statistique des recherches publiées : l'expérience clinique

---

15:30-17:00 **Invited / Sur invitation** (abstract/résumé 255) **146 (SB)**

New Developments in State-space Modeling Approaches for Ecology and Environmental Research
Nouvelles évolutions en méthodes de modèles d'espaces d'états pour l'écologie et la recherche environnementale

Chair/Président: Ying Zhang
Organizer/Responsable: Ying Zhang

15:30-16:00 **Hugh Chipman** (Acadia University), **Khurram Nadeem** (University of Guelph), **Ying Zhang** (Acadia University)
A Hierarchical State-Space Approach for Modeling Population Indices Data / Une approche hiérarchique de l'espace d'état pour la modélisation des données sur les indices de population

16:00-16:30 **Guohua Yan** (University of New Brunswick), **Xingde Duan** (Guizhou University of Finance and Economics), **Xiaolei Zhang** (Yunnan Normal University), **Renjun Ma** (University of New Brunswick), **Ying Zhang** (Acadia University)
A Simultaneous Trajectory Modelling of Weather-Related Natural Disasters in Canada / Modélisation des trajectoires simultanées des catastrophes naturelles liées aux intempéries au Canada

16:30-17:00 **Connie Stewart** (University of New Brunswick Saint John), **Shelley Lang** (Fisheries and Oceans Canada)
Measuring Repeatability in the Diet of Grey Seals (Halichoerus Grypus) / Mesure de la répétabilité dans l'alimentation des phoques gris (Halichoerus grypus)

15:30-17:00 **Contributed / Communications libres** (abstract/résumé 257) **142 (AD)**

Causal Inference: Applications and Case Studies
Inférence causale : applications et études de cas

Chair/Président: Sanjeena Dang

15:30-15:45 **Thai-Son Tang** (University of Toronto), **Keith A. Lawson** (University Health Network), **Antonio Finelli** (University Health Network), **Olli Saarela** (University of Toronto)
Causal Inference Methods for Quality-Of-Care Comparisons Involving Small Institutions / Méthodes d'inférence causale pour la comparaison de la qualité des soins entre petits établissements

15:45-16:00 **Mohammad Ehsanul Karim** (The University of British Columbia), **Menglan Pang** (McGill University), **Robert Platt** (McGill University)
Can We Train Machine Learning Methods to Outperform the High-Dimensional Propensity Score Algorithm? / Pouvons-nous entraîner des méthodes d'apprentissage machine pour surpasser l'algorithme de scores de propension de dimension élevée ?

16:00-16:15 **Shomoita Alam** (McGill University), **Shomoita Alam** (,) **Erica Moodie** (McGill University), **David Stephens** (McGill University)
Should a Propensity Score Model be Super? The Utility of Ensemble Procedures for Causal Adjustment / Un modèle de scores de propension doit-il être « Super » ? Utilité des procédures de prévision d'ensemble pour l'ajustement causal

16:15-16:30 **Sudipta Saha** (University of Toronto), **Olli Saarela** (University of Toronto), **Amy Liu** (Princess Margaret Cancer Centre)
A Causal Model for Simulating Subgroup Effects in Randomized Screening Trials / Un modèle de causalité pour simuler les effets des sous-groupes dans des essais de dépistage randomisés Ⓔ E

16:30-16:45 **Yasin Khadem Charvadeh** (Memorial University of Newfoundland), **Candemir Cigsar** (Memorial University of Newfoundland)
The Use of Propensity Score Matching Methods for Estimating Treatment Effects in Recurrent Events / Utilisation des méthodes d'appariement des scores de propension pour estimer l'effet du traitement lors d'événements récurrents Ⓔ E

16:45-17:00 **Steve Ferreira Guerra** (McGill University), **Michal Abrahamowicz** (McGill University), **Robert Platt** (McGill University)
A Novel Bootstrap Algorithm for Estimating the Variance of Longitudinal Propensity Score Matching Estimators / Un nouvel algorithme bootstrap pour estimer la variance du score de propension longitudinal des estimateurs d'appariement Ⓔ E

15:30-17:00 **Contributed / Communications libres** (abstract/résumé 261) **119 (SA)**

Modeling Time-to-Event Data
Modélisation de données de durées de vie

Chair/Président: Gyanendra Pokharel

15:30-15:45 **Xiaoming Lu** (Memorial University of Newfoundland), **Zhaozhi Fan** (Memorial University of Newfoundland)
A Joint Model of Longitudinal Quantiles and Multiple-Censored Survival Data / Un modèle conjoint de quantiles longitudinaux et des données de survie multiples et censurées Ⓔ E

15:45-16:00 **Shahedul A. Khan** (University of Saskatchewan)
A Flexible Proportional Hazards Model for Joint Analysis of Longitudinal and Time-To-Event Data / Un modèle de risques proportionnels flexible pour analyse conjointe de données longitudinales et de temps jusqu'à événement Ⓔ E

16:00-16:15 **Tingxuan Wu** (University of Saskatchewan)
Randomized Survival Probability Residual for Assessing Parametric Survival Models / Résidu de probabilité de survie randomisé pour l'évaluation des modèles de survie paramétriques Ⓔ E

16:15-16:30 **Rebecca A. Clark** (University of Alberta), **Yan Yuan** (University of Alberta)
Evaluating Model Accuracy under Sampling Frame for Time-To-Event Data / Évaluation de la précision d'un modèle dans un cadre d'échantillonnage de données de temps d'événement Ⓔ E

16:30-16:45 **Fahmida Yeasmin** (University of Calgary), **Alexander de Leon** (University of Calgary), **Hua Shen** (University of Calgary)
Conditional Dependence in Joint Modelling of Time-To-Event and Longitudinal Outcomes / Dépendance conditionnelle appliquée à la modélisation conjointe des résultats de temps d'événements et longitudinaux Ⓔ E

16:45-17:00 **Mingchen Ren** (University of Calgary), **Ying Yan** (Sun Yat-sen University), **Alexander de Leon** (University of Calgary)
Causal Mediation Analysis of a Survival Outcome with Multiple Mediators Subject to Measurement Error / Analyse de la médiation causale d'un résultat de survie avec plusieurs médiateurs faisant l'objet d'une erreur de mesure Ⓔ E

15:30-17:00 **Contributed / Communications libres** (abstract/résumé 264) **113 (SS)**

Innovations in Statistical and Data Science Education
Innovations en enseignement de la statistique et de la science des données

Chair/Président: Joel A. Dubin

15:30-15:45 **Sohee Kang** (University of Toronto Scarborough), **Sotirios Damouras** (University of Toronto Scarborough)
Effective Online Tool for Mathematics Communication / Outil en ligne efficace pour la communication mathématique

15:45-16:00 **Bethany J.G. White** (University of Toronto), **Lilin Tong** (University of Toronto), **Ming Zhao** (University of Toronto)
Exploring Students' Readiness to Engage with Statistics in Life Science Research / Examen de l'intention des étudiants à se lancer en statistique dans les sciences de la vie

16:00-16:15 **Nicholas Mitsakakis** (University of Toronto)
Teaching Machine Learning in the Health Sciences: Learning Experiences from Developing a New Graduate Course / Enseigner l'apprentissage machine en sciences de la santé : des expériences d'apprentissage tirées de l'élaboration d'un nouveau cours de cycle supérieur

16:15-16:30 **Tharshanna Nadarajah** (St. Francis Xavier University), **Asokan Variyath** (Memorial University of Newfoundland)
A Systematic Approach for Effective Assignment Problem Solving / Approche systématique pour une résolution efficace d'un problème d'affectation

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

Chair/Président: Robert Platt

Organizer/Responsable: Robert Platt

Room/Salle: 148 (ST)

Abstract/Résumé

[Monday May 27/lundi 27 mai, 08:30-09:50]

Sylvia Richardson (University of Cambridge)

Statistical Challenges in the Analysis of Complex Phenotypes in Biomedicine

Défis statistiques relatifs à l'analyse de phénotypes complexes en biomédecine

To better exploit the structure of rich sets of phenotypes, such as clinical biomarkers or genomic profiles, that are currently measured on large samples of healthy or diseased individuals, statistical models of the variations within and the interplay between different layers of genomic structures can be constructed. Generic Bayesian model building strategies and algorithms have been tailored for this purpose. In this talk, I will discuss four areas: refining genetic associations using multiple regression models coupled with variable selection; implementing joint hierarchical modelling of a large number of phenotypes and a large number of features to discover key genetic drivers; analysing tree structured ontology data for finding the underlying genetic origin of rare diseases; and characterizing network structures using fast Bayesian inference in large Gaussian graphical models. Modeling strategies and computations will be illustrated on case studies.

Afin de mieux exploiter la structure des riches ensembles de phénotypes (p. ex., des biomarqueurs cliniques ou des profils génomiques) qui sont actuellement mesurés à partir de grands échantillons d'individus malades ou en santé, il est possible de concevoir des modèles statistiques des variations internes et de l'action réciproque entre différentes couches des structures génomiques. À cette fin, des stratégies de conception de modèle bayésien générique et des algorithmes ont été conçues. Lors de cet exposé, j'examinerai quatre sujets : le raffinement d'associations génétiques grâce à des modèles de régression multiple couplés à une sélection de variable ; la mise en œuvre de la modélisation hiérarchique conjointe d'un grand nombre de phénotypes et de caractéristiques pour découvrir les forceurs génétiques clés ; l'analyse des données ontologiques structurées en arborescence pour trouver l'origine génétique sous-jacente des maladies rares ; et la caractérisation des structures de réseau au moyen d'une inférence bayésienne rapide dans de grands modèles graphiques gaussiens. Nous illustrerons les stratégies de modélisation et les calculs à partir d'études de cas.

Capital Allocation
Affectation de capitaux

Chair/Président: Hélène Cossette

Organizer/Responsable: Hélène Cossette

Room/Salle: 144 (SB)

Abstract/Résumé

[Monday May 27/lundi 27 mai, 10:20-10:50]

David Saunders (University of Waterloo) , **Dan Rosen** (d1g1t Inc.)

Conditional Simulation of Risk Factor Distributions for Stress Testing and Capital Allocation

Simulation conditionnelle des distributions des facteurs de risque pour les simulations de crise et l'imputation sur les fonds propres

We present a simple approach to create meaningful stress scenarios for risk management and investment analysis of multi-asset portfolios, which effectively combines economic forecasts and expert views with portfolio simulation methods. Examples from applications to portfolio credit risk management are presented.

Nous présentons une approche simple dans le but de créer des scénarios de crise pertinents dans le cadre de la gestion des risques et de l'analyse des investissements de portefeuille multisupport, et qui combinent efficacement les prévisions économiques et l'opinion d'experts aux méthodes de simulation de portefeuilles. Nous présenterons aussi des exemples tirés d'applications en gestion des risques de crédit de portefeuilles.

[Monday May 27/lundi 27 mai, 10:50-11:20]

Edward Furman (York University) , **Alexey Kuznetsov** (York University) , **Justin Miles** (York University)

General Risk Aggregation: Is Gamma the New Normal?

Agrégation générale des risques : Gamma représente-t-il la nouvelle norme ?

Moment matching approximations (MMAs) are arguably the most popular method to approximate the distribution of aggregate risk. The existing MMAs comprise such naïve methods as the normal and shifted-gamma approximations that, respectively, match the first two and three moments. More intricate methods are based on the mixed Erlang distributions. However, in practice the sums of risks can have numerous and just a few summands; in the latter case the normal approximation is very questionable. Also, in practice the distributions of the stand-alone risks can be light-tailed or heavy-tailed. In the latter case moments of higher orders may not exist, and so the approximation based on mixed Erlang distributions is of limited usefulness. I will reveal a refined MMA method that approximates the distributions of interest to any precision, works well for light and heavy-tailed distributions, and is fast irrespective of the number of the involved summands.

L'approximation de moments appariés (AMA) est sans doute la méthode la plus populaire pour approximer la distribution de risque agrégée. Parmi les méthodes AMA, on en compte des naïves comme les approximations normales et Gamma décalés, qui apparient les deux et trois premiers moments respectivement. Des méthodes plus complexes sont basées sur les distributions mixtes d'Erland. Cependant, en pratique, la somme des risques peut contenir beaucoup ou peu d'opérandes ; dans le deuxième cas, l'approximation normale est discutable. De plus, la queue des distributions des risques indépendants peut être légère ou lourde. Dans le deuxième cas, les moments d'ordres supérieurs pourraient ne pas exister, ce qui rend très peu utile l'approximation basée sur les distributions mixtes d'Erland. Je présenterai une méthode AMA raffinée qui approxime à différents niveaux de précisions les distributions d'intérêt, fonctionne avec les distributions à queue légère et lourde et est rapide indépendamment du nombre d'opérandes impliqués.

[Monday May 27/lundi 27 mai, 11:20-11:50]

Mélina Mailhot (Concordia University)

Capital Allocation under New Canadian Regulations for PC Insurers

Allocation de capital basée sur la nouvelle règlementation pour les compagnies d'assurance canadiennes

We are currently in a transition period for Canadian insurance companies' regulation, moving from specific guidelines given by the Canadian Institute of Actuaries and the Office of the Superintendent of Financial Institutions to the new IFRS 17 framework. The changes are more specifically related to the financial reporting and calculations of reserves. In this presentation, we will describe the impact on researchers of the new regulatory framework. We will present specific cases, where using precise dependence models and new risk measures can be useful tools for insurance companies, to help them provide more accurate reserves and risk adjustment amounts. We will also present some drawbacks, which can be addressed with recent research results. Examples of car and catastrophic insurance products will be used to illustrate the results.

Les compagnies d'assurance canadiennes vivent présentement une période de transition vers l'adoption des nouvelles normes IFRS 17, qui étaient auparavant dictées par l'Institut Canadien des Actuaires et le Bureau du Superintendant des Institutions Financières. Les changements sont particulièrement liés aux réserves et aux méthodes de calculs des marges de risque. Dans cette présentation, l'impact du nouveau cadre règlementaire sera expliqué. Des cas spécifiques spécifiques, utilisant des modèles de dépendance précise et des nouvelles mesures de risque seront présentés, afin de fournir des réserves et des ajustements de risque plus précis aux compagnies d'assurance. Certains inconvénients seront expliqués et des pistes de solutions seront partagées. Des exemples de produits d'assurance automobile et de risques catastrophiques seront illustrés.

Applications of Nonstandard Analysis to Probability Theory and Statistics
Applications de l'analyse non standard à la théorie des probabilités et à la statistique

Chair/Président: Daniel M. Roy

Organizer/Responsable: Daniel M. Roy

Room/Salle: 113 (SS)

Abstract/Résumé

[Monday May 27/lundi 27 mai, 10:20-10:50]

Haosui Duanmu (University of Toronto) , **Robert Anderson** (University of California, Berkeley) , **Aaron Smith** (University of Ottawa)

Mixing Times and Hitting Times of Markov Processes via Nonstandard Analysis

Temps de mélange et temps d'atteinte des processus de Markov au moyen d'une analyse non standard

The hitting and mixing times are two fundamental quantities associated with Markov chains. Peres, Sousi and Oliveira showed that the mixing times and worst-case hitting times of reversible Markov chains on finite state spaces are equal up to some universal multiplicative constant. We use tools from nonstandard analysis to extend this result to reversible Markov chains on general state spaces that satisfy the strong Feller property. Finally, we show that this asymptotic equivalence can be used to find bounds on the mixing times of a large class of Markov chains used in MCMC, such as typical Gibbs samplers and Metroplis-Hastings chains, even though they usually do not satisfy the strong Feller property.

Les temps d'atteinte et de mélange sont deux quantités fondamentales associées aux chaînes de Markov. Peres, Sousi et Oliveira ont montré que les temps de mélange et les temps d'atteinte dans les pires des cas de temps d'atteinte des chaînes de Markov réversibles dans les espaces d'états finis sont égaux à une constante multiplicative universelle. Nous utilisons des outils d'analyse non standards pour étendre ce résultat à des chaînes de Markov réversibles sur des espaces d'états généraux qui satisfont la solide propriété de Feller. Enfin, nous montrons qu'on peut utiliser cette équivalence asymptotique pour trouver des limites sur les temps de mélange d'une grande classe de chaînes de Markov utilisées dans la méthode de Monte Carlo par chaîne de Markov, comme les typiques échantillonneurs de Gibbs et les chaînes de Metroplis-Hastings, même si elles ne satisfont généralement pas la solide propriété de Feller.

[Monday May 27/lundi 27 mai, 10:50-11:20]

Peter A Loeb (University of Illinois, Urbana-Champaign)

Nonstandard Analysis and Boundaries

Analyse non standard et limites

Aspects of boundary theory we consider are representing measures for positive harmonic functions, Fatou boundary limit theorems, and compactifications including the Martin boundary. First, we review application of standard probability spaces on nonstandard point sets initiated with a construction of representing measures. Further applications include Anderson's construction of Brownian motion and the Ito Integral as well as Sun's treatment of many independent random variables. Next, we note an improvement for any Fatou type boundary limit result. Then we discuss recent work with Insall and Marciniak showing that a standard compactification is always produced by any equivalence relation on "remote" points in a space's nonstandard ex-

Les aspects de la théorie des limites que nous abordons représentent des mesures pour des fonctions harmoniques positives, les théorèmes de limite de Fatou et les compactifications, y compris la limite de Martin. Tout d'abord, nous examinons l'application d'espaces de probabilité standards sur des ensembles de points non standards amorcée par une construction de mesures représentatives. D'autres applications comprennent la construction du mouvement brownien d'Anderson et l'intégrale d'Ito, ainsi que le traitement de Sun de nombreuses variables aléatoires indépendantes. Ensuite, nous constatons une amélioration pour tous les résultats de limite de type Fatou. Puis, nous discutons de travaux récents avec Insall et Marciniak montrant qu'une compatification standard est toujours produit par des relations d'équivalence sur des points ≪ isolés ≫ dans l'extension non standard d'un espace. De

tension. Moreover, any Hausdorff compactification is constructed in this way. We finally consider what might be the probability theoretic equivalence relation for the Martin boundary.

plus, les compactifications de Hausdorff sont construites de cette manière. Nous examinons enfin à ce que pourrait être la relation d'équivalence de probabilité théorique pour la limite de Martin.

[Monday May 27/lundi 27 mai, 11:20-11:50]

Robert Anderson (University of California at Berkeley) , **Roberto Raimondo** (University of Melbourne)

Hyperfinite Existence Results Imply Convergence Results

Les constructions hyperfinies impliquent des théorèmes de convergence

Every hyperfinite construction of a standard object implies a convergence theorem for the corresponding finite construction. Standard proofs of these convergence results often range from very difficult to virtually incomprehensible. We illustrate this with a nonstandard proof from 2008 of the existence of Walrasian equilibrium in a continuous-time finance model. The existence of equilibrium in the finite case rests on Kakutani's Fixed Point Theorem and a perturbation argument. We establish equilibrium in a hyperfinite random walk model and extract from it an equilibrium in the continuous-time model, the first known existence proof outside of particular special cases. We derive as a corollary the convergence of every aspect of the equilibria of the discrete-time model to the corresponding aspect of the equilibria of the continuous-time model; no tractable standard proof of the convergence result is known.

Toute construction hyperfinie d'un objet standard implique un théorème de convergence pour les constructions finies correspondantes. Les preuves standard de ces résultats de convergence sont souvent très difficiles ou pratiquement incompréhensibles. Nous illustrons cela avec une preuve nonstandard datant de 2008 de l'existence d'un équilibre de Walras dans un modèle de finance en temps continu. L'existence de l'équilibre dans le cas fini repose sur le théorème de Kakutani à point fixe et sur un argument de perturbation. Nous établissons l'équilibre dans un modèle de marche aléatoire hyperfini et en extrayons un équilibre dans le modèle à temps continu, la première preuve d'existence connue en dehors de cas particuliers. Nous déduisons en corollaire la convergence de chaque aspect des équilibres du modèle à temps discret avec l'aspect correspondant des équilibres du modèle à temps continu ; aucune preuve standard du résultat de convergence traitable n'est connue.

Innovations in Data Science for Undergraduates in Canada
Innovation en science des données pour les étudiants de premier cycle au Canada

Chair/Président: Bruce Dunham

Organizer/Responsable: Bruce Dunham

Room/Salle: 102 (ICT)

Abstract/Résumé

[Monday May 27/lundi 27 mai, 10:20-10:50]

Tiffany A Timbers (University of British Columbia)

Teaching an "Introduction to Data Science" - A Discussion of Course Design Intent and Lessons Learned

Enseigner une «introduction aux sciences des données» : une discussion relative à l'élaboration de cours et aux leçons apprises

This past semester the Statistics Department at the University of British Columbia launched an introductory Data Science course targeted at first year students. This new course aimed to use both contemporary pedagogical practices (e.g., flipped classroom, pair-programming) and modern technology platforms (R in Jupyter notebooks, autograding using nbgrader, and a JupyterHub integrated with Canvas for authentication). Here we present and discuss what worked and what didn't from the perspectives of both the instructional team and the students.

Lors du dernier semestre, le département de statistiques de l'Université de la Colombie-Britannique a ouvert un cours d'introduction aux sciences des données pour les étudiants de premier cycle. L'objectif de ce nouveau cours est d'adopter les deux pratiques pédagogiques contemporaines (c.-à-d. la classe inversée et la programmation en binôme) et les plateformes de technologie moderne (R avec Jupyter Notebook, notation automatique à l'aide de Nbgrader, et un portail JupyterHub intégré avec Canvas pour l'authentification). Nous présenterons et examinerons quels facteurs ont été positifs ou négatifs selon l'opinion des enseignants et des étudiants.

[Monday May 27/lundi 27 mai, 10:50-11:20]

Xu (Sunny) Wang (Wilfrid Laurier University)

How to Design an Undergraduate Data Science Program to Meet the Needs of Modern Business and Industry?

Comment concevoir un programme de premier cycle en sciences des données pour répondre aux besoins des entreprises et de l'industrie modernes ?

Data-driven decision making is quickly becoming the standard, creating unprecedented demand for data-oriented professionals across a wide range of industries, organizations, and disciplines. Advances in hardware, software and statistical techniques are giving rise to automated methods to analyze patterns and models for all kinds of data, with applications ranging from scientific discovery to business intelligence and analytics. Given the sheer volume of the data sets in question, it is nearly impossible to extract useful information without some degree of mathematical guidance and programming expertise. As such, there are compelling reasons for an undergraduate program that carefully merges mathematics, statistics, data analytics and computing technologies and methods. In this talk, I will present the rationale and the structure of developing the Honours Bachelor of Science in Data Science Program at Wilfrid

La prise de décision orientée par les données devient rapidement la norme, ce qui crée une demande sans précédent de professionnels axés sur les données dans un large éventail de secteurs, d'organisations et de disciplines. L'évolution du matériel informatique, des logiciels et des techniques statistiques donnent lieu à des méthodes automatisées d'analyse des tendances et des modèles pour toutes sortes de données, avec des applications allant de la découverte scientifique à la veille économique et à l'analyse des données en entreprise. Étant donné le volume considérable des ensembles de données en question, il est presque impossible d'extraire de l'information utile sans un certain degré d'orientation mathématique et d'expertise en programmation. Ainsi, il y a de bonnes raisons de créer un programme de premier cycle qui fusionne avec précaution les mathématiques, les statistiques, l'analyse des données et les technologies et méthodes informatiques. Dans cet exposé, je présenterai la raison d'être et la structure de l'élaboration du programme de baccalauréat ès sciences avec

Laurier University.

spécialisation en sciences des données à l'université Wilfrid Laurier.

[Monday May 27/lundi 27 mai, 11:20-11:50]

Jim Stallard (University of Calgary)

Reflections in Building a Data Science Program

Réflexions sur l'élaboration d'un programme en science des données

Data Science continues to emerge as a predominant topic of discussion at many professional conferences. Various institutions across North America are racing to create Data Science undergraduate programs. These initiatives raise questions about undergraduate programs in statistics, and can be a potential spring board for faculty/department divisiveness as claims to the 'Data Science space' are made and competition for limited provincial post-secondary funding to support such initiatives ensues. This presentation will focus on the process of the creation of the undergraduate (minor) program in Data Science, and emphasize the need for (i) collaboration between computer science departments and statistics/mathematics and statistics departments and (ii) consultation with potential stakeholders in faculties across the university.

La science des données poursuit son émergence comme sujet de conversation prédominent dans plusieurs conférences professionnelles. Différentes institutions à travers l'Amérique du nord se pressent à créer des programmes en science des données au baccalauréat. Ces initiatives ont soulevé plusieurs questions à propos des programmes de statistique au baccalauréat. Ces dernières peuvent être source de discorde dans les facultés/départements alors que des revendications pour « l'espace de la science des données » sont faites et qu'une compétition pour l'obtention d'un financement provincial post-secondaire en support à de telles initiatives en résulte. Cet exposé sera axé sur le processus de création du programme au baccalauréat (mineure) en science des données et souligne le besoin de (i) collaboration entre les départements d'informatique et les départements de statistique/mathématiques et de statistique et (ii) consultation avec les partenaires potentiels dans les facultés à travers l'université.

Novel Statistical Methods and Applications in Genomics
Nouvelles méthodes statistiques et applications à la génomique

Chair/Président: Mireille E. Schnitzer

Organizer/Responsable: Mireille E. Schnitzer

Room/Salle: 201 (ENA)

Abstract/Résumé

[Monday May 27/lundi 27 mai, 10:20-10:40]

Jinko Graham (Simon Fraser University)

Relatedness, Inherited Traits and Trait-Influencing DNA Variants

Lien de parenté, caractères héréditaires et variants d'ADN influençant les caractères

A central idea of genetics is that similarity in relationships is reflected by similarity in inherited traits. Related individuals share segments of their genome, derived from a DNA segment in a common ancestor. Whether in pedigrees or populations, the DNA sharing arising from this relatedness underlies heritable traits. The relationships of the DNA segments cannot be observed directly but are informed by the genetic-marker data on individuals. The inferred relationships together with the traits provide information about the genomic location of trait-influencing variants. We will provide an overview of our work and present recent progress in exploring these ideas with simulated and real data.

Une idée centrale de la génétique est que la similitude dans les relations se reflète par la similitude dans les caractères héréditaires. Des individus apparentés partagent des segments de leur génome, obtenus à partir d'un segment d'ADN d'un ancêtre commun. Que ce soit dans les généalogies ou les populations, le partage de l'ADN découlant de cette parenté est à la base des caractères héréditaires. Les relations entre les segments d'ADN ne peuvent pas être observées directement, mais elles s'appuient sur les données des marqueurs génétiques des individus. Les relations déduites ainsi que les caractères fournissent de l'information sur l'emplacement génomique des variants influençant les caractères. Nous donnerons un aperçu de notre travail et présenterons les progrès récents dans l'exploration de ces idées à l'aide de données simulées et réelles.

[Monday May 27/lundi 27 mai, 10:40-11:00]

Marie-Pierre Sylvestre (Université de Montréal) , **Angelo Canty** (McMaster University) , **Shelley Bull** (University of Toronto) , **Paterson Andrew** (University of Toronto) , **Laurence Boulanger** (CRCHUM)

Methods for Genomewide Analysis of Complex Phenotypes

Méthodes d'analyse de phénotypes complexes à l'échelle du génome

I address the methodological challenges associated with conducting genomewide association studies (GWAS) when the phenotype of interest is a nonlinear curve comprising several observations per individual. I compare two approaches that go beyond random effects models and avoid resampling techniques that can become too computationally intensive in GWAS settings. First, I consider semiparametric regression that uses low-rank penalized splines to capture nonlinear changes in the phenotype and with a mixed model representation of the splines that simplifies the estimation. Next, I consider functional principal component analysis (fPCA), a data reduction technique that identifies the principal directions of variations between individual curves. fPCA summarizes the subject-specific features of the curves

J'aborde les défis méthodologiques associés aux études d'association à l'échelle du génome (GWAS) lorsque le phénotype d'intérêt est une courbe non-linéaire comprenant plusieurs observations par individu. Je compare deux approches qui vont au-delà des modèles à effets aléatoires et évitent les techniques de ré-échantillonnage qui peuvent requérir trop de calculs pour un GWAS. Je considérerai d'abord la régression semi-paramétrique qui utilise des splines pénalisées de bas ordre afin de capturer les changements non-linéaires des phénotypes et avec une représentation en modèle mixte des splines qui simplifie l'estimation. Ensuite, je considérerai l'analyse fonctionnelle en composantes principales (AfCP), une technique de réduction des données qui identifie les principales sources de variation entre les courbes individuelles. La AfCP résume les caractéristiques des courbes spécifiques au sujet en les exprimant en composantes principales

into principal component scores that can then be used as phenotypes. I illustrate the methods using data on cholesterol profiles from the Diabetes Control and Complications Trial.

qui peuvent ensuite être utilisés comme phénotypes. J'illustrerai ces méthodes sur les données de profiles de cholestérol de l'étude Diabetes Control and Complications Trial (DCCT).

[Monday May 27/lundi 27 mai, 11:00-11:20]

Jingjing Wu (University of Calgary) , **Tasnima Abedin** (Alberta Health Services)

A Mixture Model under Stochastic Dominance Constraint for Genetic Studies

Modèle de mélange avec contrainte de dominance stochastique pour études génétiques

In this research, we studied a two-component nonparametric mixture model with stochastic dominance constraint, a model that arises naturally from genetic studies. Our interest lies in both (1) the estimation of the mixing proportion and (2) classification. For this model, we proposed and studied a nonparametric estimation based on cumulative distribution functions and a maximum likelihood estimator (MLE) through multinomial approximation. In order to incorporate nicely the stochastic dominance constraint, we introduced a semiparametric model for which we proposed and investigated both MLE and minimum Hellinger distance estimation (MHDE). We also proposed a hypothesis testing approach to assess the validity of the semiparametric model. For the proposed methods, we investigated both their asymptotic properties, including consistency and asymptotic normality, and their finite-sample performance through simulation studies and real data analysis.

Dans ces recherches, nous avons étudié un modèle de mélange non paramétrique à deux composantes avec contrainte de dominance stochastique, un modèle qui s'adapte naturellement aux études génétiques. Nous nous intéressons à l'estimation de la proportion du mélange et à la classification. Pour ce modèle, nous proposons et étudions une estimation non paramétrique fondée sur des fonctions de distribution cumulative et une EMV par approximation multinomiale. Afin de bien intégrer la contrainte de dominance stochastique, nous introduisons un modèle semiparamétrique pour lequel nous proposons et étudions l'EMV et l'estimation de la distance de Hellinger minimum (EDHM). Nous proposons également un test d'hypothèse pour tester la validité du modèle semiparamétrique. Nous étudions enfin les propriétés asymptotiques (convergence et normalité asymptotique) des méthodes proposées et leur performance sur échantillon fini via des études de simulation et une analyse de données réelles.

[Monday May 27/lundi 27 mai, 11:20-11:40]

Ting-Huei Chen (Laval University) , **Hanaa Boughal** (Laval University)

A Powerful Approach to Identify the Genetic Variables Associated with Psychiatric Diseases

Une approche efficace pour l'identification de variables génétiques associées aux maladies psychiatriques

Psychiatric diseases are complex and highly heterogeneous conditions. Due to its complexity, the qualitative diagnosis often suffers high rates of misdiagnosis and oversimplifies disease phenotypes. However, the qualitative diagnosis is often used as a response variable in genetic association studies for psychiatric diseases. Endophenotypes are the quantitative traits hypothesized to underlie disease syndromes; they are believed to better capture the human behavioral abnormalities than the imprecise categorical psychiatric diagnoses. Since case-control studies often collect information on secondary phenotypes, the appropriate use of endophenotypes may enhance the statistical power to identify the disease associated genetic variables. We propose a novel method to address this challenge. Simulation results demon-

Les maladies psychiatriques sont complexes et hautement hétérogènes. En raison de sa complexité, le diagnostic qualitatif présente souvent un taux élevé de mauvais diagnostics et simplifie trop les phénotypes de la maladie. Par contre, le diagnostic qualitatif est fréquemment utilisé en tant que variable de réponse dans les études d'association génétiques pour les maladies psychiatriques. Les endophénotypes sont les caractéristiques quantitatives qui, par hypothèse, sont à la base des syndromes de maladie ; ils sont considérés comme étant meilleurs pour saisir les anomalies comportementales humaines que les diagnostics psychiatriques catégoriques imprécis. Étant donné que les études cas-témoin recueillent souvent de l'information sur les phénotypes secondaires, l'utilisation appropriée d'endophénotypes peut améliorer la puissance statistique de l'identification des variables génétiques associées aux maladies. Nous proposons une nouvelle méthode pour

strate good performance of the method. The real data analysis of the Alzheimer's disease illustrates the practical utility of the techniques.

traiter ce problème. Des résultats de simulation démontrent une bonne performance de la méthode. L'analyse de données réelles sur la maladie d'Alzheimer illustre l'utilité pratique de ces techniques.

Recent Developments in Survival Analysis with Complex Data
Récentes évolutions en analyse de survie avec données complexes

Chair/Président: Xuewen Lu

Organizer/Responsable: Xuewen Lu

Room/Salle: 122 (ICT)

Abstract/Résumé

[Monday May 27/lundi 27 mai, 10:20-10:50]

Menggang Yu (University of Wisconsin)

Cox Regression with Nonignorable Survival-Time-Dependent Missing Covariate Values

Régression de Cox avec des valeurs de covariables manquantes non ignorables dépendantes du temps de survie

When analyzing time-to-event data in clinical and epidemiological studies with missing covariate values, the missing at random assumption is commonly adopted. It assumes that missingness depends on the observed data, including the observed outcome which is the minimum of survival and censoring time. However, in certain settings, the missingness is likely related to the survival time but not to the censoring time. This occurs for example when covariates are measured at baseline and censoring is administrative. In this case, the covariate missingness mechanism is nonignorable as the survival time is censored, and it creates challenge in data analysis. We propose two different estimators to deal with such survival-time-dependent covariate missingness based on the well known Cox regression models. Our method is based on inverse propensity weighting with the propensity estimated by nonparametric kernel regression.

Lors de l'analyse des données relatives au délai avant l'événement dans les études cliniques et épidémiologiques avec des covariables manquantes, on adopte en général l'hypothèse des données manquantes aléatoirement. Elle suppose que le mécanisme responsable des données manquantes dépend des données observées, ainsi que des résultats observés qui représentent le minimum de temps de survie et de censure. Cependant, dans certains contextes, le mécanisme en question est probablement lié au temps de survie, mais pas au temps de censure. C'est par exemple ce qui se produit lorsque les covariables sont mesurées au départ et que la censure est administrative. Dans ce cas, le mécanisme de covariables manquantes ne peut être ignoré, car le temps de survie est censuré, ce qui pose un problème dans l'analyse des données. Nous proposons deux estimateurs différents pour traiter ce manque de covariables dépendantes du temps de survie selon les modèles de régression de Cox bien connus. Notre méthode est fondée sur la pondération inverse de la propension avec la propension estimée par la régression non paramétrique du noyau.

[Monday May 27/lundi 27 mai, 10:50-11:20]

Gang Li (University of California, Los Angeles) , **Eric Kawaguchi** (University of California Los Angeles) , **Marc Suchard** (University of California Los Angeles) , **Zhenqiu Liu** (Penn State University)

Scalable Sparse Cox's Regression for Large-Scale Survival Data via Broken Adaptive Ridge

Régression de Cox creuse et échelonable pour des données de survie à grande échelle au moyen d'un ridge adaptatif brisé

In this talk I will present a new sparse Cox regression method for high-dimensional massive sample size survival data. Our method is an L0-based iteratively reweighted L2-penalized Cox regression model, which inherits some appealing properties of both L0 and L2 penalized Cox regression while overcoming their limitations. We establish that it has an oracle property for selection and estimation and a grouping property for highly correlated covariates. We develop an efficient implementation for high-dimensional massive sample size survival data, which exhibits substantial speedups

Ma présentation porte sur une nouvelle méthode de régression de Cox creuse pour des données de survie d'un échantillon de données volumineuses de grande dimension. Notre méthode est une régression de Cox pénalisée par la norme L2 itérativement repondérée selon la norme L0, méthode dotée de certaines propriétés attrayantes à la fois de la régression de Cox pénalisée par les normes L0 et L2, tout en surmontant leurs limites. Nous lui reconnaissons une propriété oracle pour le choix et l'estimation ainsi qu'un propriété de regroupement pour des covariables fortement correlées. Nous développons une implémentation efficace pour des données de survie d'un échantillon de données volumineuses

over its competitor in numerical studies. The performance of our method is illustrated using simulations and real data examples.

de grande dimension qui montre d'importantes accélérations par rapport à ses concurrents dans des études numériques. Le fonctionnement de notre méthode est illustré à l'aide de simulations et d'exemples avec des données réelles.

[Monday May 27/lundi 27 mai, 11:20-11:50]

Zhigang Li (University of Florida) , **Janaka Peragaswaththe Liyanage** (University of Florida) , **Lihui Zhao** (Northwestern University)

Joint Modeling of Survival and Longitudinal Quality of Life Data with Informative Censoring in Palliative Care Studies

Modélisation conjointe des données de survie et des données longitudinales sur la qualité de vie avec censure informative dans les études sur les soins palliatifs

Palliative medicine is an interdisciplinary specialty focusing on improving quality of life (QOL) for patients with serious illness and their families. Palliative care programs are widely available or under development at US hospitals. In palliative care studies, longitudinal QOL and survival data are often highly correlated which, in the face of censoring, makes it challenging to properly analyze and interpret terminal QOL trends. Informative dropout in the study adds another level of complication of the problem. To address these issues, we propose a novel statistical approach to jointly model the terminal trend of QOL and survival data, accounting for informative dropout. We assess the model through simulation and application to establish a novel modeling approach that could be applied in future palliative care trials.

La médecine palliative est une spécialité interdisciplinaire axée sur l'amélioration de la qualité de vie des patients atteints de maladies graves et de leur famille. Les programmes de soins palliatifs sont largement disponibles ou en cours d'élaboration dans les hôpitaux américains. Dans le cadre des études sur les soins palliatifs, les données longitudinales sur la qualité de vie et les données de survie sont souvent fortement corrélées, ce qui, face à la censure, rend difficiles l'analyse et l'interprétation adéquates des tendances de la qualité de vie terminale. La perte d'information dans l'étude ajoute un autre niveau de complication au problème. Pour résoudre ces problèmes, nous proposons une nouvelle approche statistique permettant de modéliser conjointement la tendance terminale de la qualité de vie et les données sur la survie, en tenant compte de la perte d'information. Nous analysons le modèle par une simulation et une application afin d'établir une nouvelle approche de modélisation qui pourrait être appliquée dans de futurs essais en soins palliatifs.

Chair/Président: Quan Long

Room/Salle: 142 (AD)

Abstract/Résumé

[Monday May 27/lundi 27 mai, 10:20-10:35]

John R.J. Thompson (University of Western Ontario)

Estimating Fire Spread Rates from Micro-Fire Experiments

Estimation des taux de propagation du feu à partir d'expériences de micro-feux

Wildfires in North America were a hot topic in 2018, largely due to fires in California and British Columbia. Each wildfire is unique, with environmental conditions that vary within and between fires, and so is the uncertainty of fire spread rates. Research is presented on the statistical methods used to estimate fire spread rate and spread rate variability from satellite-perspective imagery data. The design of a fire smoldering apparatus is presented and is followed by results obtained in micro-fires experiments under controlled environmental conditions. The progression of fire is classified by three major areas: fuel, burning, and burnt-out. Anisotropic smoothing is applied to remove noise in images while preserving the boundaries between areas, then a data sharpening algorithm is used to estimate the state of each pixel over time. The estimation of fire spread rates is validated initially with simulated data, and then by comparison with measurements of the micro-fires.

Les feux incontrôlés en Amérique du Nord ont été un sujet brûlant en 2018, principalement en raison des incendies en Californie et en Colombie-Britannique. Chaque feu incontrôlé est unique, dans des conditions environnementales qui varient d'un feu à l'autre et à l'intérieur d'un même feu, tout comme l'incertitude des taux de propagation du feu. On présente la recherche sur les méthodes statistiques utilisées pour estimer la variabilité de la vitesse de propagation du feu et de la vitesse de propagation à partir des données d'imagerie satellitaire. On présente le modèle d'un appareil d'extinction du feu couvant, ainsi que les résultats obtenus lors d'expériences de micro-feux dans des conditions environnementales contrôlées. La progression du feu est classée selon trois grands domaines : le combustible, la combustion et la carbonisation. Le lissage anisotrope est appliqué pour supprimer le bruit dans les images tout en préservant les limites entre les zones. Ensuite, on utilise un algorithme d'affinage des données pour estimer l'état de chaque pixel dans le temps. On valide d'abord l'estimation des taux de propagation du feu par des données simulées, puis par comparaison avec les mesures des micro-feux.

[Monday May 27/lundi 27 mai, 10:35-10:50]

Nan Zheng (Marine Institute of Memorial University of Newfoundland) , **Noel Cadigan** (Centre for Fisheries Ecosystems Research, Fisheries and Marine Institute of Memorial University of Newfoundland) , **Joanne Morgan** (Fisheries and Oceans Canada)

A Spatiotemporal Von Bertalanffy Growth Model and Its Estimation When Data Are Collected Through Length-Stratified Sampling

Un modèle spatiotemporel de croissance de Von Bertalanffy et son estimation lorsque les données sont recueillies par échantillonnage stratifié dans la longueur

We develop a conditional empirical proportion likelihood approach for data collected with the response-selective stratified sampling design under the condition that no appropriate covariate distribution model is available. Neglecting this sampling scheme can lead to seriously biased estimation results. We propose a mixed effects Von Bertalanffy growth model incorporating spatial and temporal variation and correlation, between-individual variation, as well as length and age measure-

Nous élaborons une approche de la vraisemblance de la proportion conditionnelle et empirique pour les données recueillies au moyen du plan d'échantillonnage stratifié sélectif en fonction de la réponse, à la condition qu'il n'y ait aucun modèle de distribution de covariables approprié. Le fait de ne pas tenir compte de ce plan d'échantillonnage peut conduire à des résultats d'estimation très biaisés. Nous proposons un modèle de croissance de Von Bertalanffy à effets mixtes en intégrant la variation et la corrélation spatiale et temporelle, la variation entre les individus, ainsi que les

ment errors. The effect of maturation on growth is also incorporated explicitly. The bias in the age-length data due to size-selective capture of the survey gear is accounted for by a logistic selectivity model. The model and the methodology are applied to data for American plaice in NAFO Divisions 3LNO. The estimates of the distribution of size-at-age are an important input to a spatial stock assessment model with length-based catchability being developed for this stock.

erreurs de mesure de la longueur et de l'âge. L'effet de la maturation sur la croissance est également intégré explicitement. Le biais dans les données sur l'âge et la longueur en raison de la capture sélective de la taille du matériel d'enquête s'explique par un modèle de sélectivité logistique. Le modèle et la méthodologie sont appliqués aux données sur la plie canadienne dans la division 3LNO de l'Organisation des pêches de l'Atlantique Nord-Ouest (OPANO). Les estimations de la distribution de la taille en fonction de l'âge constituent une donnée importante pour un modèle d'évaluation spatiale des stocks, dont la capturabilité basée sur la longueur est en cours d'élaboration.

[Monday May 27/lundi 27 mai, 10:50-11:05]

Rolf Turner (University of Auckland) , **Kate Richards** (New Zealand Plant and Food Reseach Institute)

Adventures with Bark Beetles

Aventures avec les scolytes

I shall describe some of the problems encountered in fitting dose-response models to insect survival data obtained by researchers at the New Zealand Plant and Food Research Institute. The insects in question are bark beetles (e.g. the black pine bark beetle, Hylastes ater), which may infest shipments of logs exported from New Zealand. The fundamental practical problem is to determine a dose level for the fumigant used, e.g. methyl bromide (MeBr) or ethanedinitrile (EDN), which is as small as possible but still sufficient to kill essentially all of the insects. "As small as possible" is important since the fumigants are environmentally hazardous. Our approach is to fit generalised linear (mixed) models. We determine, using Fieller's formula, the upper endpoint of 95% confidence interval for the dose that is lethal with probability 0.999968 ("probit 9 level efficacy").

Je décrirai certains des problèmes rencontrés lors de l'ajustement des modèles dose-réponse aux données sur la survie des insectes obtenues par les chercheurs du New Zealand Plant and Food Research Institute. Les insectes en question sont les scolytes (p. ex., l'hylésine noir du pin, Hylastes ater), qui peuvent infester les envois de grumes exportés de Nouvelle-Zélande. Concrètement, le problème fondamental consiste à déterminer une dose de fumigant utilisée, comme le bromure de méthyle (MeBr) ou l'éthanedinitrile (EDN), qui est aussi faible que possible, mais suffisante pour tuer pratiquement tous les insectes. La notion d'« aussi faible que possible » est importante, car les fumigants sont dangereux pour l'environnement. Notre approche consiste à adapter des modèles (de mélange) linéaires généralisés. Nous déterminons, à l'aide de la formule de Fieller, le critère d'effet final supérieur de l'intervalle de confiance à 95 % pour la dose létale avec une probabilité de 0,999968 (« efficacité au niveau probit 9 »).

[Monday May 27/lundi 27 mai, 11:05-11:20]

Matthew R. Parker (University of Victoria) , **Vivian Pattison** (University of Victoria) , **Laura L.E. Cowen** (University of Victoria)

Estimating Population Abundance Using Counts from an Auxiliary Population and N-Mixture Models

Estimation de l'abondance d'une population à l'aide de nombres d'une population auxiliaire et de modèles de mélange N

We provide a method for estimating the adult breeding population of an Ancient Murrelet seabird colony using chicks as an auxiliary population. Adults spend much of their time at sea and are difficult to count, while chicks can be counted soon after hatching, as they leave the colony. Count data for 6 sites and 17 sampling occasions provide the framework to fit N-mixture models and to obtain auxiliary estimates. Ancient Murrelet clutch size is used to convert from auxiliary to breeding pair estimates. The breeding pair estimates are then extrapolated

Nous proposons une méthode qui permet d'estimer la population adulte reproductrice d'une colonie de guillemots à cou blanc en utilisant les oisillons comme population auxiliaire. Les adultes passent beaucoup de temps en mer et sont donc difficiles à compter, tandis que les oisillons peuvent être dénombrés peu de temps après l'éclosion, alors qu'ils quittent la colonie. Des données de dénombrement de 6 sites et 17 cycles d'échantillonnage fournissent le cadre d'ajustement de modèles de mélange N afin d'obtenir des estimations auxiliaires. Nous utilisons la taille de la nichée des guillemots à cou blanc pour convertir de l'estimation

to the total colony using an area expansion. The population is seen to decrease over the period 1995 to 2006, from 901 to 761. We compare our estimates to the Canadian Wildlife Service (CWS) survey estimates of 1995 and 2006, indicating a smaller abundance than the CWS estimates for the year 1995, and larger for the year 2006. These methods can be applied to other difficult to count populations having links to auxiliary populations.

auxiliaire à l'estimation des couples reproducteurs. Ces dernières estimations sont ensuite extrapolées à la colonie totale par une expansion de la zone. On note que la population diminue de 901 à 761 sur la période 1995-2006. Nous comparons nos estimations aux estimations d'enquête du Service canadien de la faune (SCF) pour 1995 et 2006 : nos chiffres sont inférieurs à ceux de la SCF pour 1995 et plus élevés pour 2006. Ces méthodes peuvent être appliquées à d'autres populations difficiles à compter mais liées à des populations auxiliaires.

[Monday May 27/lundi 27 mai, 11:20-11:35]

Guowen Huang (Centre for Global Health Research) , **Patrick Brown** (St. Michael's Hospital; University of Toronto)

Daily Mortality and Air Quality: Using Multivariate Time Series with Seasonally-Varying Covariances

Mortalité quotidienne et qualité de l'air : utilisation d'une série temporelle multivariée avec des covariances variables selon les saisons

We studied the association of daily mortality with short-term variations in the ambient concentrations of PM2.5, NO2 and O3 in Vancouver and Toronto. Firstly, a multivariate time series model within a Bayesian framework was proposed for exposure assessment. This was a mixture of Gamma and Half-Cauchy models, with the latter being used to capture the heavy tail of the data distribution. Seasonally varying covariance among pollutants was allowed, and the pollution model was implemented with parallel subthreads to speed computation. Then a case-crossover design and conditional logistic regression disease model was used to relate exposure to mortality data for 1981 to 2012.

Nous avons étudié l'association entre la mortalité quotidienne et les variations à court terme des concentrations ambiantes de particules fines PM2.5, de dioxyde d'azote (NO2) et d'ozone (O3) à Vancouver et Toronto. Un modèle de série temporelle multivariée dans un cadre d'étude bayésien a d'abord été proposé pour l'évaluation de l'exposition. Ce modèle est un mélange entre les modèles Gamma et demi-Cauchy, ce dernier étant utilisé pour capturer la queue lourde de la distribution des données. Une covariance variable selon les saisons parmi les polluants était permise et le modèle de pollution a été implémenté avec des sous-unités d'exécution parallèles pour accélérer le calcul. Un modèle d'étude croisée de cas et de régression logistique conditionnelle pour la maladie a ensuite été utilisé pour lier l'exposition aux données sur la mortalité entre 1981-2012.

[Monday May 27/lundi 27 mai, 11:35-11:50]

Inesh Munaweera (University of Manitoba) , **Saman Muthukumarana** (University of Manitoba) , **Darren Gillis** (University of Manitoba) , **Douglas Watkinson** (Fisheries and Oceans Canada) , **Colin Charles** (Fisheries and Oceans Canada)

Understanding Lake Winnipeg Basin Walleye Fish Movement Patterns Using Bayesian State-Space Models

Analyse des schémas de mouvement des dorés jaune du bassin du lac Winnipeg par modèles espace-état bayésiens

State-space models (SSMs) are frequently used to model dynamic systems which involve hidden or unobservable states. SSMs have been increasingly favored in studying animal movements and population dynamics in ecology since they can account for both process variation and observational error. Our study is based on the dataset consisting of detection records of tagged fish (walleye) which were collected using a grid of acoustic receivers laid in the bottom of Lake Winnipeg under the "Lake Winnipeg Basin Fish Movement Project" which is being conducted by Fisheries and Oceans Canada. We will assess walleye movement patterns by employing broad summaries and individual movement path re-

On utilise souvent des modèles espace-état (MEE) pour modéliser des systèmes dynamiques dont certains états sont cachés ou non-observables. C'est le cas, de plus en plus, pour l'étude des mouvements d'animaux et de la dynamique des populations en écologie, puisqu'ils permettent de tenir compte à la fois des variations de processus et des erreurs d'observation. Notre étude repose sur un jeu de données composé d'enregistrements de détection de poissons (dorés jaunes) marqués, collectés grâce à une grille de récepteurs acoustiques disposés au fond du lac Winnipeg dans le cadre du projet « Lake Winnipeg Basin Fish Movement Project » de Pêches et Océans Canada. Nous évaluons les schémas de mouvement des dorés en employant des synthèses larges et une reconstruction de trajectoires individuelles. Dans cette étude,

construction. In this study, the true fish positions are unobserved; we only have the positions of the acoustic receivers detecting them. Hence, we will use the Bayesian State-space modeling approach to reconstruct the true fish movement model by combining it with the observation model describing fish detections.

les positions réelles des poissons ne sont pas observées : nous ne connaissons que les positions des récepteurs acoustiques qui les détectent. Nous utilisons donc une approche de modélisation espace-état bayésienne pour reconstruire le mouvement réel des poissons en le combinant avec le modèle d'observation qui décrit la détection des poissons.

Chair/Président: Shamsia Sobhan

Room/Salle: 101 (ENA)

Abstract/Résumé

[Monday May 27/lundi 27 mai, 10:20-10:35]

Tristan Watson (ICES) , **Laura Rosella** (University of Toronto) , **Kathy Kornas** (University of Toronto) , **Catherine Bornbaum** (University of Toronto)

Age-Standardizing Prevalence Estimates from Combined Cycles of the Canadian Community Health Survey

Estimations de la prévalence normalisées selon l'âge à partir d'une combinaison de cycles de l'Enquête sur la santé dans les collectivités canadiennes

Survey data from the Canadian Community Health Survey (CCHS) is commonly used by health system organizations to estimate the prevalence of behavioral, health status, and other risk factors, in order to inform health needs and priorities for people in different populations. However, two issues typically present challenges. First, a single cycle of the CCHS may not have a large enough sample size to reliably compute estimates for a small region or population. Second, crude prevalence estimates between two populations might be confounded by the underlying age structure of the two groups. Thomas and Wannell (2009) proposed a 'pooled approach' method for combining CCHS cycles, in order to increase the power and sample size for analysis. We illustrate the application of this pooled approach and demonstrate how to produce age-standardized prevalence estimates and confidence intervals generated with bootstrap weights from pooled CCHS data using procedures in SAS.

Les données de l'Enquête sur la santé dans les collectivités canadiennes (ESCC) sont fréquemment utilisées par les organismes du système de santé pour estimer la prévalence de facteurs de risque comportemental, de statut de santé et autres et ainsi mieux comprendre les besoins et priorités de santé des membres de différentes populations. Cependant, deux questions se posent souvent. Premièrement, un seul cycle de l'ESCC ne produit pas forcément un échantillon assez grand pour calculer avec fiabilité des estimations pour une petite région ou population donnée. Deuxièmement, les estimations de prévalence brutes pour deux populations peuvent être faussées par la structure d'âge sous-jacente des deux groupes. Thomas et Wannell (2009) avaient proposé une « approche de regroupement » pour combiner plusieurs cycles de l'ESCC et ainsi augmenter la puissance d'une analyse et la taille de l'échantillon. Nous illustrons l'application de cette approche de regroupement et montrons comment produire des estimations de la prévalence normalisées selon l'âge et des intervalles de confiance générés avec des poids bootstrap à partir de données regroupées de l'ESCC avec des procédures SAS.

[Monday May 27/lundi 27 mai, 10:35-10:50]

Joshua Gutoskie (Statistics Canada)

Simulating Survey Design Weights to Account for Sample Coordination Between Surveys

Pondérations simulées de conception d'enquête pour prendre en compte la coordination échantillonnale entre les questionnaires

In an effort to reduce respondent burden, sample coordination often takes place to remove overlapping units between surveys. The removed units are often replaced with other eligible units that were not included in the pre-coordinated sample. Calculating the first-order inclusion probabilities is often quite difficult when trying to account for the removal and replacement of units within a sample. Statistics Canada's Farm Management Survey's unique sample design required several stages of sample coordination not only with other agriculture

Dans le but d'alléger le fardeau du répondant, l'échantillonnage est souvent coordonné pour éliminer le chevauchement des unités entre les questionnaires. Les unités éliminées sont souvent remplacées par d'autres admissibles qui ne sont pas comprises dans l'échantillonnage pré-coordination. Le calcul des probabilités d'inclusion de premier ordre se révèle souvent difficile lorsque nous tentons de prendre en compte l'élimination et le remplacement d'unités dans un échantillonnage. La conception unique de l'Enquête sur la gestion des fermes de Statistique Canada a exigé plusieurs étapes de coordination de l'échantillonnage, non seule-

surveys but also within its own sample, which made calculating the first-order inclusion probabilities challenging. In this paper, we use the Monte Carlo simulation-based approach to estimate first-order inclusion probabilities, as proposed by Thompson and Wu (2008), applying its use to account for the sample coordination done for the Farm Management Survey.

ment par rapport à d'autres enquêtes portant sur l'agriculture, mais à l'intérieur même de son propre échantillonnage, compliquant ainsi le calcul des probabilités d'inclusion de premier ordre. Pour estimer ces probabilités, nous faisons appel à une méthode de simulation de Monte-Carlo, comme l'ont proposé Thompson et Wu (2008), en appliquant son utilisation à la prise en compte de la coordination de l'échantillonnage pour l'Enquête sur la gestion des fermes.

[Monday May 27/lundi 27 mai, 10:50-11:05]

Vanessa McNealis (Université de Montréal) , **Christian Léger** (Université de Montréal)

Smoothed Bootstrap Estimator of the Variance of a Quantile Estimator in a Finite Population Context

Estimation de bootstrap lissé de la variance d'un estimateur de quantile dans le contexte d'une population finie

As shown in several simulation studies, existing bootstrap methods for survey data lead to relatively poor confidence intervals and variance estimators in small to moderate samples when the functional of interest is a quantile. In an i.i.d. context, Hall, DiCiccio and Romano (1989) have shown that resampling from a smoothed estimate of the distribution function instead of the usual empirical distribution function can improve the estimator convergence rate in the case of the variance of quantiles. The smoothed bootstrap has yet to be used in survey methodology, although it would be notably feasible in pseudo-population bootstrap methods. Given a kernel function and a bandwidth, it would consist in smoothing the pseudo-population from which bootstrap samples are drawn using the original sampling design. This study aims to assess the performance of the approach for variance estimation and confidence intervals for quantiles while providing the user a guide for selecting the bandwidth.

Comme montré dans plusieurs études de simulation, les méthodes bootstrap existantes pour des données d'enquête mènent à de mauvaises estimations des intervalles de confiance et de la variance dans des échantillons de taille petite ou modérée, lorsque la fonctionalité d'intérêt est un quantile. Dans le cas de données i.i.d., Hall, DiCiccio et Romano (1989) ont montré que la vitesse de convergence de l'estimateur de la variance d'un quantile connaît un gain si l'on rééchantillonne à partir d'une estimation lissée de la fonction de répartition au lieu de la fonction de répartition usuelle. La méthode de bootstrap lissé n'a pas encore été utilisée en méthodologie d'enquête, bien qu'elle serait notamment faisable dans les méthodes de bootstrap par pseudo-population. Avec une fonction de noyau et un paramètre de lissage h, la méthode consiste à lisser une pseudo-population à partir des échantillons bootstrap obtenus selon le plan de sondage initial. Nous évaluerons la performance de cette approche pour l'estimation de la variance et des intervalles de confiance des quantiles, tout en fournissant un guide quant à la sélection du paramètre de lissage.

[Monday May 27/lundi 27 mai, 11:05-11:20]

Yuxiang Gao (University of Toronto) , **Lauren Kennedy** (Columbia University) , **Daniel Simpson** (University of Toronto)

Incorporating Structured Priors into Multilevel Regression and Poststratification

Intégrer des lois a priori structurées dans la régression multi-niveaux et la post-stratification

A central theme in the field of survey statistics is estimating population-level quantities through data coming from potentially non-representative samples of the population. Multilevel Regression and Poststratification (MRP), a model-based approach, is gaining traction against the traditional weighted approach for survey estimates. MRP uses partial pooling through random effects, thus shrinking model estimates to an overall mean and reducing potential overfitting. Despite MRP's straightforward specification of prior distributions, the estimates are susceptible to bias if there is an underlying

Un thème central dans le domaine des sondages statistiques est l'estimation de quantités au niveau de la population par des données provenant d'échantillons potentiellement non-représentatifs de la population. La régression multi-niveaux et la post-stratification (RMP), une approche basée sur le modèle, gagne du terrain face à l'approche pondérée traditionnelle pour l'estimation par sondage. RMP utilise le regroupement partiel par effets aléatoires, rétrécissant ainsi les estimations du modèle à une moyenne générale et réduisant le sur-ajustement potentiel. Malgré la spécification simple des lois a priori de la RMP, les estimations sont exposées aux biais s'il y a une structure sous-jacente que la loi a priori ne

structure that the prior does not capture. This work aims to provide a new framework for specifying structured prior distributions that lead to more robust estimates for MRP. We use simulation studies to explore the benefit of these priors and demonstrate them on US survey data.

saisit pas. Cet exposé présente un nouveau cadre pour définir la structure des lois a priori qui entraine des estimations plus robustes pour RMP. Nous utilisons des études de simulation pour explorer les avantages de ces lois a priori et nous les démontrons avec des données d'un sondage américain.

[Monday May 27/lundi 27 mai, 11:20-11:35]

Yilin Chen (University of Waterloo) , **Changbao Wu** (University of Waterloo) , **Pengfei Li** (University of Waterloo)

Estimation of Population Proportions with Non-Probability Survey Samples

Estimation des proportions de la population à l'aide d'échantillons d'enquêtes non probabilistes

In survey questionnaires, binary responses such as yes/no, agree/disagree, satisfied/not satisfied, are common. Collected binary data are then used to estimate proportions of the population with certain characteristics. In this project, we propose to estimate population proportions with samples from non-probability based surveys. We propose a pseudo-empirical likelihood (PEL) inferential procedure, and show that resulting point estimator for a population proportion has desirable doubly robust property. Two methods of constructing PEL ratio confidence intervals for the population proportion are proposed, one is based on the limiting distribution of adjusted PEL ratio statistic and the other uses the bootstrap calibrated PEL. A simulation study shows that, when sample size is small, PEL ratio based confidence intervals have better coverage rate and more balanced tail error rate than the commonly-used Wald-type confidence intervals.

Dans les questionnaires d'enquête, les réponses binaires comme oui-non, d'accord-en désaccord, satisfait-pas satisfait, sont courantes. Les données binaires recueillies sont ensuite utilisées pour estimer les proportions de la population présentant certaines caractéristiques. Dans le cadre de ce projet, nous proposons d'estimer les proportions de la population à partir d'échantillons provenant d'enquêtes non probabilistes. Nous proposons une procédure d'inférence de pseudo-vraisemblance empirique et montrons que l'estimateur ponctuel, qui en résulte pour une proportion de la population, a une propriété doublement robuste souhaitable. On propose deux méthodes de construction des intervalles de confiance du rapport de pseudo-vraisemblance empirique pour la proportion de la population : l'une est fondée sur la distribution limite de la statistique du rapport de pseudo-vraisemblance empirique ajusté, et l'autre utilise la pseudo-vraisemblance empirique bootstrap calibrée. Une étude de simulation montre que, lorsque la taille de l'échantillon est petite, les intervalles de confiance basés sur le rapport de pseudo-vraisemblance empirique ont un meilleur taux de couverture et un taux d'erreur plus équilibré dans les queues que les intervalles de confiance de type Wald couramment utilisés.

[Monday May 27/lundi 27 mai, 11:35-11:50]

Jules J. S. de Tibeiro (Université de Moncton) , **Filipe Afonso** (Symbad, Symbolic Data Lab) , **Edwin Diday** (Paris Dauphine University)

Analyzing Aggregated Household Consumption with Symbolic Data Analysis

Analyse de la consommation des ménages agrégée par analyse de données symboliques

This paper introduces an advanced setting of data analysis in the context of external information in Symbolic Data Analysis (SDA) of a dataset K = [X—Z]. A descriptive approach is applied on the whole dataset before we partition the dataset into two parts: the first part is data subset X for which we focus on explanation and prediction; the second part is denoted as data subset Z which consists of additional qualitative variables. More precisely, X contains homogenous records of expenses and is to be explained by Z (qualitative data), but Z might be quantitative as it is said to be socio–demographic variables such as age and income.

Cette présentation introduit une situation avancée d'analyse de données dans le contexte d'informations externes se présentant dans l'analyse de données symboliques (ADS) d'un jeu de données K = [X—Z]. Nous appliquons une approche descriptive à l'ensemble du jeu de données avant de le diviser en deux parties : la première est le sous-ensemble de données X pour lequel nous privilégions l'explication et la prédiction ; la deuxième, désignée sous-ensemble de données Z, consiste en variables qualitatives supplémentaires. Plus précisément, X contient des enregistrements de dépenses homogènes et est à expliquer par Z (« données qualitatives »), mais Z peut être quantitatif puisqu'il contient des « variables sociodémographiques » comme l'âge et le revenu.

In this study, we are therefore confronted with symbolic data where several types of symbolic variables appear simultaneously: certain interval variables and others with histogram values. Classes of individuals are described by taking into account the variation in the values of variables that characterize them.

Dans cette étude, nous sommes donc confrontés à des données symboliques pour lesquelles plusieurs types de variables symboliques apparaissent simultanément : certaines, des intervalles, et d'autres avec des valeurs d'histogramme. Nous décrivons des classes d'individus en tenant compte de la variation des valeurs qui les caractérisent.

Chair/Président: John Joseph Koval

Room/Salle: 109 (SS)

Abstract/Résumé

[Monday May 27/lundi 27 mai, 10:20-10:35]

François A Marshall (Queen's University)

A Statistical Characterization of Multitaper Statistical Detectors using Monte Carlo

Une caractérisation statistique des détecteurs statistiques ¡¡multitaper¿¿ à l'aide de Monte Carlo

In the multitaper spectral analysis of a physical process, statistical detectors are designed using the assumption that the distribution of the eigen-coefficients is multivariate complex-normal. The assumption is motivated by numerous examples in the physical sciences where the empirical distribution of the multitaper spectrum estimate agrees with a central chi-squared distribution. Moreover, there exist numerous central-limit theorems for the eigen-coefficients of the output of a first-order autoregressive filter given a typical IID innovations sequence. The results of a simulation study will be presented which demonstrate preservation of the complex-normal behaviour of these eigencoefficients upon changing: the filter type; and both the distribution and the correlation span of the input process. Emphasis will be placed on the selection of input processes based on what can be inferred from a multitaper estimate of the power spectrum.

Dans l'analyse spectrale ¡¡multitaper¿¿ d'un processus physique, les détecteurs statistiques sont conçus en partant de l'hypothèse que la distribution des coefficients propres est multivariée complexe-normale. Cette hypothèse est motivée par de nombreux exemples en sciences physiques où la fonction de répartition empirique de l'estimation du spectre ¡¡multitaper¿¿ correspond à une distribution chi-carré centrale. En outre, il existe de nombreux théorèmes de limite centrale pour les coefficients propres de la sortie d'un filtre autorégressif de premier ordre, à partir d'une séquence d'innovations IID typique. Les résultats d'une étude de simulation illustrant la préservation du comportement complexe-normal de ces coefficients propres lors du changement : du type de filtre ; et à la fois la distribution et la durée de corrélation du processus d'entrée. Nous insisterons sur la sélection des processus d'entrée fondée sur l'inférence à l'aide d'un estimateur ¡¡multitaper¿¿ du spectre de puissance.

[Monday May 27/lundi 27 mai, 10:35-10:50]

Feiyu Zhu (University of Waterloo) , **Martin Lysy** (University of Waterloo)

Particle Physics Representation of a Continuous Stationary Gaussian Process

Représentation en physique des particules d'un processus gaussien stationnaire continu

The Generalized Langevin Equation (GLE) is a fundamental result in statistical mechanics describing the stochastic evolution of observables in interacting-particle systems. In the linear case, the GLE corresponds to a continuous stationary Gaussian process, a particularly useful first-order approximation in many single-molecule biophysical experiments. However, dynamics of the observed particle position process must be deconvoluted from those of a latent force, rendering statistical modeling and parameter estimation typically intractable. Here, we prove that any continuous stationary Gaussian process can be expressed as a GLE, providing an explicit inverse representation of the force in terms of position. In addition to the immediate up-

L'équation généralisée de Langevin (EGL) est un résultat fondamental en mécanique statistique qui décrit l'évolution stochastique de données observables dans des systèmes d'interaction de particules. Dans le cas linéaire, l'EGL correspond à un processus gaussien stationnaire continu, une approximation de premier ordre particulièrement utile dans plusieurs expériences biophysiques à molécule unique. Par contre, la dynamique du processus de position de la particule observée doit être déconvoluée à partir de celle d'une force latente, rendant la modélisation statistique et l'estimation de paramètres typiquement insolubles. Nous prouvons ici que n'importe quel processus gaussien stationnaire continu peut s'exprimer comme une EGL, produisant une représentation inverse explicite de la force en termes de la position. Outre le résultat immédiat de la modélisation directe des EGLs linéaires dans le

shot of straightforward linear GLE modeling directly in the observable domain, the result provides insights for modeling nonlinear stationary physical processes.

domaine observable, le résultat offre un aperçu de la modélisation de processus physiques stationnaires non linéaires.

[Monday May 27/lundi 27 mai, 10:50-11:05]

Armin Hatefi (Memorial University of Newfoundland) , **Mohammad Jafari Jozani** (University of Manitoba) , **Omer Ozturk** (Ohio State University)

Finite Mixture Modeling Based on Multi-Observer Sampling Design

Modèles de mélange finis fondés sur un plan d'échantillonnage d'ensembles ordonnés multi-observateurs

We study the problem of finite mixture modeling based on multi-observer ranked-set sampling design. Ranked-set sampling is used in situations where obtaining the measurements of variables of interest is costly but rank information about sampling units can be obtained easily. We consider the cases where single or multiple observers with different ranking potentials are available to assign judgmental ranks to sampling units. We propose maximum likelihood estimation and model-based classification procedures using multi-observer ranked-set samples from finite mixture models. A suitable expectation-maximization algorithm is developed to incorporate the rank information in the estimation of the model parameters. Through simulation studies, we investigate the performance of estimation and classification methods. The proposed methods are finally used for analysis of a real dataset.

Nous étudions le problème des modèles de mélange finis fondés sur un plan d'échantillonnage d'ensembles ordonnés multi-observateurs. L'échantillonnage d'ensembles ordonnés est utilisé dans des situations où obtenir les mesures des variables d'intérêt est coûteux mais l'information sur le rang des unités d'échantillonnage peut être facilement obtenue. Nous examinons les cas où des observateurs uniques ou multiples avec des potentiels de classement différents sont disponibles pour assigner les rangs subjectifs aux unités d'échantillonnage. Nous proposons une estimation par la méthode du maximum de vraisemblance et des procédures de classification fondées sur le modèle en utilisant des échantillons d'ensembles ordonnés multi-observateurs provenant de modèles de mélange finis. Un algorithme d'espérance-maximisation adapté est développé pour inclure l'information sur le rang dans l'estimation des paramètres du modèle. Nous évaluons la performance de l'estimation et des méthodes de classification par l'entremise d'études de simulation. Les méthodes proposées sont finalement utilisées pour l'analyse d'un ensemble de données réelles.

[Monday May 27/lundi 27 mai, 11:05-11:20]

Jeffrey D. Picka (University of New Brunswick)

Ontological Aspects of Statistical Modelling

Aspects ontologiques de la modélisation statistique

When complex models are fitted to big data, some process of validation is necessary. This process involves mathematics and computation, but the process also must evaluate the kind of knowledge that can be constructed from the model. Evaluation of validation processes requires examining how statistical models are assigned to different classes of knowledge, and involves confronting the consequences of the careless use of misleading mathematical analogies during the validation process. Simple examples of knowledge classes for statistical models will be presented, together with appropriate mathematical analogies for their use.

Lorsque des modèles complexes sont ajustés à des mégadonnées, un processus de validation est nécessaire. Ce processus fait appel aux mathématiques et au calcul, mais il doit aussi évaluer le type de connaissances que l'on peut bâtir à partir du modèle. L'évaluation des processus de validation exige d'examiner comment les modèles statistiques sont attribués aux différentes classes de connaissances et de faire face aux conséquences de la mauvaise utilisation d'analogies mathématiques trompeuses au cours du processus de validation. Nous présenterons des exemples simples de classes de connaissances pour les modèles statistiques, ainsi que des analogies mathématiques appropriées pour leur utilisation.

[Monday May 27/lundi 27 mai, 11:20-11:35]

Anthony Coache (Université du Québec à Montréal) , **François Watier** (Université du Québec à Montréal)

Stochastic Algorithms for Solving a Multi-Period Quantile-Based Portfolio Optimization Problem

Algorithmes stochastiques pour résoudre un problème d'optimisation multi-périodique de portefeuille basé sur un quantile

In financial asset allocation, while many risk measures are proposed, care must be taken to ensure that their properties reflect an observed investor behavior. Furthermore, the investor should also be able to adjust his strategy according to market fluctuations during the investment period. Thus we focus our analysis on a class of multiperiod portfolio optimization problems in which the investor wants to minimize a quantile-based function of the terminal wealth and where the stock prices follow a binomial tree model. We explore various stochastic approaches in finding an optimal or near-optimal strategy and compare their efficiency through simulation studies.

En théorie moderne du portefeuille, bien que plusieurs mesures de risques soient proposées, il faut prendre bien soin que celles-ci comportent des propriétés traduisant un comportement observé des investisseurs. De plus, l'investisseur devrait pouvoir réagir aux fluctuations du marché et adapter sa stratégie financière en conséquence pendant la période d'investissement. Ainsi, nous concentrons notre analyse sur une classe de problèmes d'optimisation multi-périodique de portefeuille pour laquelle l'investisseur souhaite minimiser une fonction basée sur un quantile de la richesse terminale et où le prix des actifs financiers suit un modèle d'arbre binomial. Nous explorons diverses approches stochastiques pour trouver une stratégie optimale ou quasi-optimale et nous comparons leur efficacité par des études de simulation.

[Monday May 27/lundi 27 mai, 11:35-11:50]

Nirodha Mihirani Epasinghege Dona (University of Manitoba) , **Brad Johnson** (University of Manitoba)

Estimating Random Walk Centrality

Estimation de la centralité par marche aléatoire

Centrality measures play an important role in determining the importance of nodes in networks. For strongly connected networks, the random walk centrality measures how easy it is to reach a given state from another randomly chosen state. This measure requires calculating a generalized group inverse for the transition matrix, which can be computationally difficult for large state spaces. It is known that the random walk centrality for a particular state can be written as a function of the first and second moments of the first passage times for that state. In this study, using realization of random walks, we estimate the distributions of first passage times by using a number of statistical methods, including Bayesian bootstrap and two Poisson mixture model approaches. Finally, we compare the resulting estimates of the random walk centrality measures to the true values.

Les mesures de centralité jouent un rôle important afin de déterminer l'importance des nœuds dans les réseaux. Pour les réseaux fortement connectés, la centralité par marche aléatoire mesure à quel point il est facile d'atteindre un état donné à partir d'un autre état choisi au hasard. Cette mesure nécessite le calcul d'un inverse généralisé de groupe de la matrice de transition, ce qui peut être difficile à calculer pour les grands espaces d'état. On sait que la centralité par marche aléatoire pour un état particulier peut être écrite en fonction des premier et deuxième moments des premiers temps de passage pour cet état. Dans le cadre de cette étude, nous utilisons la réalisation de marches aléatoires pour estimer les distributions des temps de premier passage au moyen d'un certain nombre de méthodes statistiques, dont le bootstrap bayésien et deux approches de modèle de mélange de Poisson. Enfin, nous comparons les estimations qui découlent des mesures de centralité par marche aléatoire aux valeurs réelles.

Case Study 1: Counting Cells from Microscopic Images
Étude de cas 1 : Comptage de cellules dans des images microscopiques

Chair/Président: Pingzhao Hu

Room/Salle: 103Z (ST)

Abstract/Résumé

Case Study 2: Risk of Cardiovascular Disease among Osteoarthritis Patients: Exploring the Relationship in a National Health Survey
Étude de cas 2 : Risque de malade cardiovasculaire chez les patients souffrant d'arthrose : étude de la relation dans une enquête nationale sur la santé

Chair/Président: Pingzhao Hu

Room/Salle: 103Z (ST)

Abstract/Résumé

Stochastic Processes and Applications
Processus et applications stochastiques

Chair/Président: Mary E. Thompson

Organizer/Responsable: Mary E. Thompson

Room/Salle: 144 (SB)

Abstract/Résumé

[Monday May 27/lundi 27 mai, 13:30-14:00]

Lam Ho (Dalhousie University) , **Vu Dinh** (University of Delaware) , **Frederick A. Matsen** (Fred Hutchinson Cancer Research Center) , **Marc Suchard** (University of California, Los Angeles)

On the Consistency of the MLE for the Transition Rate of a Two-State Symmetric Markov Process on a Tree

La convergence des estimateurs du maximum de vraisemblance pour le taux de transition d'un processus de Markov symétrique à deux états sur un arbre

Maximum likelihood estimators (MLEs) are used extensively to estimate unknown parameters of stochastic models of phenotypic evolution. Although the consistency of the MLE has been proven for independent data, we cannot appeal to this result because phenotypes of different species are highly correlated due to the fact that species are related to each other according to an evolutionary tree. In this work, we investigate the asymptotic properties of the MLE for estimating the transition rate of a binary phenotype evolved along a phylogenetic tree according to a two-state symmetric Markov process.

Les estimateurs du maximum de vraisemblance sont souvent utilisés pour estimer les paramètres inconnus des modèles stochastiques de l'évolution phénotypique. Bien qu'on ait prouvé la convergence de l'estimateur du maximum de vraisemblance pour les données indépendantes, nous ne pouvons pas recourir à ce résultat, car les phénotypes de différentes espèces sont fortement corrélés du fait que les espèces sont apparentées les unes aux autres selon un arbre évolutif. Dans le cadre de ces travaux, nous étudions les propriétés asymptotiques de l'estimateur du maximum de vraisemblance pour estimer le taux de transition d'un phénotype binaire évolué le long d'un arbre phylogénétique selon un processus de Markov symétrique à deux états.

[Monday May 27/lundi 27 mai, 14:00-14:30]

Hélène Guérin (Université du Québec à Montréal) , **Ninon Fétique** (Université de Tours) , **Florent Malrieu** (Université de Tours)

Long Time Behavior of Interacting Zig-Zag Particles

Le comportement en temps long des particules Zig Zag en interaction

The Zig-Zag process is a generalization of the telegraph process belonging to the class of piecewise deterministic Markov processes. It has been recently introduced to model the behavior of flagelated bacteria, as E-Coli, in their environment. Indeed the behavior of such bacteria is composed of run phases with constant velocity and of tumble phases where we observe a quick change of the velocity. Under some good assumptions this process converges exponentially fast to an explicit invariant measure. New MCMC algorithms based on the Zig-Zag process have been recently introduced and studied. In this talk we will consider some interacting Zig-Zag processes. We will introduce a system of Zig-Zag particles attracted by the mean position of the system. The question of propagation of chaos will be studied and the

Le processus Zig-Zag est un processus du Télégraphe généralisé, qui appartient à la classe des processus de Markov déterministes par morceaux. Il a été récemment introduit pour modéliser le comportement des bactéries flagellées, comme les E-Coli, dans leur environnement. En effet, le comportement de ces bactéries est composé d'une succession de mouvements linéaires et de changements rapides de la vitesse. Sous de bonnes hypothèses, ce processus converge exponentiellement et rapidement vers une mesure invariante explicite. De nouveaux algorithmes MCCM ont été récemment introduits à l'aide de ce processus. Dans cet exposé, nous introduirons un système de particules Zig-Zag attirées par la position moyenne du système. La question de la propagation du chaos sera étudiée et les propriétés du processus de limite non-linéaire telles que son comportement en temps long seront présentées. Ce travail est une collaboration avec N. Fétique et F.

properties of the nonlinear limit process such as its long time behavior will be presented. This is a joint work with N. Fétique and F. Malrieu (Université de Tours, France).

Malrieu (Université de Tours, France).

[Monday May 27/lundi 27 mai, 14:30-15:00]

Yaozhong Hu (University of Alberta) , **Khoa Le** (Imperial College London)

Density of Parabolic Anderson Random Variable

Densité de la variable aléatoire parabolique d'Anderson

The solution u(t,x) to a stochastic partial differential equation $\frac{\partial}{\partial t}u(t,x) = \frac{1}{2}\Delta u(t,x) + u \diamond \dot{W}(t,x)$ is a random variable for any fixed t and x, where $\dot{W}$ is a general Gaussian noise and $\diamond$ denotes the Wick product. We will prove that this random variable has a probability density $\rho(t,x;y)$ and we will be mainly concerned with the asymptotic behavior of $\rho(t,x;y)$ when $y \to \infty$ or when $t \to 0+$. Both upper and lower bounds are obtained and these two bounds match each other modulo some multiplicative constants. If the initial data is positive, then $\rho(t,x;y)$ is supported on the positive half line $y \in [0,\infty)$ and in this case we show that $\rho(t,x;0+) = 0$ and obtain an upper bound for $\rho(t,x;y)$ when $y \to 0+$.

La solution $u(t,x)$ d'une équation aux dérivées partielles stochastique $\frac{\partial}{\partial t}u(t,x) = \frac{1}{2}\Delta u(t,x) + u \diamond \dot{W}(t,x)$ est une variable aléatoire pour tous t et x fixes, où $\dot{W}$ est un bruit gaussien général et $\diamond$ dénote le produit de Wick. Nous prouvons que cette variable aléatoire présente une densité de probabilité $\rho(t,x;y)$ et nous nous concentrons sur le comportement asymptotique de $\rho(t,x;y)$ quand $y \to \infty$ ou quand $t \to 0+$. Nous obtenons les limites supérieures et inférieures, qui se correspondent modulo certaines constantes multiplicatives. Si les données initiales sont positives, alors $\rho(t,x;y)$ vaut sur la demi-droite positive $y \in [0,\infty)$ et dans ce cas nous montrons que $\rho(t,x;0+) = 0$ et obtenons une limite supérieure pour $\rho(t,x;y)$ quand $y \to 0+$.

Measuring the Quality of Multisource Statistics
Évaluation de la qualité des statistiques multisources

Chair/Président: Katherine Jenny Thompson

Organizer/Responsable: Wesley Yung

Room/Salle: 201 (ENA)

Abstract/Résumé

[Monday May 27/lundi 27 mai, 13:30-14:00]

John Eltinge (US Census Bureau)

Assessment of Inferential Quality in the Integration of Multiple Data Sources

Évaluation de la qualité inférentielle dans l'intégration de sources de données multiples

Statistical agencies are focusing attention on integration of multiple data sources (e.g., sample surveys; administrative and commercial records; and social-media traces). In this evolving environment, it is important to have a realistic assessment of inferential properties of the resulting estimators. This paper addresses those issues through review and synthesis of four complementary streams of literature: (1) Assessment of variability encountered with one or more data sources (e.g., multiple-frame, propensity and superpopulation approaches), (2) Extension of total survey error models to include decomposition of estimation error arising in multiple-source cases, (3) Evaluation of the impact of point-estimation bias on properties of interval estimators, and (4) Implications of the preceding for transparency, reproducibility and replicability of procedures for integration of multiple sources. The talk will close with comments on realistic communication with data users.

L'attention des organismes de statistique se porte sur l'intégration de sources de données multiples (par ex. : sondages par échantillonnage, dossiers administratifs et commerciaux et traces laissées sur les réseaux sociaux). Dans cet environnement évolutif, il est important de procéder à une évaluation réaliste des propriétés inférentielles des estimateurs qui en résultent. Cet article s'intéresse à ces questions en passant en revue et en synthétisant quatre courants complémentaires dans la documentation : (1) Évaluation de la variabilité observée avec une ou plusieurs sources de données (par ex. : approches multicadres, prophilie et superpopulation) (2) Extension des modèles d'erreur totale d'enquête pour y inclure la décomposition de l'erreur d'estimation découlant de cas à sources multiples (3) Évaluation de l'impact du biais d'estimation ponctuelle sur les propriétés des estimateurs d'intervalles (4) Implications des points (1)-(3) pour la transparence, la reproductibilité et la réplicabilité des procédures pour l'intégration de sources multiples. L'article conclut avec des commentaires sur une communication réaliste avec les utilisateurs de données sur les points (1)-(4).

[Monday May 27/lundi 27 mai, 14:00-14:30]

Susie Fortier (Statistique Canada) , **Martin Beaulieu** (Statistics Canada) , **Ryan Chepita** (Statistics Canada)

Defining, Measuring and Communicating Quality in a Multi-Source Environment

Définir, mesurer et transmettre la qualité dans un environnement multisources

The methods and language for measuring and communicating data quality mostly originate from sampling theory. While the basic concepts were extended to take into consideration non-sampling errors in the Total Survey Error Framework, the question of how to fully define and measure quality in the new/big data paradigm in a multi-source environment is still open. This talk will cover two areas from the viewpoint of official statistics and in the specific context of Statistics Canada. It covers (a) recent exploration of the theoretical framework and (b) how to address the immediate needs of data users

Les méthodes et le langage utilisés pour mesurer et transmettre la qualité des données découlent surtout de la théorie de l'échantillonnage. Bien que les notions de base aient été élargies pour prendre en compte les erreurs non dûes à l'échantillonnage dans le cadre de l'erreur totale d'enquête, une question demeure : comment définir et mesurer entièrement la qualité du paradigme des données nouvelles ou volumineuses dans un environnement multisources. Cet article présente deux sujets selon la perspective statistique officielle et dans le contexte propre à Statistique Canada. Il aborde (a) une exploration récente du cadre théorique et (b) le moyen de répondre aux besoins immédiats des utilisateurs

and producers. The research work builds on the well-known dimensions of quality in the context of official statistics and, amongst others goals, aims to explore a whole-of-government quality framework.

et producteurs de données. Le travail de recherche se développe à partir des dimensions bien connues de la qualité dans le contexte de la statistique officielle et a notamment pour objectif d'explorer un cadre de la qualité pour l'ensemble du gouvernement.

[Monday May 27/lundi 27 mai, 14:30-15:00]

Rachel Skentelbery (Office for National Statistics United Kingdom) , **Hannah Finselbach** (Office for National Statistics UK)

Measuring the Quality of Multisource Statistics

Mesurer la qualité des statistiques multisources

The Office for National Statistics (ONS) is transforming to put administrative and alternative data sources at the core of our statistics. Combining new sources with surveys will allow us to meet the ever-increasing user demand for improved and more detailed statistics. Like many national statistics organizations, we want to exploit the rich data sources available from government agencies and digital platforms. However, using these data involve addressing a range of statistical challenges, including the need to measure and communicate quality and uncertainty. ONS has established a methodological research programme to develop a theoretical framework to effectively use and integrate new data sources. This talk will present a high-level view of, and progress against, our current research priorities, including frameworks for measuring and/or describing: the quality of source data errors or uncertainty introduced in processing, and quality indicators or measures of final output.

Le Bureau des statistiques nationales (BSN) se transforme actuellement, pour mettre les sources de données administratives et autres au cœur de nos statistiques. La combinaison de nouvelles sources avec les sondages nous permettra de répondre à la demande toujours croissante des utilisateurs de statistiques améliorées et plus détaillées. Comme bon nombre d'organismes de statistiques nationales, nous souhaitons exploiter les riches sources de données qu'offrent les organismes gouvernementaux et les plateformes numériques. Cependant, l'utilisation de ces données exige que soient abordés divers défis statistiques, notamment la capacité à mesurer et à communiquer la qualité et l'incertitude. Le BSN a créé un programme de recherche méthodologique chargé d'élaborer un cadre théorique pour l'utilisation et l'intégration efficaces de nouvelles sources de données. Nous présenterons une analyse de haut niveau et l'état d'avancement des travaux par rapport à nos priorités de recherche actuelles, et notamment les cadres permettant de mesurer et/ou de décrire : • la qualité des données sources ; • les erreurs ou l'incertitude créée par le traitement • les indicateurs de qualité ou les mesures des produits finaux.

Measurement Error Models and Its Impacts in Health Sciences
Modèles d'erreur de mesure et impacts sur les sciences de la santé

Chair/Président: Mahmoud Torabi

Organizer/Responsable: Mahmoud Torabi

Room/Salle: 101 (ENA)

Abstract/Résumé

[Monday May 27/lundi 27 mai, 13:30-14:00]

Grace Yi (University of Waterloo)

Analysis of Multi-State Models with Misclassified States

Analyse des modèles multi-états avec erreurs de classification des états

Multi-state models are commonly used in studies of disease progression. Methods developed under this framework, however, are often challenged by misclassification in states. In this talk, I will discuss issues concerning continuous time progressive multistate models with state misclassification. I will describe inference methods using both the likelihood and pairwise likelihood methods that are based on the joint modelling of the transition and misclassification processes. The performance of the estimation procedures is evaluated by numerical studies.

Les modèles multi-états sont couramment utilisés dans les études sur la progression de la maladie. Toutefois, les méthodes élaborées dans ce cadre sont souvent remises en question en raison d'erreurs de classification dans les états. Dans cet exposé, j'aborderai les questions concernant les modèles multi-états progressifs à temps continu avec une classification erronée des états. Je décrirai les méthodes d'inférence à l'aide de méthodes de vraisemblance et de vraisemblance par paire qui sont fondées sur la modélisation conjointe des processus de transition et de la classification erronée. L'efficacité des procédures d'estimation est évaluée au moyen d'études numériques.

[Monday May 27/lundi 27 mai, 14:00-14:30]

Liqun Wang (University of Manitoba) , **Lin Xue** (University of Manitoba) , **Hengjian Cui** (Capital Normal University)

Variable Selection and Estimation in Generalized Linear Models with Measurement Error

Sélection et estimation de variables dans les modèles linéaires généralisés avec erreur de mesure

Regularization methods for high-dimensional variable selection and estimation have been intensively studied in recent years and most of them are developed in the framework of linear regression models where the predictor variables are assumed to be accurately measured. However, in real data analysis it is common that some predictors cannot be measured directly or precisely. While it is well known that measurement error in predictors causes attenuation in parameter estimation, its impact on variable section is not well studied. We study this problem in the framework of generalized linear models and propose an instrumental variable approach to correct the bias in variable section and estimation. We present some theoretical results as well as numerical examples.

Les méthodes de régularisation de la sélection et de l'estimation des variables de grande dimension ont fait l'objet d'études intensives ces dernières années, la plupart étant développées dans le cadre de modèles de régression linéaire où on assume que les variables explicatives sont bien mesurées. Cependant, en analyse de données il est fréquent que certaines variables explicatives ne puissent être mesurées directement ou avec précision. Or même si l'on sait que l'erreur de mesure de ces variables cause une atténuation de l'estimation des paramètres, l'impact de cette erreur sur la sélection de variables n'a pas été étudiée de manière satisfaisante. Nous étudions ce problème dans le cadre des modèles linéaires généralisés et proposons une approche de variables instrumentales pour corriger le biais dans la sélection et l'estimation des variables. Nous présentons des résultats théoriques et des exemples numériques.

[Monday May 27/lundi 27 mai, 14:30-15:00]

Juxin Liu (University of Saskatchewan) , **Annshirley Afful** (University of Saskatchewan)

Joint Misclassification Errors in Both Response and Explanatory Variables

Erreurs de classification conjointes dans les variables de réponse et explicatives

It is commonly encountered in many applications that variables are not measured perfectly, which is known as errors in variables (EIV) in the statistical literature. It has been long recognized that simply ignoring EIV leads to misleading inference results. There has been extensive work focusing on EIV in explanatory variables only. EIV in both response and explanatory variables has received very limited attention especially when they are discrete/categorical. Our work focuses on the joint misclassification errors in a binary response variable and a binary explanatory variable. To account for the possible dependence between the misclassification errors in both variables, we introduce the dependence parameters following the notion in Vogel et al. (2005). We conduct sensitivity analysis to check the consequence of fitting an independent misclassification model to data actually generated from dependent misclassification models.

Il est commun, dans de nombreuses applications, que les variables ne soient pas mesurées parfaitement ce qui est connu comme erreurs dans les variables (EDV) dans la littérature statistique. Il est reconnu que de simplement ignorer les EDV donne lieu à des résultats d'inférence trompeurs. Un travail considérable a été effectué se concentrant seulement sur les EDV dans les variables explicatives. Par contre, peu d'attention a été consacrée aux EDV à la fois dans les variables de réponse et dans les variables explicatives, particulièrement quand elles sont discrètes/catégorielles. Notre travail est axé sur les erreurs de classification conjointes dans la variable de réponse binaire et dans la variable explicative binaire. Pour tenir compte de la dépendance possible entre les erreurs de classification dans les deux variables, nous introduisons des paramètres de dépendance suite au concept dans Vogel et al. (2005). Nous menons une analyse de sensibilité pour identifier les conséquences de l'ajustement d'un modèle d'erreurs de classification indépendant à des données générées à partir de modèles d'erreurs de classification dépendants.

Rocky and Atlantic Collaborations in the Health Sciences
Collaborations en sciences de la santé dans les Rocheuses et l'Atlantique

Chair/Président: John Braun

Organizer/Responsable: John Braun

Room/Salle: 122 (ICT)

Abstract/Résumé

[Monday May 27/lundi 27 mai, 13:30-15:00]

John Braun (The University of British Columbia) , **Rasika Rajapakshe** (BC Cancer Agency) , , **Yingqi Wang** (University of Calgary) , , **Andrew Jirasek** (University of British Columbia) , , **Renjun Ma** (University of New Brunswick) , , **Henrik Stryhn** (University of Prince Edward Island)

Rocky and Atlantic Collaborations in the Health Sciences

Collaborations en sciences de la santé dans les Rocheuses et l'Atlantique

Health Science Collaborating Centres (HSCC) have been now been established at several sites across Canada, with the common purposes of increasing research collaborations between health and statistical scientists and training the next generation of statisticians. We report on two HSCCs, one from the West and the one from the Atlantic region. The session will highlight the collaboration experience from different perspectives: a trainee who studied expert-informed regression modelling with the LASSO, a medical physicist who has begun collaborating intensively with statisticians and computer scientists, and of course, biostatisticians.

Des Centres de collaboration en sciences de la santé (CCSS) ont été établis sur divers sites partout au pays en vue de multiplier les collaborations de recherche entre scientifiques de la santé et de la statistique et de former la prochaine génération de statisticiens. Nous présentons deux CCSS, l'un de l'Ouest et l'autre de la région Atlantique. La séance portera sur différents aspects de l'expérience de collaboration : un stagiaire qui a étudié la modélisation de la régression informée par experts à l'aide du LASSO, un physicien médical qui a commencé d'intenses collaborations avec des statisticiens et des informaticiens et, bien entendu, des biostatisticiens.

Computational challenges in statistical learning for complex data
Défis computationnels en apprentissage statistique pour données complexes

Chair/Président: Teng Zhang

Room/Salle: 102 (ICT)

Abstract/Résumé

[Monday May 27/lundi 27 mai, 13:30-14:00]

Marianna Pensky (University of Central Florida) , **Rasika Rajapakshage** (University of Central Florida)

Clustering in Statistical Ill-Posed Linear Inverse Problems

Regroupement dans les problèmes inverses linéaires statistiques mal posés

In many statistical linear inverse problems, one needs to recover classes of similar curves from their noisy images under an operator that does not have a bounded inverse. Problems of this kind appear in many areas of application. Routinely, in such problems clustering is carried out at the pre-processing step and then the inverse problem is solved for each of the cluster averages separately. As a result, the errors of the procedures are usually examined for the estimation step only. The objective of this talk is to examine, both theoretically and via simulations, the effect of clustering on the accuracy of the solutions of general ill-posed linear inverse problems.

Dans bon nombre de problèmes inverses linéaires statistiques, il faut identifier des classes de courbes similaires à partir d'images bruitées avec un opérateur dont il n'existe pas d'inverse borné. On trouve des problèmes de ce type dans de nombreux domaines d'application. Habituellement dans ces problèmes, on procède au regroupement lors de l'étape de prétraitement puis on résout le problème inverse pour chacune moyenne de regroupement. Par conséquent, les erreurs des procédures ne sont généralement examinées qu'à l'étape de l'estimation. L'objectif de cette présentation est d'examiner, à la fois théoriquement et via des simulations, l'effet du regroupement sur l'exactitude des solutions de problèmes inverses linéaires généraux mal posés.

[Monday May 27/lundi 27 mai, 14:00-14:30]

Yuwen Gu (University of Connecticut) , **Simon Fontaine** (University of Montreal) , **Yi Yang** (McGill University) , **Wei Qian** (University of Delaware) , **Bo Fan** (University of Oxford)

A Unified Approach to Sparse Tweedie Modeling of Multi-Source Insurance Claims Data

Une approche unifiée pour modélisation éparse Tweedie de données de réclamations d'assurance à sources multiples

Actuarial practitioners now have access to multiple sources of insurance data corresponding to various situations: multiple business lines, umbrella coverage, multiple hazards, and so on. Despite the wide use and simple nature of single-target approaches, modeling these types of data may benefit from a simultaneous approach. We propose a unified algorithm to perform sparse learning of such fused insurance data under the Tweedie (compound Poisson) model. By integrating ideas from multi-task sparse learning and sparse Tweedie modeling, our algorithm produces flexible regularization that balances predictor sparsity and between-sources sparsity. When applied to simulated and real data, our approach clearly outperforms single-target modeling in both prediction and selection accuracy, notably when the sources do not have exactly the same set of predictors. An efficient implementation of the proposed algorithm is provided in our R package MStweedie.

Les actuaires ont maintenant accès à de multiples sources de données d'assurance concernant diverses situations : plusieurs secteurs d'activité, couverture parapluie, risques multiples, et ainsi de suite. Malgré la grande utilisation et la nature simple des approches à cible unique, la modélisation de ce genre de données peut profiter d'une approche simultanée. Nous proposons un algorithme unifié pour accomplir un apprentissage épars de ces données d'assurance fusionnées sous le modèle Tweedie (Poisson composé). En intégrant des idées de l'apprentissage épars à tâches multiples et du modèle épars Tweedie, notre modèle produit une régularisation flexible qui balance l'éparsité du prédicteur et l'éparsité entre sources. Lorsqu'elle est appliquée à des données réelles et simulées, notre approche surpasse nettement la modélisation avec cible unique dans la précision de la prédiction et de la sélection, notamment quand les sources n'ont pas exactement le même nombre de variables prédictives. Une exécution efficace de l'algorithme proposé est fournie dans notre bibliothèque R MStweedie.

[Monday May 27/lundi 27 mai, 14:30-15:00]

Computational challenges in statistical learning for complex data
Défis computationnels en apprentissage statistique pour données complexes

Asad Haris (McGill University) , **Robert Platt** (McGill University)

A Targeted Approach to Confounder Selection for High-Dimensional Data

Approche ciblée dans le choix de variables de confusion pour des données de grande dimension

We consider the problem of selecting confounders for adjustment from a potentially large set of covariates. Recently, the high-dimensional Propensity Score (hdPS) method was developed for this task; hdPS ranks potential confounders by estimating an importance score for each variable and selects the top few variables. However, this ranking procedure is limited: it requires all variables to be binary. We propose an extension of the hdPS score to general types of variables. We also develop a group importance score, allowing us to rank groups of potential confounders. The main challenge is that our parameter requires either the propensity score or response model; both vulnerable to model misspecification. We propose a targeted maximum likelihood estimator (TMLE) which allows the use of nonparametric, machine learning tools for estimating the intermediate models. We establish asymptotic normality of our estimator, which consequently allows constructing confidence intervals.

Nous réfléchissons au problème de la sélection des variables de confusion pour un ajustement à partir d'un ensemble de covariables possiblement volumineux. Récemment, la méthode du score de propension de dimension élevée (SPDE) a été élaborée pour cette tâche ; le SPDE classe les variables de confusion potentielles en estimant un score d'importance pour chacune et sélectionne quelques variables en tête de classement. Cette procédure de classement est cependant limitée et requiert que toutes les variables soient binaires. Nous proposons d'étendre le SPDE à des types généraux de variables. Nous développons aussi un score d'importance de groupes permettant de classer les groupes de variables de confusion potentielles. Le problème principal est que notre paramètre demande soit un score de propension, soit un modèle de réponse, qui tous deux sont vulnérables aux erreurs de spécification du modèle. Nous proposons un estimateur du maximum de vraisemblance ciblé (EMVC) permettant l'utilisation d'outils d'apprentissage automatique non paramétriques pour l'estimation de modèles intermédiaires. Nous établissons une normalité asymptotique de notre estimateur qui permet conséquemment de déterminer des intervalles de confiance.

Diagnostic Tests and Prediction Models
Tests diagnostiques et modèles prédictifs

Chair/Président: Farouk Nathoo

Room/Salle: 142 (AD)

Abstract/Résumé

[Monday May 27/lundi 27 mai, 13:30-13:45]

Alexander de Leon (University of Calgary) , **Joyce Raymond Punzalan** (University of the Philippines Diliman) , **Hua Shen** (University of Calgary)

Estimation of Diagnostic Accuracy Measures of Correlated Diagnostic Tests for Paired Organs in the Absence of a Gold Standard

Estimation des mesures d'exactitude de diagnostic de tests diagnostiques correlés d'organes pairs en l'absence d'un étalon de référence

Paired diagnostic data arise when fellow organs undergo diagnostic tests to predict the presence of certain diseases. In the absence of a gold standard, we adopt a finite mixture of extended common correlation models for the correlated diagnostic outcomes of multiple tests on fellow organs, to account for conditional dependence among the tests. We discuss identifiability of our joint model and outline an ES algorithm, a variant of the EM algorithm, for estimating parameters of the joint model, including the tests' accuracy measures. We provide simulations on the finite-sample bias and efficiency of the resulting estimates and the impact on estimation of incorrectly assuming conditional independence of the tests. We analyze real diagnostic data on dry eye disease, for which no gold standard is available, to illustrate our methodology.

On obtient des données diagnostiques appariées lorsque des tests diagnostiques sont menés sur des organes pairs afin de prédire la présence de certaines maladies. En l'absence d'un étalon de référence, nous adoptons un mélange fini de modèles étendus de corrélation commune pour les résultats diagnostiques correlés de tests multiples sur les organes pairs, afin de prendre en compte la dépendance conditionnelle entre les tests. Nous traitons de l'identifiabilité de notre modèle conjoint et présentons un algorithme espérance-solution (ES), une variante de l'algorithme espérance-maximisation (EM), afin d'estimer les paramètres du modèle conjoint, y compris les mesures d'exactitude des tests. Nous présentons des simulations sur le biais d'échantillon fini et de l'efficacité des estimations obtenues et aussi l'impact sur l'estimation d'une présomption erronée de l'indépendance conditionnelle des tests. Nous illustrons notre méthodologie avec une analyse des données de diagnostic réelles sur la sécheresse oculaire, une affection pour laquelle aucun étalon de référence n'est établi.

[Monday May 27/lundi 27 mai, 13:45-14:00]

Meng Yuan (University of Waterloo) , **Changbao Wu** (University of Waterloo) , **Pengfei Li** (University of Waterloo)

Semiparameteric Inference of the Youden Index and Optimal Cut-Off Point

Inférence semiparamétrique de l'index de Youden et de la valeur-seuil optimale

In biomedical research, the receiver operating characteristic (ROC) curve is a widely used statistical tool to evaluate the discriminatory effectiveness of a biomarker for distinguishing diseased individuals from healthy ones. The Youden Index is a popular summary statistic of the ROC curve. In this paper, we propose a semiparametric estimator of the Youden Index and its optimal cut-off point based on the density ratio model (DRM). However, data with a fixed lower limit of detection (LOD) often arise in biomedical research. We also propose DRM-based estimators in the case that samples are with the LOD. Asymptotic properties of all estimators are de-

En recherche biomédicale, la courbe ROC (efficacité du récepteur) est un outil statistique largement utilisé pour évaluer l'efficacité d'un biomarqueur à distinguer des individus malades ou en bonne santé. L'index de Youden est une statistique sommaire populaire de la courbe ROC. Dans cette présentation, nous proposons un estimateur semiparamétrique de l'index de Youden et de sa valeur-seuil optimale en fonction du modèle de rapport de densités (MRD). Cependant, les recherches biomédicales produisent souvent des données avec une limite inférieure de détection (LID) fixe. Nous proposons également des estimateurs fondés sur le MRD pour les échantillons avec LID. Nous dérivons les propriétés asymptotiques de tous ces estimateurs. Nous effectuons

rived. Simulation studies are conducted over a range of distributional scenarios and sample sizes to compare the performance of some existing methods and our semi-parametric method. In terms of relative bias and mean squared error, our proposed estimators have a remarkable efficiency gain.

des études de simulation sur divers scénarios de distribution et de tailles d'échantillons pour comparer la performance de certaines méthodes existantes et la nôtre. En termes de biais relatif et d'erreur quadratique moyenne, les estimateurs que nous proposons présentent un net gain en efficacité.

[Monday May 27/lundi 27 mai, 14:00-14:15]

Wanhua Su (MacEwan University)

Determining the Optimal Break-Point(s) Based on Precision-Recall Curves

Détermination du point de rupture optimal par courbes de précision-rappel

Two diagnostic tools for binary classification predictive modelling problems are Receiver Operating Characteristic (ROC) curves and Precision-Recall curves. ROC curves are appropriate when the numbers of observations of the two classes are balanced, whereas precision-recall curves are more appropriate for imbalanced datasets. One appealing feature of ROC curves is to facilitate the search of the optimal cut-point(s). The lack of this feature for Precision-Recall curves limits their applications in other disciplines outside the information retrieval community. To fill the gap, several methods were proposed to determine the best break-point(s) based on Precision-Recall curves. The effectiveness of the proposed methods will be illustrated through simulation studies and real data applications.

Les deux outils de diagnostic principaux des problèmes de modélisation prédictive à deux classes sont les courbes ROC (efficacité du récepteur) et les courbes de précision-rappel. Les courbes ROC sont adaptées lorsque le nombre d'observations dans chacune des deux classes est équilibré, tandis que les courbes de précision-rappel sont mieux adaptées aux jeux de données déséquilibrés. L'une des caractéristiques intéressantes des courbes ROC est qu'elles facilitent la recherche du/des point(s) de rupture optimal/optimaux. L'absence de cette caractéristique dans les courbes de précision-rappel limite leurs applications aux disciplines autres que la récupération d'informations. Plusieurs méthodes ont été proposées pour combler cette lacune et déterminer le meilleur point de rupture en se fondant sur les courbes de précision-rappel. Nous en illustrons l'efficacité par des études de simulation et des applications sur données réelles.

[Monday May 27/lundi 27 mai, 14:15-14:30]

Junhan Fang (University of Waterloo) , **Grace Yi** (University of Waterloo)

Regularized Matrix-Variate Regression with Misclassification in Binary Responses

Régression matrice-variables régularisée avec erreurs de classification dans des réponses binaires

Matrix-variate logistic regression is useful in facilitating the relationship between the binary response and complex-featured matrix-variates commonly arising from medical imaging research. However, such a model is impaired by the presence of response misclasification as well as inactive covariates. It is imperative to account for misclassification effects and select active covatiates when employing matrix-variate logistic regression to handle such data. In this paper, we develop inferential methods based on unbiased estimating functions in combination with penalty functions to deal with the sparsity of the matrix-variate data together with response misclassification. We examine the biases induced from the naive analysis which ignores the misclassification of responses. The proposed methods are justified both theoretically and empirically. We analyze the Breast Cancer Wisconsin (Prognostic) data with the

La régression logistique matrice-variables est utile pour faciliter la relation entre la réponse binaire et les matrices-variables complexes qui surviennent fréquemment dans la recherche en imagerie médicale. Par contre, un tel modèle est affaibli par la présence d'erreurs de classification des réponses ainsi que par des covariables inactives. Il est essentiel de tenir compte des effets des erreurs de classification et de sélectionner des covariables actives lorsque la régression logistique matrice-variables est utilisée pour traiter de telles données. Dans cet exposé, nous développons des méthodes d'inférence fondées sur des fonctions d'estimation non biaisées en combinaison avec des fonctions de pénalité pour traiter l'éparsité des données de la matrice-variables avec les erreurs de classification des réponses. Nous examinons le biais causé par l'analyse naïve qui ignore les erreurs de classification des réponses. Les méthodes proposées sont justifiées à la fois théoriquement et empiriquement. Nous analysons les données du Breast Cancer Wisconsin (Prognostic) avec les méthodes pro-

proposed methods.

posées.

[Monday May 27/lundi 27 mai, 14:30-14:45]

Ali Karimnezhad (University of Ottawa) , **Pearl Campbell** (Ottawa Hospital Research Institute) , **Bryan Lo** (University of Ottawa) , **David J. Stewart** (University of Ottawa) , **Theodore J. Perkins** (University of Ottawa)

An Empirical Bayes Variant Calling Algorithm Designed for Next-Generation Sequencing Data Analysis

Un algorithme bayésien empirique d'identification de variantes pour analyse de données de séquençage de nouvelle génération

The rapid rise of DNA sequencing technologies has led to new statistical problems in identifying genuine DNA mutations (or variants) from noisy sequencing data. Many analysis methods and software have been introduced, but they are mainly based on conventional Bayesian or frequentist methods. In this work, we deal with the task of variant calling as a multiple hypothesis testing procedure. Our proposed procedure utilizes the empirical Bayes principle as well as an internal recursive algorithm which optimally learns parameters of the underlying model with no prior information. We compare the performance of our proposed algorithm against four popular variant callers from the literature using a variety metrics, including sensitivity and precision, on a set of four replicates of a synthetic DNA dataset. We find that the proposed variant caller reliably detects mutations and performs very well compared to existing variant callers.

La croissance rapide des technologies de séquençage d'ADN a mené à de nouveaux problèmes statistiques dans l'identification des véritables mutations d'ADN (ou variantes) dans des données bruitées de séquençage. Plusieurs méthodes et logiciels d'analyse ont été présentés mais ils se basent principalement sur des méthodes conventionnelles bayésiennes ou fréquentistes. Dans cet exposé, nous traitons de l'identification de variantes comme une procédure de test d'hypothèses multiples. Notre procédure utilise le principe empirique bayésien ainsi qu'un algorithme interne récursif qui apprend de façon optimale les paramètres du modèle sous-jacent sans information préalable. Nous comparons la performance de notre algorithme avec quatre identificateurs de variantes populaires de la littérature à l'aide d'une variété de mesures, telles que la sensibilité et la précision, sur un ensemble de quatre copies d'ADN synthétique. Nous constatons que l'identificateur de variantes proposé détecte de manière fiable les mutations et performe bien comparativement aux identificateurs de variantes existants.

Recent Advances in Clinical Trials and Experimental Design and Inference
Percées récentes dans les essais cliniques et en conception et inférence expérimentales

Chair/Président: Judy-Anne W. Chapman

Room/Salle: 119 (SA)

Abstract/Résumé

[Monday May 27/lundi 27 mai, 13:30-13:45]

Su Hwan Kim (University of Alberta) , **Keumhee Chough Carriere** (University of Alberta)

Optimal Crossover Designs with Baselines and Proportional Carryover Effects

Plans d'étude croisés optimaux avec effets de report proportionnels et de base

Crossover designs allow researchers to use within-subject differences to estimate the effect of treatments being tested. These designs gain advantages over parallel designs in terms of estimation efficiency and sample size as the within-subject variability is in general smaller than between-subject variability in repeated measures data. Furthermore, introducing baseline measurements in the analysis is known to improve statistical efficiency. In crossover trials, the effect of treatment administered in the previous period may remain on the current period and these carryover effects are highly correlated with direct treatment effects and can be represented as a remaining proportion of direct treatment effect. This proportional approach allows parameter reduction and therefore can be useful in crossover designs where the number of subjects is small. We construct optimal crossover designs in the presence of baseline outcomes and show the optimality in proportional models.

Les plans d'étude croisés permettent aux chercheurs d'utiliser les différences à l'intérieurs des sujets pour estimer l'effet des traitements testés. Ces plans d'étude ont l'avantage sur les plans parallèles en termes d'efficacité d'estimation et de taille d'échantillon étant donnée que la variabilité propre aux sujets est en général plus petite que la variabilité entre les sujets dans des données de mesures répétées. De plus, l'introduction des mesures de base dans l'analyse est reconnue pour améliorer l'efficacité statistique. Dans des études croisées, l'effet du traitement administré dans la période précédente peut subsister dans la période actuelle et ces effets de report sont hautement corrélés avec les effets directs du traitement et peuvent être représentés comme une proportion restante des effets directs du traitement. Cette approche proportionnelle permet la réduction des paramètres et peut donc être utile dans les concepts croisés où le nombre de sujets est petit. Nous élaborons des concepts croisés optimaux en présence de résultats de base et nous démontrons l'optimalité des modèles proportionnels.

[Monday May 27/lundi 27 mai, 13:45-14:00]

Xinyi Ge (Queen's University) , **Yingwei Peng** (Queen's University) , **Dongsheng Tu** (NCIC Clinical Trials; Queen's University)

A single-Index Threshold Linear Mixed Model for Identification of Treatment-Sensitive Subsets in a Clinical Trial Based on Longitudinal Outcomes and a Continuous Covariate

Un modèle linéaire mixte de seuil à indice unique pour l'identification des sous-ensembles sensibles à un traitement dans un essai clinique fondé sur des variables longitudinales et une covariable continue

Identification of a subset of patients who may be sensitive to a specific treatment is an important problem in clinical trials. In this paper, we consider the case where the treatment effect is measured by longitudinal outcomes, such as quality of life scores assessed over the duration of the clinical trial, and the subset is determined by a continuous covariate, such as a biomarker. A single-index threshold linear mixed model is introduced, and a smoothing maximum likelihood method is proposed to obtain the estimation of the parameters in

L'identification d'un sous-ensemble de patients qui sont susceptibles d'être sensibles à un traitement spécifique est un problème important dans les essais cliniques. Dans cet exposé, nous examinons le cas où l'effet du traitement est mesuré par des variables longitudinales, telles que la qualité de vie évaluée durant toute la durée de l'essai clinique, et le sous-ensemble est déterminé par une covariable continue, telle qu'un biomarqueur. Un modèle linéaire mixte de seuil à indice unique est présenté et une méthode de lissage par maximum de vraisemblance est proposée pour estimer les paramètres du modèle. L'approche Broyden-Fletcher-Goldfarb-

the model. Broyden-Fletcher-Goldfarb-Shanno (BFGS) is employed to maximize the proposed smoothing likelihood function. The proposed procedure is evaluated through simulation studies and application to the analysis of data from a randomized clinical trial on patients with advanced colorectal cancer.

Shanno (BFGS) est utilisée pour maximiser la fonction de lissage par vraisemblance proposée. La performance de la méthode est évaluée par l'entremise d'études de simulations et par son application à l'analyse de données d'un essai clinique aléatoire sur des patients souffrant d'un cancer colorectal avancé.

[Monday May 27/lundi 27 mai, 14:00-14:15]

Keyue Ding (Queen's University)

Group Sequential Test for Nested Multiple Populations According to Biomarker Expression Levels

Test séquentiel groupé pour populations multiples emboîtées selon les niveaux d'expression des biomarqueurs

A cancer biomarker refers to a substance or process that is indicative of the presence of cancer in the body. Usually, such biomarkers can be assayed in bio-samples from patients. Targeted therapies are developed to target the cancer biomarkers, and biological rationale assumes that targeted therapy is more effective in patients with higher levels of the biomarker expression. Clinical trials are conducted to evaluate treatment effects in multiple nested populations according to the levels of the biomarker expression. In this paper, we proposed weighted group sequential Bonferroni tests with closed testing procedures to control the type I error, while taking account of the correlation among the test statistic across the test populations to enhance statistical power. As an application, we demonstrate the design with a cancer clinical trial conducted by the Canadian Cancer Trial Group.

Un biomarqueur du cancer est une substance ou processus qui indique la présence d'un cancer dans le corps. Habituellement, de tels biomarqueurs peuvent être évalués dans les échantillons biologiques des patients. Des traitements ciblés sont développés pour cibler les biomarqueurs du cancer et le raisonnement biologique suppose que le traitement ciblé est plus efficace chez les patients avec des niveaux plus élevés de l'expression des biomarqueurs. Des essais cliniques sont menés pour évaluer les effets du traitement dans des populations multiples emboîtées selon les niveaux d'expression des biomarqueurs. Dans cet exposé, nous présentons des tests Bonferroni séquentiels groupés pondérés avec des procédures de test fermées pour contrôler l'erreur de type I tout en tenant compte de la corrélation entre la statistique de test à travers les populations testées pour augmenter la puissance statistique. Comme application, nous démonterons ce concept avec des essais cliniques sur le cancer menés par le Canadian Cancer Trial Group.

[Monday May 27/lundi 27 mai, 14:15-14:30]

Eric Sanders (University of British Columbia) , **Paul Gustafson** (University of British Columbia) , **Mohammad Ehsanul Karim** (University of British Columbia)

Incorporating Partial Adherence into the Principal Stratification Analysis Framework

Intégration de l'adhésion partielle dans le cadre de l'analyse de stratification principale

Participants in pragmatic clinical trials often partially adhere to treatment. Simple statistical analyses of binary adherence (receiving either full or no treatment) introduce biases in the presence of partial adherence. We developed a framework which expands the principal stratification approach to allow partial adherers to have their own principal stratum and treatment level. We derived consistent estimates for bounds on population values of interest. A Monte Carlo posterior sampling method is derived which is computationally faster than Markov Chain Monte Carlo sampling, with confirmed equivalent results. Simulations indicate that the two methods agree with each other and are superior in most cases to the biased estimators created through standard

Les participants aux essais cliniques pragmatiques ont souvent tendance à adhérer partiellement au traitement. Des analyses statistiques simples de l'adhésion binaire (c'est-à-dire subir un traitement complet ou aucun traitement) présentent des biais en présence de l'adhésion partielle. Nous avons donc conçu un cadre qui élargit l'approche de stratification principale pour permettre aux adhérents partiels d'avoir une strate principale et un niveau de traitement qui leur sont propres. Nous avons aussi dérivé des estimations convergentes des limites relatives aux valeurs d'intérêt de la population. Une méthode d'échantillonnage a posteriori de Monte Carlo est dérivée et donc plus rapide sur le plan calculatoire que l'échantillonnage par la méthode de Monte Carlo par chaîne de Markov, ce qui est confirmé par des résultats équivalents. Des simulations indiquent que les méthodes s'accordent entre elles et

principal stratification. The results suggest that these new methods may lead to increased accuracy of inference in settings where study participants only partially adhere to assigned treatment.

sont dans la plupart des cas supérieures par rapport aux estimateurs biaisés produits à partir d'une stratification principale standard. Les résultats semblent indiquer que ces nouvelles méthodes pourront améliorer la précision de l'inférence dans le cadre d'études où les participants n'adhèrent que partiellement à un traitement donné.

[Monday May 27/lundi 27 mai, 14:30-14:45]

Gurbakhshash Singh (Central Connecticut State University) , **Mark Lowerison** (University of Calgary) , **Ayoola Ademola** (University of Calgary) , **Bijoy K. Menon** (University of Calgary) , **Michael D. Hill** (University of Calgary) , **Tolulope Sajobi** (University of Calgary)

On Covariate Adaptive Randomization in Clinical Trials

De la randomisation adaptée aux covariables dans les essais cliniques

Covariate adaptive randomization (CAR) methods are commonly used to achieve covariate balance across treatment groups in trials with important baseline prognostic covariates. This study compares the performance of CAR methods and block randomization methods (BRM) in terms of covariate imbalance, treatment imbalance and allocation predictability. Monte Carlo methods were used to compare Fair Coin (FC), permuted block, Efron's biased coin, Simon Pocock minimization (MIN), and minimal sufficient balance (MSB) randomization. Simulation conditions include distribution of covariates, sample size, degree of biased coin, and number of covariates.CAR methods had about 11% lower chance of covariate imbalance than the other methods. The FC and BRM had lower treatment allocation imbalance. MIN and BRM exhibited on average 5-25% higher allocation predictability than MSB and FC methods. Recommendations for guiding the choice of an appropriate randomization method for a trial will be discussed.

Les méthodes de randomisation adaptée aux covariables (RAC) sont couramment utilisées pour obtenir un équilibre des covariables dans l'ensemble des groupes de traitement au cours d'essais avec d'importantes covariables prognostiques de base. Cette étude compare la performances des méthodes RAC et des méthodes d'échantillonnage aléatoire par blocs (MEAB) sur le plan du déséquilibre des covariables, du déséquilibre du traitement et de la prévisibilité de l'attribution. Des méthodes de Monte-Carlo ont été utilisées pour comparer divers types de randomisation : pièce de monnaie (PM), blocs de permutation, pièces biaisées d'Efron, minimisation Simon Pocock (MIN) et équilibre suffisant minimal (ESM). Les critères de simulation comprennent la distribution des covariables, la taille de l'échantillon, le degré de biais d'une pièce et le nombre de covariables. Le risque de déséquilibre des covariables avec les méthodes RAC était inférieur de 11 % à celui d'autres méthodes. Le déséquilibre de l'attribution de traitement était plus faible avec les méthodes PM et MEAB. En moyenne, la prévisibilité de l'attribution avec les méthodes MIN et MEAB était de 5 % à 25 % plus élevée qu'avec les méthodes ESM et PM. Nous discuterons des recommandations pour orienter le choix d'une méthode de randomisation appropriée pour un essai clinique.

[Monday May 27/lundi 27 mai, 14:45-15:00]

Anthony Greco (Brock University) , **Xiaojian Xu** (Brock University)

Active Learning and Optimal Experimental Design

Apprentissage actif et conception expérimentale optimale

Active learning is a useful learning process to accurately classify data, in that a small number of training datasets can be used if the learner is able to select the data in which it learns from. The design of the training data can be determined to minimize errors caused from a possibly misspecified model. We review past literature on optimal and robust active learning, and propose new methods for robust active learning in a linear regression

L'apprentissage actif est un processus d'apprentissage utile pour classer des données correctement, en ce sens qu'un petit nombre d'ensembles de données d'apprentissage peut être utilisé si l'apprenant est capable de sélectionner les données à partir desquelles il apprend. La conception des données d'apprentissage peut être définie pour minimiser les erreurs causées par un modèle mal spécifié. Nous examinons la littérature sur l'apprentissage actif robuste et optimal et nous proposons de nouvelles méthodes pour un

setting. Applying methods from design theory, analytical forms for robust designs are given. Monte Carlo simulations show improved efficiencies for the proposed methods relative to other active learning methods. Some practical examples are used to demonstrate the proposed methods. Comparisons and recommendations among those proposed and existing methods are also presented.

apprentissage actif robuste dans un cadre de régression linéaire. Des formes analytiques pour concepts robustes sont présentées en utilisant des méthodes de la théorie de la conception. Des simulations Monte Carlo démontrent une efficacité améliorée pour les méthodes proposées comparativement à d'autres méthodes d'apprentissage actif. Des exemples concrets sont utilisés pour démontrer les méthodes proposées. De plus, des comparaisons et recommandations entre ces méthodes proposées et existantes sont présentées.

Missing Data Methods and Applications
Données manquantes : méthodes et applications

Chair/Président: Shakhawat Hossain

Room/Salle: 109 (SS)

Abstract/Résumé

[Monday May 27/lundi 27 mai, 13:30-13:45]

Yang YZ Zhao (University of Regina) , **Meng Liu** (University of Regina)

Consistent Estimation in Multiple Imputation for Regression Models with Missing Data

Estimation convergente sous imputation multiple pour les modèles de régression avec données manquantes

Multiple imputation (MI) is widely applied in statistical analysis with missing data. Consistent estimation in MI depends on the imputation models. We develop a new diagnostic method for comparing MI models and testing the consistency of MI estimates. The new diagnostic method is efficient as it uses all the observed data and is not affected by the percentage of missing values. In case when the MI estimates are not consistent, we propose a method to compute consistent estimates of regression parameters. The new method is a combination of MI and a unified approach (Chen and Chen 2000) that uses a parametric working model to extract information from imputed data. The consistency of the new estimator depends neither on the MI model nor on the parametric working regression model. We examine the finite sample performance of the proposed methods in simulation studies and illustrate the methods in a study of predicting significant coronary disease and associated factors.

L'imputation multiple est largement utilisée dans l'analyse statistique en présence des données manquantes. L'estimation convergente sous imputation multiple dépend des modèles d'imputation. Nous élaborons une nouvelle méthode de diagnostic pour comparer les modèles d'imputation multiple et vérifier la convergence des estimations issues de l'imputation multiple. La nouvelle méthode de diagnostic est efficace, car elle utilise toutes les données observées et n'est pas touchée par le pourcentage de valeurs manquantes. Dans le cas où les estimations issues de l'imputation multiple ne sont pas convergentes, nous proposons une méthode pour calculer les estimations convergentes des paramètres de régression. La nouvelle méthode est une combinaison de l'imputation multiple et d'une approche unifiée (Chen et Chen, 2000) qui utilise un modèle paramétrique pour extraire l'information des données imputées. La convergence du nouvel estimateur ne dépend ni du modèle d'imputation multiple ni du modèle paramétrique de régression de travail. Nous examinons l'efficacité des méthodes proposées sous échantillons finis à l'aide d'études de simulation et illustrons les méthodes utilisées dans une étude de prédiction des maladies coronariennes significatives et des facteurs associés.

[Monday May 27/lundi 27 mai, 13:45-14:00]

David Luke Thiessen (University of Regina) , **Yang Zhao** (University of Regina)

Non-Monotone Missing Covariates in Cox Regression

Covariables manquantes non-monotones dans la régression Cox

The Cox regression model is one of the most widely used models in survival analysis and much research has been done to overcome difficulties in applying it. One of the most common of these issues is missing data. Missingness appears in survival data in many ways, but in particular we consider when covariates are not observed either by design or by chance. Early work focused on dealing with monotone missing patterns and when data are missing completely at random. More recent work has focused on weakening those assumptions and developing methods for more general situations. We re-

Le modèle de régression Cox est un des modèles les plus utilisés dans l'analyse de survie et beaucoup d'études ont été menées pour surmonter les difficultés lors de son application. L'une de ces difficultés la plus répandue concerne les données manquantes. Les données manquantes apparaissent dans les données de survie de plusieurs façons mais nous examinons en particulier le cas où les covariables ne sont pas observées par conception ou par hasard. Des travaux antérieurs se sont concentrés sur les schémas monotones de manque et sur le cas où les données sont manquantes de façon totalement aléatoire. Les travaux plus récents sont axés sur l'assouplissement de ces hypothèses et sur le développement de

view methods for fitting the Cox regression model with non-monotone missing data under the missing at random assumption. We demonstrate these methods with a real data set and compare their efficiency with simulated data.

méthodes pour des situations plus générales. Nous examinons des méthodes pour ajuster le modèle de régression Cox avec données manquantes non-monotones sous l'hypothèse de données manquantes de façon aléatoire. Nous démontrons ces méthodes avec un ensemble de données réelles et nous comparons leur efficacité avec des données simulées.

[Monday May 27/lundi 27 mai, 14:00-14:15]

Shixiao Zhang (University of Waterloo) , **Peisong Han** (University of Michigan) , **Changbao Wu** (University of Waterloo)

A Multiply Robust Mann-Whitney Test for Non-Randomized Pretest-Posttest Studies with Missing Data

Un test Mann-Whitney multi-robuste pour études non randomisées menées avant et après l'essai avec des données manquantes

Pretest-posttest trials are widely used to study the treatment effect of an intervention. While the average treatment effect is usually of primary interest, testing the equivalence of the marginal distributions of the potential outcomes between the two intervention groups is also of great research importance. We propose an empirical likelihood (EL) based Mann-Whitney test in non-randomized pretest-posttest studies where the posttest outcomes are also subject to missingness. The proposed test is multiply robust, in the sense that multiple working models for the propensity score of treatment assignment, the missingness probability and the outcome regressions can be accommodated, and the test maintains the correct type I error if a certain combination of those multiple working models are correctly specified. The proposed method is a major extension to the work of Zhang, Han and Wu (2018).

Des études avant et après l'essai sont souvent utilisées pour analyser l'effet thérapeutique d'une intervention. Bien que la moyenne de l'effet du traitement soit l'objectif principal, tester l'équivalence des distributions marginales des résultats potentiels entre les deux groupes d'intervention est aussi très important pour la recherche. Nous proposons une vraisemblance empirique (VE) fondée sur le test Mann-Whitney dans des études non randomisées menées avant et après l'essai où les résultats après l'essai sont aussi soumis à des données manquantes. Le test proposé est multi-robuste, en ce sens qu'il peut tenir compte de plusieurs modèles de travail pour le score de propension relatif à l'attribution du traitement, de la probabilité de données manquantes et de la régression des résultats, et le test préserve une erreur de type I appropriée si une certaine combinaison de ces modèles de travail est correctement spécifiée. La méthode proposée est une importante extension du travail de Zhang, Han et Wu (2018).

[Monday May 27/lundi 27 mai, 14:15-14:30]

Sophie Castel (Trent University) , **Melissa Van Bussel** (Trent University) , **Wesley Burr** (Trent University)

Imputation of Missing Values in Time Series Data

Imputation de valeurs manquantes dans des séries chronologiques

The spectrum of a given time series is a characteristic function describing its frequency properties. Estimators of the spectrum often require the data to be contiguous for optimal performance. This poses a fundamental challenge when considering real-world data (of strong scientific interest) that is often plagued by missing values, and/or irregularly recorded measurements. Imputation (often referred to as interpolation in this context) seeks to repair the original time series, inserting estimated values for the missing quantities. Recent work has led to a number of algorithms that have proven successful for the interpolation of large gaps of missing data, but largely are applicable for use on stationary time series only. In this talk we will explore the performance

Le spectre d'une série chronologique est une fonction caractéristique qui décrit ses propriétés fréquentielles. Les estimateurs du spectre nécessitent fréquemment des données contiguës pour obtenir une performance optimale. Ceci représente un défi fondamental lorsque des données réelles sont étudiées (d'un grand intérêt scientifique). En effet, ces données sont souvent confrontées à des valeurs manquantes et/ou à des mesures enregistrées de manière irrégulière. L'imputation (communément appelé interpolation dans ce contexte) vise à restaurer la série chronologique originale en remplaçant les quantités manquantes par des valeurs estimées. Des travaux récents ont mené à un certain nombre d'algorithmes qui se sont avérés efficaces pour l'interpolation de grands écarts dans des données manquantes. Toutefois, ces algorithmes ne sont en général appropriés à des fins d'utilisation que dans des

of a recently proposed imputation algorithm, the Hybrid Wiener Interpolator, and compare it to several other modern approaches from the signal processing, time series, and general statistics communities.

séries chronologiques stationnaires. Dans cet exposé, nous examinerons la performance d'un algorithme récemment présenté, l'interpolateur hybride Wiener, et nous le comparerons à plusieurs autres approches modernes du traitement des signaux, des séries chronologiques et des communautés statistiques en général.

[Monday May 27/lundi 27 mai, 14:30-14:45]

Md. Shaddam Hossain Bagmar (University of Calgary) , **Hua Shen** (University of Calgary)

Causal Inference with Missingness in Confounders

Inférence causale accompagnée de facteurs de confusion comportant des lacunes

Causal inference is the process of uncovering causal connection between the effect variable and disease outcome in public health and medical research. Confounders that influence both the effect variable and outcome need to be accounted for when obtaining the causal effect in observational studies. In addition, missing data often arise in the data collection procedure, working with complete cases often results in biased estimates. We consider the estimation of causal effect in the presence of missingness in the confounders under the missing at random assumption. We investigate how different estimation methods perform when applying complete-case analysis or multiple imputation. We then propose an expectation-maximization (EM) algorithm to estimate the expected values of the missing confounder and utilize weighting approach in the effect estimation to obtain robust estimators. Simulation studies are conducted to show that it provides more consistent estimates of the causal effect.

En recherche médicale et en santé publique, l'inférence causale désigne le processus par lequel on peut établir une relation de causalité entre l'effet d'une variable et l'issue d'une maladie. Il faut tenir compte des facteurs de confusion qui influencent à la fois les effets de la variable et le résultat pour obtenir l'effet causal dans des études observationnelles. De plus, il n'est pas rare que l'on remarque des données manquantes lors de la procédure de collecte de données, et travailler sur des cas complets entraîne souvent des estimations biaisées. Nous examinons l'estimation de l'effet causal en présence de lacunes dans les facteurs de confusion selon l'hypothèse de valeur manquante aléatoire. Nous examinons comment différentes méthodes d'estimation fonctionnent lorsque l'on applique une analyse de cas complète ou une imputation multiple. Nous proposons ensuite un algorithme espérance-maximisation (EM) pour estimer les valeurs attendues du facteur de confusion manquant et nous servir d'une approche de pondération dans l'estimation de l'effet dans le but d'obtenir des estimateurs robustes. Des études de simulation sont menées pour démontrer que la convergence des estimations de l'effet causal obtenu par l'algorithme est supérieure.

[Monday May 27/lundi 27 mai, 14:45-15:00]

Menglu Che (University of Waterloo) , **Jerry Lawless** (University of Waterloo) , **Peisong Han** (University of Michigan)

Empirical and Conditional Likelihoods for Two-Phase Studies with Response-Dependent Samples

Vraisemblances empiriques et conditionnelles pour les études biphasées avec échantillonnage dépendant de la réponse

Two-phase, response-dependent sampling is often used in applications involving expensive covariate measurements. In phase 1, response and inexpensive covariates are measured in a sample, and in a response-dependent phase 2 sub-sample, the expensive covariate is measured. In a missing data framework, unobserved covariate values are missing-at-random (MAR). Conditional maximum likelihood (CML) is an attractive approach for MAR covariates as it avoids modeling the covariate distribution. Scott and Wild (2011) gave a semiparametric efficient estimator, referred to as the SW estimator, for regression with binary response and categorical covariates. We consider a general regression model and

L'échantillonnage dépendant de la réponse dans les études biphasées est souvent utilisé dans des applications faisant appel à des mesures de covariables coûteuses. Dans la phase 1, on obtient les mesures de la réponse et des covariables peu coûteuses d'un échantillon, et dans un sous-échantillon de phase 2 dépendant de le réponse. Dans un cadre de données manquantes, les valeurs des covariables inobservées sont manquantes aléatoirement (MA). Le maximum de vraisemblance conditionnel (MVC) est une approche attrayante pour les covariables MA, car il évite la modélisation de la loi des covariables. Scott et Wild (2011) ont présenté un estimateur semi-paramétrique efficace, appelé estimateur SW, pour une régression à réponse binaire et covariables catégorielles. Nous prenons un modèle de régression généralisé pour montrer qu'un

show that an estimator of the same form as SW has identical efficiency to two empirical likelihood estimators, and that they dominate the CML estimator. Thus the SW estimator is appealing in more general settings, and avoids the sometimes difficult computation of empirical likelihood estimators.

estimateur de même forme que le SW a une efficacité identique à deux estimateurs de vraisemblance empiriques et qu'il domine l'estimateur MVC. D'où l'attrait de l'estimateur SW pour un cadre plus général et pour éviter le calcul parfois difficile des estimateurs de vraisemblance empiriques.

Implementation, Advances and Precision in Mixture Model-Based Classification
Mise en œuvre, progrès et précision en classification fondée sur les modèles de mélange

Chair/Président: Jeffrey L. Andrews

Organizer/Responsable: Brian Franczak

Room/Salle: 113 (SS)

Abstract/Résumé

[Monday May 27/lundi 27 mai, 13:30-14:00]

Antonio Punzo (University of Catania)

On the Use of the Contaminated Normal Distribution in Model-Based Clustering

L'utilisation de la loi normale contaminée dans le regroupement modélisé

The multivariate contaminated normal distribution represents a simple heavy-tailed generalization of the multivariate normal distribution to model elliptical contoured scatters in the presence of mild outliers. Once the model is fitted to the available data, a classification of the observations as typical points and (mild) outliers can be automatically obtained. The price to pay for these advantages is two additional parameters, both with a specific and useful interpretation: the proportion of good observations and degree of contamination. In this work we provide a review of the approaches using the contaminated normal distribution in model-based clustering.

La loi normale multivariée contaminée représente une simple généralisation à queue lourde de la loi normale multivariée pour modéliser les diffuseurs elliptiques en présence de faibles valeurs aberrantes . Une fois que le modèle a été ajusté aux données disponibles, on peut obtenir automatiquement une classification des observations, comme les points typiques et les valeurs aberrantes (faibles). La conséquence est qu'on a deux paramètres supplémentaires, tous deux avec une interprétation spécifique et utile : la proportion de bonnes observations et le degré de contamination. Dans ces travaux, nous passons en revue les approches utilisant la loi normale contaminée dans le regroupement modélisé.

[Monday May 27/lundi 27 mai, 14:00-14:30]

Cristina Tortora (San Jose State University) , **Antonio Punzo** (University of Catania)

Advances in Model-Based Clustering and Outlier Detection

Progrès en matière de regroupement fondé sur un modèle et de détection des valeurs aberrantes

Model-based clustering assumes that the data were generated from a convex combination of densities. The choice of the density function is crucial; the multivariate contaminated normal distribution (MCN) was proposed to model datasets characterized by the presence of outliers. The MCN is a two-component Gaussian mixture; one of the components, with a large prior probability, represents the good observations, and the other, with an inflated covariance matrix, represents the outliers. Mixtures of MCN distributions can detect outliers and perform cluster analysis improving the clustering performance when compared to normal. However, the CND uses univariate parameters to model the outliers, i.e., they are the same for all the variables. This is a limit because the outliers may be different in each dimension. To overcome this issue, we propose a multiple scaled contaminated normal distribution with p-dimensional proportion of outliers and degrees of con-

Le regroupement fondé sur un modèle suppose que les données ont été générées à partir d'une combinaison convexe de densités. Le choix de la fonction de densité est crucial ; on a proposé la distribution normale contaminée à plusieurs variables pour modéliser des ensembles de données caractérisés par la présence de valeurs aberrantes. La distribution normale contaminée à plusieurs variables est un mélange gaussien à deux composants : l'un des composants, avec une forte probabilité a priori, représente les bonnes observations, et l'autre, avec une matrice de covariance gonflée, représente les valeurs aberrantes. Les mélanges de distributions normales contaminées à plusieurs variables peuvent détecter les valeurs aberrantes et effectuer des analyses de regroupement, améliorant ainsi l'efficacité du regroupement par rapport à la normale. Cependant, la distribution normale contaminée utilise des paramètres univariés pour modéliser les valeurs aberrantes, c'est-à-dire qu'ils sont les mêmes pour toutes les variables. Il s'agit d'une limite, car les valeurs aberrantes peuvent être différentes dans chaque dimension. Pour résoudre ce problème, nous proposons une distribution

tamination, with p number of variables.

normale contaminée à échelles multiples avec une proportion p dimensionnelle de valeurs aberrantes et de degrés de contamination, ainsi qu'un nombre p de variables.

[Monday May 27/lundi 27 mai, 14:30-15:00]

Hua Shen (University of Calgary)

Mixture Model with Inaccurate Measurements

Modèle de mélange accompagné de mesures imprécises

The mixture model is often used to address heterogeneity in the pooled population when sub-populations are present. It often requires the accurate measurement of the observations though does not require the observed data to identify the sub-population attributions. When observations are measured with error, the naive approach can lead to biased results in the parameter estimation and subgroup identification. In a simulation study we assess the consequences of having inaccurate observations in the finite mixture model and evaluate the performance of the proposed method to handle this complication.

On emploie souvent le modèle de mélange pour aborder l'hétérogénéité dans la population groupée en présence de sous-populations. Il faut habituellement avoir des mesures précises des observations pour cerner les attributions d'une sous-population, toutefois les données observées ne sont pas nécessaires. En mesurant les observations avec des erreurs, une approche naïve peut produire des résultats biaisés dans l'estimation du paramètre et dans l'identification du sous-groupe. Dans le cadre d'une étude de simulation, nous examinons les conséquences que peuvent entraîner des observations imprécises dans le modèle de mélange fini et évaluons la performance de la méthode proposée pour gérer cette complication.

New Directions in Mathematical Finance
Nouvelles orientations en mathématiques financières

Chair/Président: Jean-François Bégin

Organizer/Responsable: Alexandru Badescu

Room/Salle: 144 (SB)

Abstract/Résumé

[Monday May 27/lundi 27 mai, 15:30-16:00]

Cody Hyndman (Concordia University) , **Anastasis Kratsios** (ETH Zurich)

The NEU Meta-Algorithm for Geometric Learning

Le méta-algorithme NEU pour l'apprentissage géométrique

We introduce a meta-algorithm, called non-Euclidean upgrading (NEU), which learns algorithm specific geometries to improve the training and validation set performance of a wide class of learning algorithms. Our approach is based on iteratively performing local reconfigurations of the space in which the data lie. These reconfigurations build universal approximation and universal reconfiguration properties into the new algorithm that is learned. This allows any set of features to be learned by the new algorithm to arbitrary precision. The training and validation set performance of NEU is investigated through implementations predicting the relationship between select stock prices, and finding low-dimensional representations of the German Bond yield curve.

Nous présentons un méta-algorithme appelé valorisation non euclidienne (NEU en anglais), pour l'apprentissage de géométries propres à l'algorithme afin d'améliorer la performance d'un ensemble de formation et de validation d'une vaste catégorie d'algorithmes d'apprentissage. Notre approche se fonde sur des reconfigurations locales de performances itératives de l'espace dans lequel se trouvent les données. Ces reconfigurations produisent des propriétés d'approximation universelle et de reconfiguration universelle pour le nouvel algorithme qui est enseigné. Il est ainsi possible d'apprendre par ce nouvel algorithme un ensemble de caractéristiques de précision arbitraire. Nous examinons la performance d'un ensemble de formation et de validation NEU par des implémentations prédisant la relation entre les prix des actions, tout en découvrant des représentations de faible dimension de la courbe de rendement des obligations allemandes.

[Monday May 27/lundi 27 mai, 16:00-16:30]

Mark Reesor (Wilfrid Laurier) , **Xinghua Zhou** (Morgan Stanley)

Calibration of Capital Structure Models

Calibration de modèles de structure du capital

Capital structure models treat equity as a call option on firm value and hence traded equity options are viewed as compound options (CO) on firm value. Using the CO interpretation, recent work (Geske et al 2016) has shown that prices of traded equity options depend on a firm's capital structure. This work is done in the Merton framework in which default and liquidation of the firm is allowed only at one specific future date. In our work, we extend the CO analysis to the first-passage time (FPT) framework in which default occurs the first time that firm value breaches a barrier. We derive valuation equations and show that the FPT framework provides greater flexibility in fitting option-implied volatility curves as compared to the Merton framework. As part of the calibration, we obtain market-implied lever-

Les modèles de structure du capital considèrent une action comme une option d'achat de la valeur de l'entreprise, et conséquemment les options sur actions échangées sont considérées comme des options composées (OC) de la valeur de l'entreprise. Au moyen de l'interprétation de l'OC, une étude récente (Geske et coll. 2016) a démontré que les prix des options sur actions échangées varient selon la structure du capital de l'entreprise. Cette étude est réalisée selon le cadre Merton, dans lequel la défaillance et la liquidation d'une entreprise ne sont permises qu'à une date à venir. Dans le cadre de notre étude, nous élargissons l'étendue des analyses des OC à un cadre de temps de premier passage (TPP), dans lequel la défaillance se produit la première fois que la valeur de l'entreprise franchit un seuil. Nous dérivons les équations d'évaluation et démontrons que le cadre TPP est bien plus polyvalent pour ajuster les courbes de volatilité implicites sur option par rapport au

age and firm volatility, which can be used in other corporate finance applications.

cadre Merton. Grâce à la calibration, on obtient un levier financier implicite au marché et une volatilité d'entreprise, que l'on peut appliquer à d'autres financements des entreprises.

[Monday May 27/lundi 27 mai, 16:30-17:00]

Jinniao Qiu (University of Calgary)

Viscosity Solutions of Stochastic Hamilton-Jacobi-Bellman Equations and their Applications in Mathematical Finance

Solutions de viscosité des équations stochastiques de Hamilton-Jacobi-Bellman et applications en mathématiques financières

We shall talk about the fully nonlinear stochastic Hamilton-Jacobi-Bellman (HJB) equation for the optimal stochastic control problem of stochastic differential equations with random coefficients. The notion of viscosity solution is introduced, and the value function of the optimal stochastic control problem is proved to be the unique viscosity solution of the associated stochastic HJB equation. Applications in mathematical finance will be discussed as well.

Nous parlerons de l'équation stochastique non linéaire de Hamilton-Jacobi-Bellman (HJB) pour le problème de contrôle optimal stochastique des équations différentielles stochastiques à coefficients aléatoires. Nous introduirons la notion de solution de viscosité et prouverons que la fonction de valeur du problème de contrôle optimal stochastique est la solution de viscosité unique de l'équation stochastique de HJB associée. Nous discuterons également d'applications en mathématiques financières.

Making Sense of Complex Featured Data with Statistical Methods
Exploitation de données à caractéristiques complexes par méthodes statistiques

Chair/Président: Grace Yi

Organizer/Responsable: Grace Yi

Room/Salle: 102 (ICT)

Abstract/Résumé

[Monday May 27/lundi 27 mai, 15:30-15:50]

Mireille E. Schnitzer (Université de Montréal) , **Lucie Blais** (Université de Montréal) , **Robert Platt** (McGill University) , **Madeleine Durand** (Centre Hospitalier de l'Université de Montréal and the Research Center of Centre Hospitalier de l'Université de Montréal)

Dealing with Time-Varying Eligibility for Exposure Using the Target Trials Approach to Causal Inference

Traitement de l'admissibilité variant dans le temps de l'exposition à l'aide de l'approche des essais ciblés pour l'inférence causale

Improper handling of time-dependent exposures can result in ill-defined causal effects, residual time-dependent confounding, and immortal time bias. Marginal structural models were proposed for time-varying exposures where confounders may be affected by prior treatment. However, typical definitions of exposure effects no longer apply when eligibility for exposure varies over time. We demonstrate how "treatment strategies" in the context of a target trial (Hernán and Robins, 2016) provide a solution using two real examples. The first example involves post-baseline contraindications to Direct Oral Anticoagulants (DOACs) in a study that aims to contrast DOACs with warfarin in patients with atrial fibrillation. The second involves contrasting trimester-specific exposures during pregnancy when some women deliver in the second trimester. For both of these examples, we define effects based on "treatment strategies" and describe alternate implementations of existing causal inference methods.

Un mauvais traitement des expositions dépendant du temps peut entraîner des effets causaux mal définis, une confusion résiduelle en fonction du temps et un biais de temps immortel. Des modèles structurels marginaux ont été proposés pour les expositions variant dans le temps, dans lesquelles les facteurs de confusion peuvent être touchés par un traitement préalable. Toutefois, les définitions typiques des effets de l'exposition ne s'appliquent plus lorsque l'admissibilité à l'exposition varie avec le temps. Nous montrons comment les ≪ stratégies de traitement ≫ dans le contexte d'un essai ciblé (Hernán et Robins, 2016) apportent une solution au moyen de deux exemples concrets. Le premier exemple porte sur les contre-indications aux anticoagulants oraux directs (AOD) dans une étude visant à comparer les AOD avec la warfarine chez les patients atteints de fibrillation auriculaire. Le deuxième porte sur les expositions contrastantes spécifiques au trimestre pendant la grossesse, lorsque certaines femmes accouchent au cours du deuxième trimestre. Dans ces deux exemples, nous définissons les effets en fonction des ≪ stratégies de traitement ≫ et décrivons d'autres applications des méthodes d'inférence causale existantes.

[Monday May 27/lundi 27 mai, 15:50-16:10]

Wenqing He (University of Western Ontario) , **Grace Yi** (University of Waterloo) , **Junhan Fang** (University of Waterloo)

Prediction for Error-Contaminated Image Data with an Application of the Prostate Cancer Imaging Study

Prédiction des données d'images contaminées par des erreurs à l'aide d'une application de l'étude d'imagerie du cancer de la prostate

Prostate cancer is the most commonly diagnosed cancer and the third highest cause of cancer-related mortality in men. Treatment for prostate cancer is quite successful, with about a 95% 5 year survival rate for patients with cancer stage below 3. However, this success hinges on an early stage diagnosis and confirmation. While it is imperative to build a powerful predictive model for

Le cancer de la prostate est le cancer le plus souvent diagnostiqué et la troisième cause de mortalité liée au cancer chez les hommes. Le traitement du cancer de la prostate est très efficace, avec un taux de survie d'environ 95 % à cinq ans chez les patients dont le stade de cancer est inférieur à 3, mais ce succès dépend d'un diagnostic précoce et d'une confirmation. Bien qu'il soit impératif d'élaborer un modèle prédictif puissant pour les

prostate cancer imaging data, existing methods cannot be applied due to their inadequacy of accommodating the unique features of prostate cancer imaging data. In particular, data imbalance, spatial correlation, and outcome misclassification present great challenges in data analysis. In this talk, I will discuss various statistical approaches to building an effective prediction model. I will examine the data from multiple angles with their features accommodated differently.

données d'imagerie du cancer de la prostate, les méthodes actuelles ne peuvent pas être appliquées en raison de leur incapacité à prendre en compte les caractéristiques uniques des données d'imagerie du cancer de la prostate. Notamment, le déséquilibre des données, la corrélation spatiale et la mauvaise classification des résultats posent de gros problèmes dans l'analyse des données. Dans cet exposé, j'aborderai diverses approches statistiques pour créer un modèle de prédiction efficace. J'examinerai les données et leurs caractéristiques sous de multiples angles et traiterai leurs caractéristiques.

[Monday May 27/lundi 27 mai, 16:10-16:30]

Trevor Thomson (Simon Fraser University) , **John Braun** (University of British Columbia - Okanagan) , **Joan Hu** (Simon Fraser University)

On Time to First Spot Fire

Délai avant premier feu disséminé

Protecting communities from wildfires is of primary concern to wildfire management agencies. Under certain environmental and wildfire conditions, a burning ember can travel beyond a fuel break, such as a river or road, and produce a new fire, known as a spot fire. This phenomenon allows a wildfire to overcome barriers, which can put a strain on resources and put communities at risk. In this talk, we formulate the process of spot fire development and derive the distribution of the time to the first spot fire occurring beyond a fuel break. A simulator is developed in the framework to generate burning embers from an active wildfire that may result in a spot fire. With the generated data, we demonstrate how to estimate the rate of developing spot fires and identify significant covariates based on data in two practical formats. This is a joint work with John Braun (UBC-O) et Joan Hu (SFU).

La protection des communautés contre les feux de forêt est une préoccupation majeure des organismes de gestion des feux de forêt. Dans certaines conditions environnementales et d'incendie, une braise enflammée peut franchir un coupe-feu (rivière ou route) et produire un nouvel incendie, appelé feu disséminé. Ce phénomène permet aux feux de forêt de franchir les barrières, ce qui peut menacer les ressources et mettre en danger les communautés. Dans cette présentation, nous formulons le processus de développement d'un feu disséminé et dérivons la distribution du délai avant premier feu disséminé au-delà d'un coupe-feu. Nous développons un simulateur dans le cadre pour générer des braises à partir d'un feu actif qui peuvent donner lieu à des feux disséminés. Avec les données ainsi générées, nous montrons comment estimer la vitesse de développement des feux disséminés et identifions des covariables significatives tirées des données, dans deux formats pratiques. Ces travaux ont été réalisés en collaboration avec John Braun (UBC-O) et Joan Hu (SFU).

[Monday May 27/lundi 27 mai, 16:30-16:50]

Gabrielle Simoneau (McGill University) , **Erica Moodie** (McGill University) , **Laurent Azoulay** (McGill University) , **Robert Platt** (McGill University)

Estimating Optimal Dynamic Treatment Regimes with Survival Outcomes : An Application to the Treatment of Type 2 Diabetes

Estimation optimale des plans dynamiques de traitements avec les issues de survie : une application sur le diabète de type 2

The statistical study of precision medicine is concerned with dynamic treatment regimes (DTRs) in which treatment decisions are tailored to evolving patient-level information. An optimal DTR is the sequence of treatment decisions that yields the best expected outcome. Statistical methods for optimal DTRs of observational data are theoretically complex and hardly accessible to researchers, especially when the outcome is survival time subject to right censoring. We propose a doubly-robust

Les plans dynamiques de traitements (PDT) sont l'étude statistique de la médecine de précision où les décisions de traitements sont adaptées aux caractéristiques des patients. Un PDT optimal est la séquence de décisions qui mène à la meilleure réponse espérée. Les méthodes statistiques pour l'identification d'un PDT optimal à partir de données observationnelles sont théoriquement complexes et difficiles à appliquer, surtout lorsque la réponse est un temps de survie sujet à la censure. Nous proposons une méthode doublement robuste et accessible, un modèle de survie

method, called dynamic weighted survival modeling, for estimating optimal DTRs for such endpoints. Our method is available in the DTRreg R package, thus facilitating its application by researchers. We illustrate our novel method with an application to the treatment of type 2 diabetes, estimating a DTR for second- and third-line diabetes therapies that maximizes the time until the occurrence of a major cardiovascular event in patients for whom metformin, the preferred first-line treatment, has failed.

dynamique et pondérée, pour estimer un PDT optimal avec des données de survie. Notre méthode est implémentée dans le paquet R DTRreg. Nous illustrons notre méthode avec une application sur le diabète de type 2 en estimant un PDT pour les traitements de seconde et tierce lignes qui maximisent le temps de survie jusqu'à l'occurrence d'un évènement cardiovasculaire néfaste parmi les patients pour lesquels metformin, le traitement de première ligne recommandé, n'a pas fonctionné.

Recent Developments in Statistical Analysis of Nutrition Data
Récentes évolutions en analyse statistique de données nutritionnelles

Chair/Président: Dominique Ibañez

Organizer/Responsable: Dominique Ibañez

Room/Salle: 119 (SA)

Abstract/Résumé

[Monday May 27/lundi 27 mai, 15:30-16:00]

Hassan Vatanparast , **Rashmi Prakash Patil** (College of Pharmacy and Nutrition, University of Saskatchewan) , **Naorin Islam** (College of Pharmacy and Nutrition, University of Saskatchewan) , **Seyed Hamzeh Hosseini** (College of Pharmacy and Nutrition, University of Saskatchewan) , **Zeinab Hosseini** (College of Pharmacy and Nutrition, University of Saskatchewan) , **Arash Shamloo** (College of Pharmacy and Nutrition, University of Saskatchewan) , **Pardis Keshavarz** (College of Pharmacy and Nutrition, University of Saskatchewan)

National Nutrition and Health Survey Data Analyses, Challenges and Opportunities

Analyses de données d'enquêtes nationales sur la nutrition et la santé, défis et opportunités

Canada has the advantage of nationally representative nutrition and health survey data through Canadian Community Health Surveys (CCHS) and Canadian Health Measures Surveys (CHMS). However, challenges exist in statistical approaches for data analyses. Aim: To provide an overview of the opportunities and challenges that exist in national nutrition and health survey data analyses requiring statical expertise. Method: Nutritional health of populations is assessed at nutrient, food and food group, and dietary pattern levels (priori and posteriori). Providing examples from CCHS 2004 and 2015 and CHMS, we discuss the existing approaches, advantages, disadvantages, and the new horizons on novel statistical procedures needed in the field of nutritional epidemiology. Conclusion: In the field of nutritional epidemiology, when analyzing national nutrition and health survey data, there is a great opportunity for statisticians to address the substantial need for novel statistical approaches.

Le Canada a l'avantage de disposer de données d'enquêtes représentatives sur la nutrition et la santé au niveau national par le biais des Enquêtes sur la santé dans les collectivités canadiennes (ESCC) et des Enquêtes canadiennes sur les mesures de la santé (ECMS). Cependant, les approches statistiques pour l'analyse des données posent des problèmes. Objectif : Donner un aperçu des opportunités et des défis qui existent dans les analyses de données d'enquêtes nationales sur la nutrition et la santé nécessitant une expertise statistique. Méthode : La santé nutritionnelle des populations est évaluée au niveau des éléments nutritifs, des aliments et des groupes d'aliments et des habitudes alimentaires (a priori et a posteriori). En fournissant des exemples tirés de l'ESCC 2004 et 2015 et de l'ECMS, nous discutons des approches existantes, des avantages, des inconvénients et des nouveaux horizons sur les nouvelles procédures statistiques nécessaires dans le domaine de l'épidémiologie nutritionnelle. Conclusion : Dans le domaine de l'épidémiologie nutritionnelle, l'analyse des données d'enquêtes nationales sur la nutrition et la santé offre aux statisticiens une excellente occasion de répondre au besoin important de nouvelles approches statistiques.

[Monday May 27/lundi 27 mai, 16:00-16:30]

Dominique Ibañez (Health Canada) , **Karelyn Davis** (Health Canada) , **Alejandro Gonzalez** (Health Canada) , **Lidia Loukine** (Health Canada) , **Cunye Qiao** (Health Canada) , **Alireza Sadeghpour** (Health Canada) , **Michel Vigneault** (Health Canada) , **Kuan Chiao Wang** (Health Canada)

Early Experience with the National Cancer Institute (NCI) Method for Estimating Usual Intakes Using the Canadian Community Health Survey

Premières expériences avec la méthode d'estimation des apports usuels du National Cancer Institute (NCI) sur l'Enquête sur la santé dans les collectivités canadiennes

Health Canada needs information on the consumption of particular foods or nutrients (usual intake) for evidence-

Santé Canada a besoin d'information probante sur la consommation alimentaire (apport usuel) pour les politiques et programmes

based nutrition policies and programs. This is used to estimate the prevalence of inadequacy or excess for nutrients. The NCI method was investigated to evaluate usual intakes of nutrients using data from the Canadian Community Health Survey – Nutrition. Eating habits vary within and between individuals. The NCI method accounts for such variations to evaluate usual intake distributions. It also permits analysis of usual intakes by covariates and of episodically consumed foods. The NCI method performed well in the Canadian context. Increase in the prevalence of inadequacy was observed from 2004 to 2015 for some nutrients (e.g. calcium) and decreased for others (e.g. sodium). A drawback of the NCI method is that it requires long computational times. The NCI method is an innovative and powerful tool for the evaluation of usual intakes using Canadian nutrition data.

de nutrition. Elle est utilisée pour estimer la prévalence d'insuffisance ou d'excès en nutriments. Les apports usuels basés sur les données de l'Enquête sur la santé dans les collectivités canadiennes – Nutrition ont été évalués par la méthode NCI. Les habitudes alimentaires varient selon les individus et entre eux. La méthode NCI tient compte de ces variations. Elle permet aussi l'analyse par covariables et pour les aliments peu consommés. Cette méthode fonctionne bien dans le contexte canadien. Une augmentation de la prévalence de l'insuffisance a été observée de 2004 à 2015 pour certains nutriments (ex. calcium) et réduite pour d'autres (ex. sodium). Un inconvénient de la méthode NCI est qu'elle nécessite un long temps de calcul. La méthode NCI est un outil novateur et puissant d'évaluation des apports usuels pour les données nutritionnelles canadiennes

[Monday May 27/lundi 27 mai, 16:30-17:00]

Alireza Sadeghpour (Health Canada) , **Karelyn Davis** (Health Canada) , **Nadine Kebbe** (Health Canada) , **Isabelle Rondeau** (Health Canada) , **Michel Vigneault** (Health Canada) , **Dominique Ibañez** (Health Canada)

The Effect of the Food Model Booklet on Reported Foods in the Canadian Community Health Survey (CCHS) – Nutrition

L'effet du livret de modèles de portions (LMP) sur les aliments déclarés dans l'Enquête sur la santé dans les collectivités canadiennes (ESCC) – Nutrition

The 2004 and 2015 CCHS - Nutrition are two national surveys providing detailed information on food consumed by Canadians. Analysis of 2015 data showed decreases of 245 kcal in energy, 109 mg in calcium and 418 mg in sodium over 2004 values. Food model booklets contained dishware images useful in quantifying the amount of food consumed. The 2004 booklet was copied from an existing one from the US. The 2015 version was revised to better reflect Canadian dishware sizes. These revisions were investigated as a potential cause of the observed decreases in reported nutrient intakes between 2004 and 2015. Food amounts reported in 2004 CCHS were adjusted to mimic the revised booklet and then were compared to 2015 values. After adjustments, the booklet revisions accounted for 11% to 37% of the difference in energy, 23% to 58% in calcium and 12% to 34% in sodium. Revisions to the 2015 booklet are not negligible; differences observed between the two surveys should be interpreted with caution.

Les ESCC de 2004 et 2015 fournissent des informations détaillées sur les aliments consommés au Canada. Par rapport à 2004, les apports de 2015 ont diminué de 245 kcal en énergie, de 109 mg en calcium et de 418 mg en sodium. Les LMP contiennent des images de vaisselles qui sont utilisées pour quantifier les aliments consommés. Le LMP de 2004 a été copié d'une version américaine. Le LMP de 2015 a été révisé pour mieux refléter la taille de la vaisselle canadienne. Ces révisions ont été examinées pour expliquer les réductions observées entre 2004 et 2015. Les quantités d'aliments déclarées en 2004 dans l'ESCC ont été ajustées pour imiter le LMP révisé. Les apports ajustés de 2004 ont ensuite été comparés à ceux de 2015. Les révisions du LMP représentent de 11% à 37% de la différence en énergie, de 23% à 58% en calcium et de 12% à 34% en sodium. Les révisions apportées au LMP de 2015 ne sont pas négligeables. Les différences observées entre les enquêtes doivent être interprétées avec prudence.

Recent Statistics Research of New Investigators Across Canada
Récentes recherches des nouveaux chercheurs en statistique au Canada

Chair/Président: Reza Ramezan

Organizer/Responsable: Reza Ramezan

Room/Salle: 146 (SB)

Abstract/Résumé

[Monday May 27/lundi 27 mai, 15:30-16:00]

Félix Camirand Lemyre (Université de Sherbrooke) , **Aurore Delaigle** (University of Melbourne) , **Raymond J. Carroll** (Texas AM University)

Non-Parametric Estimation of the Distribution of Episodically Consumed Food Measured with Error

Estimation non paramétrique de la distribution de l'apport habituel d'aliments épisodiquement consommés

In national surveys, dietary data are typically collected to estimate the distribution of the usual intake of various nutrients and food groups in populations and subpopulations. Dietary intakes are often assessed using self-report instruments that allow capturing food and nutrient intakes for a single day only. Since this snapshot cannot accurately reflect the usual intake, it has long been recognized that such observations are versions of long-term average intakes contaminated by measurement errors. When the food/nutrient is consumed daily, a vast literature on measurement errors shows how to estimate the distribution of the individuals' usual intake. However, the classical methods cannot be used when considering the usual intake of episodically consumed food (e.g., fish or whole fruits) because in that case a significant proportion of reported intake is equal to zero. In this presentation, we address this problem by using a non-parametric approach.

Dans des enquêtes nationales, il est souvent d'intérêt d'estimer la distribution de l'apport habituel de nutriments ou d'aliments dans les populations ou sous-populations. Dans ce contexte, les apports alimentaires sont souvent rapportés au moyen d'outils d'auto-évaluation qui mesurent la quantité d'aliments et de nutriments consommés dans une journée. Puisque cette mesure constitue un reflet imprécis de l'apport habituel, il est désormais reconnu que de telles observations peuvent être considérées comme des versions de l'apport habituel contaminées par des erreurs de mesure. Dans le cas d'un apport quotidien, il est possible d'utiliser les méthodes classiques d'erreurs de mesure pour estimer la distribution de l'apport moyen à partir de ce type de données. Or, ces méthodes ne peuvent plus être utilisées lorsque l'aliment est consommé épisodiquement (p. ex., poisson ou fruits) puisque les données récoltées présenteront typiquement un excès de zéros. Nous abordons ici ce problème en adoptant une approche non paramétrique.

[Monday May 27/lundi 27 mai, 16:00-16:30]

Jeffrey L. Andrews (University of British Columbia — Okanagan)

On Overfitting in Cluster Analysis

Du surajustement en analyse typologique

Several simulated examples of overfitting in the context of unsupervised learning will be discussed, including mixture model-based as well as other popular clustering methodologies. An approach at combatting this issue based on incorporating the nonparametric bootstrap within each model's optimization procedure will be put forward, and the pros and cons of such an approach will be emphasized. Applications to real data will be provided along with future research directions.

Nous discuterons de plusieurs exemples simulés de surajustement dans le contexte de l'apprentissage non supervisé, notamment des méthodes fondées sur les modèles de mélange et d'autres méthodes de regroupement populaires. Nous proposerons une approche pour résoudre ce problème qui intègre le bootstrap non paramétrique dans la procédure d'optimisation de chaque modèle et insisterons sur les avantages et les inconvénients d'une telle approche. Nous présenterons des applications à des données réelles et identifierons des futurs axes de recherche.

[Monday May 27/lundi 27 mai, 16:30-17:00]

Jonathan Jalbert (Polytechnique Montreal) , **Luc Perreault** (Institut de recherche d'Hydro-Québec)

Interpolation of Extreme Precipitation of Multiple Durations in Eastern Canada
Interpolation d'extrêmes de précipitations de multiples durées dans l'Est du Canada

Intensity-duration-frequency (IDF) curves for extreme precipitation events are widely used to design civil infrastructures like sewers and dikes. An important issue raised by the Canadian Standards Association concerns the sparsity of the location where IDF curves are available. The goal of the present work consists in properly interpolating the extreme precipitation of several durations in order to compute the IDF curves everywhere in Eastern Canada. In order to gather physical information where no observation is available, a reconstruction of the historical meteorology is used as the covariate for interpolating extreme precipitation characteristics. Such a covariate is included in a hierarchical Bayesian spatial model for the extreme precipitation. This spatial model is especially suited for the covariate gridded structure, hence enabling fast and precise computations. This model provides reliable IDF curves over the whole spatial domain.

Les courbes d'intensité-durée-fréquence (courbes d'IDF) pour les extrêmes de précipitations sont communément utilisées dans la conception d'infrastructures civiles comme les égouts et les digues. L'un des problèmes importants soulevés par l'Association canadienne de normalisation concerne le peu de sites pour lesquels ces courbes d'IDF sont disponibles. L'objectif des présentes recherches est de bien interpoler les extrêmes de précipitations de plusieurs durées afin de calculer les courbes d'IDF pour l'ensemble de l'Est du Canada. En vue de la collecte d'informations physiques en l'absence de toute observation, nous utilisons une reconstruction de la météorologie historique comme covariable pour l'interpolation de caractéristiques des extrêmes de précipitations. Ces covariables sont incluses dans un modèle spatial bayésien hiérarchique des extrêmes de précipitations. Ce modèle spatial est particulièrement bien adapté à la structure de grille de covariables, permettant des calculs rapides et précis. Ce modèle produit des courbes d'IDF fiables sur l'ensemble du domaine spatial.

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

Chair/Président: Alison L. Gibbs

Room/Salle: 142 (AD)

Abstract/Résumé

[Monday May 27/lundi 27 mai, 15:30-15:45]

Steven Wu (Shopify)

People Analytics: How Shopify Uses Statistical Methods to Make Better Decisions for Its Employees

People analytics : comment Shopify utilise-t-il des méthodes statistiques pour prendre de meilleures décisions pour ses employés

Shopify's People Analytics team collects and analyzes data about the Shopify team to (a) help leadership make data informed decisions and (b) help make our workplace more efficient and engaging. We regularly use statistical methods to influence decisions around hiring, performance, retention, engagement, learning, culture, and more. Through our work we get to contribute to open source tooling to benefit the Python statistical community as well! In this talk I'll introduce the field of people analytics, go over some interesting case studies, and give a peek into our workflow of how a statistician / data scientist operates at Shopify. As a recent graduate, I hope to illustrate the burgeoning opportunities for statisticians that I see in this field and decision science in general.

L'équipe de la people analytics de Shopify recueille et analyse des données sur l'équipe de Shopify pour (a) aider la direction à prendre des décisions basées sur les données et (b) aider à rendre le milieu de travail plus efficace et engageant. Nous utilisons régulièrement des méthodes statistiques pour influencer les décisions sur l'embauche, la performance, la rétention, l'implication, l'apprentissage, la culture et plus. Notre travail nous permet aussi de contribuer à des logiciels libres d'accès pour aider la communauté statistique Python ! Dans cet exposé, je présenterai le domaine de la people analytics, je résumerai quelques études de cas intéressants et je donnerai un aperçu du travail d'un statisticien/scientifique des données chez Shopify. Récemment diplômé, j'espère illustrer les opportunités grandissantes pour les statisticiens dans ce domaine et dans la science décisionnelle en générale.

[Monday May 27/lundi 27 mai, 15:45-16:00]

Shan Shi (University of Victoria) , **Farouk Nathoo** (University of Victoria)

Feature Learning and Classification in Neuroimaging: Predicting Cognitive Impairment from Magnetic Resonance Imaging

Apprentissage et classification des caractéristiques en neuroimagerie : prédiction des troubles cognitifs causés par l'imagerie par résonance magnétique

Due to the rapid innovation of technology and the desire to find and employ biomarkers for neurodegenerative disease, high-dimensional data classification problems are routinely encountered in neuroimaging studies. To avoid over-fitting and to explore relationships between disease and potential biomarkers, feature learning and selection plays an important role in classifier construction and is an important area in machine learning. We review several important feature learning and selection techniques including lasso-based methods, principal components analysis, the two-sample t-test, and stacked auto-encoders. We compare these approaches using a numerical study involving the prediction of Alzheimer's disease from magnetic resonance imaging.

En raison de l'innovation rapide de la technologie et du désir de trouver et d'utiliser des biomarqueurs pour les maladies neurodégénératives, on rencontre régulièrement des problèmes de classification des données de grande dimension dans les études de neuroimagerie. L'apprentissage et la sélection des caractéristiques jouent un rôle important dans l'élaboration du classificateur, constituent un domaine important de l'apprentissage automatique, et ils permettent d'éviter le surapprentissage et d'explorer les relations entre la maladie et les biomarqueurs potentiels. Nous examinons plusieurs techniques importantes d'apprentissage et de sélection des caractéristiques, ainsi que les méthodes de type lasso, l'analyse en composantes principales, le test t à deux échantillons et les auto-encodeurs empilés. Nous comparons ces approches à l'aide d'une étude numérique portant sur la prédiction de la maladie d'Alzheimer par imagerie par résonance magnétique.

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

[Monday May 27/lundi 27 mai, 16:00-16:15]

Christopher Salahub (University of Waterloo) , **Wayne Oldford** (University of Waterloo)

About 'Her Emails'

À propos des « courriels d'Hillary Clinton »

Interactive visual tools designed to facilitate the visual exploration and investigation of emails sent over Secretary Hillary Clinton's controversial and politically important private email server are presented. These tools are implemented, alongside data extracted from the emails using text processing, in an interactive, public web application at https://shiny.math.uwaterloo.ca/sas/clinton to allow those interested to use their functionality. This functionality is used to view email volumes, keywords, and key people involved in Clinton's tenure as United States Secretary of State, in particular during the Libyan Civil War in 2011 and her divisive response to the attack on the American embassy in Benghazi. Patterns present, such as email gaps, are examined in light of these events and compared to less controversial times in an attempt to remove personal bias. A clear picture emerges of the conclusions supported by the data and those which cannot be explained.

Nous présentons des outils visuels interactifs conçus pour faciliter l'exploration visuelle et l'étude des courriels envoyés via le serveur de messagerie si controversé et politiquement sensible de la Secrétaire d'État Hillary Clinton. Pour quiconque souhaite utiliser leur fonctionnalité, ces outils sont mis en œuvre, avec des données extraites des courriels par traitement de texte, sur une application Web publique interactive à https ://shiny.math.uwaterloo.ca/sas/clinton/. Ils permettent de visualiser le volume de courriels, les mots-clés et les personnes importantes impliquées pendant le mandat de Secrétaire d'État des États-Unis Clinton, notamment pendant la guerre civile en Libye de 2011, et sa réaction controversée à l'attentat contre l'ambassade américaine à Benghazi. Nous examinons les tendances comme l'absence de courriels à la lumière de ces évènements et les comparons à des périodes moins controversées, pour tenter d'éliminer tout biais personnel. Une idée claire ressort quant aux conclusions supportées par les données et celles qui ne peuvent être expliquées.

[Monday May 27/lundi 27 mai, 16:15-16:30]

Usama Zafar Ansari (The University of British Columbia) , **Chengkai Zhang** (University of British Columbia, Okanagan)

Statistical Analysis of Vessel Motion Patterns in the Ports and Waterways Using Automatic Identification System (AIS)

Analyse statistique des mouvements des navires dans les ports et sur les voies navigable en utilisant le système d'identification automatique (SIA)

The research is devoted to statistical analysis of vessel motion patterns in the ports and waterways using Automatic Identification System (AIS). AIS is an automatic ship self-reporting system used for maritime transportation purposes. The AIS system broadcasts vessel sailing information like position, speed, and status, which can then be received by other ships or Vessel Traffic Service (VTS) centers. From historic AIS data we extract the trajectory patterns which are then used to construct the prediction models to determine vessel trajectory. The challenges faced are unlike those for vehicles driving on the road; a ship vessel can move from one port to the other port with various speeds and different trajectories, which makes it difficult to analyze which port is its destination. The capability to accurately forecast the movement of ships globally would potentially enable the maximization of business trading profits and safety of life at sea.

Cette étude est consacrée à l'analyse statistique des mouvements des navires dans les ports et sur les voies navigable en utilisant le système d'identification automatique (SIA). SIA est un système automatique d'autodéclaration des vaisseaux utilisé pour le transport maritime. Le système SIA diffuse l'information de navigation du navire telle que la position, la vitesse et le statut, qui peut alors être reçue par d'autres navires ou par les centres de service du trafic maritime (STM). À partir de données historiques du SIA, nous avons extrait des schémas des trajectoires qui sont ensuite utilisés pour élaborer des modèles de prédiction pour déterminer la trajectoire des navires. Le défi rencontré est que contrairement aux véhicules qui circulent sur les routes, un navire peut se déplacer d'un port à l'autre à différentes vitesses et en employant différentes trajectoires, rendant difficile la détermination du port de destination. La capacité de prédire avec précision le mouvement des navires pourrait potentiellement permettre la maximisation des profits du commerce et la sécurité de la vie en mer.

[Monday May 27/lundi 27 mai, 16:30-16:45]

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

Gabriel C. Phelan (Simon Fraser University) , **David A. Campbell** (Simon Fraser University)

Geographically Aware Latent Dirichlet Allocation via Random Effects

Allocation de Dirichlet latente géographiquement consciente par l'entremise d'effets aléatoires

Since its inception in 2003, Latent Dirichlet Allocation (LDA) (Blei et. al., 2003) has become an increasingly important tool within natural language processing for extracting latent topics from a set of documents. Various extensions have been proposed to incorporate covariates into the LDA framework; we are particularly interested in those which address the geographical diversity of language. Building on recent advances (Eisenstein et. al., 2010, Yin et. al., 2011), we consider topic probabilities as varying across a set of geographical regions based on a latent random effects model. These geo-specific topic probabilities are then used within standard LDA, and fit using state-of-the-art Markov Chain Monte Carlo (MCMC) methods such as Hamiltonian Monte Carlo (HMC). Our model is employed to study geographically-driven beer preferences across Canada, where regions correspond to Canadian provinces and "documents" are taken to be comments left by users about particular Canadian beers.

Depuis ses débuts en 2003, l'allocation de Dirichlet latente (ADL) (Blei et. al., 2003) est un outil qui a pris de plus en plus d'importance en traitement automatique du langage naturel pour l'extraction de thématiques latentes à partir d'un ensemble de documents. Plusieurs extensions ont été proposées pour intégrer les covariables dans le cadre d'une ADL, mais nous sommes tout particulièrement intéressés par celles abordant la diversité géographique d'une langue. En nous basant sur de récentes avancées (Eisenstein et coll., 2010, Yin et coll., 2011), nous considérons que les probabilités thématiques varient à travers un ensemble de régions géographiques basé sur un modèle latent à effets aléatoires. Nous nous servons ensuite de ces probabilités thématiques géospécifiques à l'intérieur d'une ADL standard, puis les ajustons au moyen de méthodes Monte Carlo par chaînes de Markov (MCCM) de pointe comme le Monte Carlo hamiltonien (MCH). Notre modèle est employé pour étudier les préférences en matière de bière en fonction des régions canadiennes, où les régions correspondent aux provinces canadiennes et les «documents» sont tirés de commentaires concernant certaines bières que des utilisateurs ont publiés.

[Monday May 27/lundi 27 mai, 16:45-17:00]

Bo Chen (DLSPH, University of Toronto) , **Keith A. Lawson** (University Health Network) , **Antonio Finelli** (University Health Network) , **Olli Saarela** (University of Toronto)

Four-Way Causal Variance Decompositions for Evaluating Hospital and Surgeon Performance

Décompositions de variance causale à quatre sens pour évaluer la performance des hôpitaux et des chirurgiens

Disease-specific quality indicators are used to compare institutions and health care providers in terms processes or outcomes relevant to treatment of a particular condition. In the context of surgical cancer treatments, the performance variations can be due to hospital or surgeon level differences in facilities and practices. We consider how the observed variation in care received at patient level can be decomposed into that causally explained by the hospital performance, the surgeon performance, the patient case-mix, and the unexplained (residual) variation. For this purpose, we derive a four-way variance decomposition, with particular attention to the causal interpretation of the components. For estimation, we use inputs from a mixed-effect model with random hospital/surgeon-specific effects, and a multinomial logistic model for the hospital/surgeon-specific patient populations. We demonstrate the use of our methods in administrative data on kidney cancer care in On-

Les indicateurs de qualité spécifiques à une maladie sont utilisés pour comparer des institutions et des prestataires de soin en termes de processus ou de résultats pertinents pour le traitement d'une condition particulière. Dans le contexte des traitements chirurgicaux du cancer, les variations dans la performance peuvent être dues aux différences au niveaux des hôpitaux ou des chirurgiens dans les pratiques et les installations. Nous examinons la façon dont la variation observée dans les soins reçus au niveau du patient peut être décomposée par ce qui est causalement expliqué par la performance de l'hôpital, la performance du chirurgien, et la classification « case-mix » du patient, et la variation inexpliquée (résiduelle). À cette fin, nous élaborons une décomposition de la variance à quatre sens et nous portons une attention particulière à l'interprétation causale des composantes. Pour l'estimation, nous utilisons des entrées provenant d'un modèle à effet mixte avec effets aléatoires spécifiques aux hôpitaux et aux chirurgiens, et un modèle logistique multinomiale pour les populations de patients spécifiques aux hôpitaux et aux chirurgiens. Nous démontrons

tario.

l'utilisation de nos méthodes avec des données sur les soins du cancer du rein en Ontario.

Software Development and Computationally-Intensive Methods
Mise au point de logiciels et méthodes à forte intensité de calculs

Chair/Président: David Saunders

Room/Salle: 201 (ENA)

Abstract/Résumé

[Monday May 27/lundi 27 mai, 15:30-15:45]

Song Cai (Carleton University) , **Laura Dumitrescu** (Victoria University of Wellington) , **JNK Rao** (Carleton University) , **Golshid Chatrchi** (Carleton University)

A Simple and Effective Variable Selection Method for Two-Fold Sub-Area Models in Small Area Estimation

Méthode de sélection de variables simple et efficace pour modèles doubles de sous-domaine en estimation pour petits domaines

This work aims to develop a simple yet effective method for variable selection in a two-fold sub-area-level model for small area estimation. A two-fold model consists of a sampling model describing the relationship between direct estimators of sub-area means and the true sub-area means, and a linking model connecting the unknown true sub-area means to covariates with additive random effects. Due to this complex structure, existing variable selection methods for regression models are not directly applicable. We propose to first transform the linking model into a linear model with IID errors. We then transform the sampling model accordingly and use the observed data and the transformed sampling model to approximate an information criterion such as AIC or BIC for the transformed linking model. Simulation studies show that the proposed method outperforms simple competitors especially when the variance of the area-level random effect in the linking model is substantial.

Cette présentation vise à développer une méthode simple mais efficace pour la sélection de variables dans un modèle double de sous-domaines en estimation pour petits domaines. Un modèle double consiste d'un modèle d'échantillonnage, qui décrit la relation entre les estimateurs directs des moyennes des sous-domaines et les moyennes réelles des sous-domaines, et un modèle de liaison, qui relie la moyenne réelle inconnue du sous-domaine à des covariables avec effets aléatoires additifs. En raison de cette structure complexe, les méthodes de sélection de variables existantes pour les modèles de régression ne sont pas directement applicables. Nous proposons, dans un premier temps, de transformer le modèle de liaison en un modèle linéaire avec erreurs iid. Ensuite, nous transformons le modèle d'échantillonnage en conséquence, puis utilisons les données observées et le modèle d'échantillonnage transformé pour estimer un critère d'information comme AIC ou BIC pour le modèle de liaison transformé. Nous montrons, par des études de simulation, que la méthode proposée fonctionne mieux que ses concurrents simples, surtout quand la variance de l'effet aléatoire du niveau du domaine est substantiel dans le modèle de liaison.

[Monday May 27/lundi 27 mai, 15:45-16:00]

Martin Lysy (University of Waterloo)

The 'msde' Package: Fast Inference for Stochastic Differential Equations in R

La bibliotèque R ≪msde≫ : une inférence rapide appliquée aux équations différentielles stochastiques dans R

Stochastic Differential Equations (SDEs) are used to model a wealth of stochastic phenomena in the natural sciences, engineering, and finance. However, parameter inference typically requires high-dimensional integration of state-space models. While many sophisticated algorithms have been developed for this purpose, few possess computer implementations applicable beyond toy examples. The purpose of the 'msde' R package is to bridge this gap. At its core, 'msde' provides efficient C++ implementations of several state-of-the-art in-

Les équations différentielles stochastiques (EDS) servent à modéliser un grand nombre de phénomènes stochastiques dans les domaines des sciences naturelles, de l'ingénierie et des finances. Toutefois, l'inférence de paramètre nécessite généralement une intégration de haute dimension des modèles spatiotemporels. Bien que plusieurs algorithmes sophistiqués aient été conçus pour y arriver, très peu d'entre eux possèdent des implémentations informatiques applicables au-delà d'exemples jouets. L'objectif de la bibliothèque R ≪msde≫ est donc de réunir ces deux conditions. À la base, ≪msde≫ procure d'efficaces implémentations C++ de

ference algorithms, parallelizing calculations whenever possible. At the interface level, users can integrate their SDE models directly with the underlying code base, with next to zero knowledge of C++ required. The extensive package documentation includes numerous SDE examples, many of which serve to showcase the main 'msde' features.

nombreux algorithmes d'inférences parmi les plus récents, tout en parallélisant les calculs si possible. Au niveau de l'interface, les utilisateurs, sans même connaître C++, peuvent intégrer leurs modèles EDS directement au moyen du code de base sous-jacent. La documentation complète concernant la bibliotèque comprend de nombreux exemples d'EDS, dont beaucoup servent à mettre en valeur les principales fonctions de ≪msde≫.

[Monday May 27/lundi 27 mai, 16:00-16:15]

Dan Richard (MacEwan University) , **Karen Buro** (MacEwan University) , **Wanhua Su** (MacEwan University)

Zero Order vs (Semi) Partial Correlation Test and Confidence Interval

Test de corrélation d'ordre zéro vs corrélation (semi) partielle et intervalle de confiance

A bootstrap test was created to measure the significance of a partial or semi-partial correlation equaling its zero order correlation. This was implemented in R and accepted by CRAN as an open source package named zeroEQpart, which can be accessed by external researchers. In addition to a hypothesis test, the functionality of the R package was extended to include confidence intervals for the parameter (zero order minus partial). The test was administered retrospectively on an honours psychology thesis to analyze the effect of metacognitions on chronic pain and health anxiety.

Un test bootstrap a été créé pour mesurer l'importance d'une corrélation partielle ou semi-partielle égalant sa corrélation d'ordre zéro. Ceci fut implémenté dans R et accepté par CRAN comme composant logiciel en libre accès nommé zeroEQpart qui est accessible aux chercheurs externes. Outre le test d'hypothèses, la fonctionnalité de la bibliothèque R a été élargie pour inclure les intervalles de confiance pour le paramètre (ordre zéro moins partiel). Le test a été utilisé rétrospectivement sur une thèse en psychologie pour analyser l'effet de la métacognition sur la douleur chronique et l'anxiété.

[Monday May 27/lundi 27 mai, 16:15-16:30]

Avinash Prasad (University of Waterloo) , **Marius Hofert** (University of Waterloo) , **Mu Zhu** (University of Waterloo)

Quasi-Random Number Generators for Multivariate Distributions Based on Generative Neural Networks

Générateurs de nombres quasi-aléatoires pour distributions multivariées fondés sur des réseau de neurones génératifs

Generative moment matching networks are introduced as quasi-random number generators for multivariate distributions. So far, quasi-random number generators for non-uniform multivariate distributions require a careful design, often need to exploit specific properties of the distribution or quasi-random number sequence under consideration, and are limited to few models. Utilizing generative neural networks, in particular, generative moment matching networks, allows one to construct quasi-random number generators for a much larger variety of multivariate distributions without such restrictions. Once trained, the presented generators only require independent quasi-random numbers as input and are thus fast in generating non-uniform multivariate quasi-random number sequences from the target distribution. Various numerical examples are considered to demonstrate the approach, including applications inspired by risk management practice.

Les réseaux d'appariement de moments génératifs sont présentés comme des générateurs de nombres quasi-aléatoires pour des distributions multivariées. Jusqu'à présent, les générateurs de nombres quasi-aléatoires pour distributions multivariées non-uniformes nécessitent une conception minutieuse, doivent souvent exploiter des propriétés spécifiques de la distribution ou de la séquence de nombres quasi-aléatoires à l'étude et sont limités à seulement quelques modèles. L'utilisation des réseaux de neurones génératifs, plus particulièrement des réseaux d'appariement de moments génératifs, permet l'élaboration de générateurs de nombres quasi-aléatoires pour une plus vaste variété de distributions multivariées sans ces limitations. Une fois formés, les générateurs présentés exigent seulement, comme entrées, des nombres quasi-aléatoires indépendants provenant de la distribution ciblée. Plusieurs exemples numériques sont examinés pour démontrer l'approche, dont des applications inspirées da la pratique de la gestion du risque.

[Monday May 27/lundi 27 mai, 16:30-16:45]

Peter D.M. Macdonald (McMaster University)

Bootstrapping Finite Mixture Distributions
Bootstrap de distributions de mélange finies

My R package, mixdist, fits univariate finite mixture distributions to grouped data by maximum likelihood. The algorithm uses one or two EM steps followed by numerical quasi-Newton optimization. To bootstrap grouped data, simply resample the multinomial distribution using either observed frequencies (for a nonparametric bootstrap) or fitted frequencies (for a parametric bootstrap). The bootstrap function in mixdist displays the resampling histograms and fits as they are produced and teaches a valuable lesson by showing how variable different samples from the same population can be, and how easy it is to over-interpret a single sample. Likelihood inference in finite mixtures is notoriously nonlinear so it is interesting to see examples where the bootstrap standard errors do or do not agree with standard errors computed from the Fisher information. The presentation will include a live demonstration of the package.

Ma bibliothèque R, mixdist, ajuste des distributions de mélange finies univariées pour des données groupées par vraisemblance maximale. L'algorithme utilise une ou deux étapes EM suivie d'une optimisation numérique quasi-Newton. Pour appliquer le bootstrap à des données groupées, il suffit de ré-échantillonner la distribution multinomiale en utilisant soit les fréquences observées (pour un bootstrap non-paramétrique) ou les fréquences ajustées (pour un bootstrap paramétrique). La fonction bootstrap dans mixdist affiche les histogrammes du ré-échantillonnage et ajuste au fur et à mesure qu'ils sont réalisés et enseigne une leçon importante en montrant à quel point des échantillons provenant de la même population peuvent varier et comment il est facile de surinterpréter un échantillon unique. L'inférence de vraisemblance dans des mélanges finis est non-linéaire, il est donc intéressant de voir des exemples où les erreurs types du bootstrap concordent ou pas avec les erreurs types calculées à partir de l'information Fisher. L'exposé se terminera avec une démonstration en directe du bibliothèque R.

[Monday May 27/lundi 27 mai, 16:45-17:00]

Jun Yang (University of Toronto) , **Zhou Zhou** (University of Toronto)

Spectral Inference under Complex Temporal Dynamics
Inférence spectrale en dynamique temporelle complexe

We develop unified theory and methodology for the inference of evolutionary Fourier power spectra for a general class of locally stationary and possibly nonlinear processes. In particular, simultaneous confidence regions (SCR) with asymptotically correct coverage rates are constructed for the evolutionary spectral densities on a nearly optimally dense grid of the joint time-frequency domain. A simulation-based bootstrap method is proposed to implement the SCR. The SCR enables researchers and practitioners to visually evaluate the magnitude and pattern of the evolutionary power spectra with an asymptotically accurate statistical guarantee. The SCR also serves as a unified tool for a wide range of statistical inference problems in time-frequency analysis ranging from tests for white noise, stationarity and time-frequency separability to the validation for nonstationary linear models.

Nous développons une théorie et une méthodologie unifiées pour l'inférence de spectres de puissance de Fourier évolutionnaires pour une classe générale de processus localement stationnaires et possiblement non linéaires. En particulier, nous construisons des régions de confiance simultanée (RCS) avec des taux de couverture asymptotiquement corrects pour les densités spectrales évolutionnaires sur un réseau de densité presque optimale du domaine de temps-fréquence commun. Nous proposons une méthode de bootstrap fondée sur la simulation pour mettre en œuvre les RCS. La RCS permet aux chercheurs et aux praticiens d'évaluer visuellement l'ampleur et les caractéristiques des spectres de puissance évolutionnaires avec une garantie statistique asymptotiquement correcte. La RCS sert aussi d'outil unique pour divers problèmes d'inférence statistique en analyse temps-fréquence, de l'identification de bruit blanc, de stationnarité et de séparabilité temps-fréquence à la validation de modèles linéaires non stationnaires.

Patient-Focused Statistical Methods
Méthodes statistiques axées sur le patient

Chair/Président: Ali Karimnezhad

Room/Salle: 101 (ENA)

Abstract/Résumé

[Monday May 27/lundi 27 mai, 15:30-15:45]

Zhihui (Amy) Liu (Princess Margaret Cancer Centre) , **Olli Saarela** (University of Toronto) , **Wei Xu** (Princess Margaret Cancer Centre, University Health Network)

Swimmers and Sinkers: Statistical Basis of the Swimmer Plot

Nageurs et pesées : base statistique du graphique du nageur (swimmer plot)

Recently the so-called swimmer plot has been used in cancer research to depict duration of treatment, response to treatment, and disease progression at the patient level. The follow-up history of the patients, the "swimmers", is displayed along parallel horizontal lines over the time axis. In a randomized clinical trial, the swimmer plot can be used to visually contrast clinical endpoints such as duration of response or duration of stable disease under treatment arms. However, such an individual-level presentation quickly becomes cluttered with increasing study size. In this talk, our objective is to explore and establish statistical underpinnings of such a presentation. In particular, we point out connections to probabilities in a multi-state model, and to the so-called population-time plot, a visualization of incidence density that is useful in describing the follow-up of a larger study.

Le graphique du nageur a récemment été utilisé dans la recherche sur le cancer pour illustrer la durée du traitement, la réponse au traitement et la progression de la maladie au niveau du patient. L'historique de suivi des patients, des « nageurs », est présenté par des lignes horizontales parallèles à l'axe du temps. Dans un essai clinique randomisé, le graphique du nageur peut être utilisé pour contraster visuellement les résultats cliniques tels que la durée de la réponse ou la durée de stabilité de la maladie des groupes de traitement. Par contre, une présentation au niveau individuel devient rapidement encombrée lorsque la taille de l'étude augmente. Dans cet exposé, notre objectif est d'explorer et d'établir le fondement statistique d'une telle présentation. Nous démontrons notamment les liens avec les probabilités dans un modèle multi-états et avec le graphique temps-population, une visualisation de la densité d'incidence qui est utile pour décrire le suivi d'une étude plus vaste.

[Monday May 27/lundi 27 mai, 15:45-16:00]

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

Dynamic Treatment Regimes with Interference

Régimes de traitement dynamiques avec interférence

In the context of precision medicine, Dynamic Treatment Regimes (DTRs) are sequences of decision rules that take individual patient data as input and produce treatment recommendations. As in the wider causal inference literature, DTR estimation relies upon the key assumption of no interference: that the outcome of one individual is unaffected by the treatment assignment of others. In many social network contexts, this assumption is questionable. We consider a DTR analysis framework with a continuous outcome, and focus on Dynamic Weighted Ordinary Least Squares (dWOLS), which is an easy-to-use and doubly robust estimation approach for estimating optimal DTRs. However, the validity of the double robustness of dWOLS is impacted by inter-

Dans le contexte de la médecine de précision, les régimes de traitement dynamiques (RTD) consistent en séquences de règles décisionnelles avec comme point d'entrée les données sur un patient désigné et comme résultats des recommandations de traitement. Dans la littérature sur l'inférence causale, au sens large, l'estimation RTD s'appuie sur l'hypothèse fondamentale de la non interférence : les résultats pour un patient ne sont pas affectés par l'attribution d'un traitement à d'autres. Dans plusieurs contextes de réseaux sociaux, cette hypothèse est mise en question. Nous envisageons un cadre d'analyse avec des données de sortie continues, en mettant l'accent sur une méthode de moindres carrés pondérés dynamiques (MCPd), une approche d'estimation facile à utiliser et doublement robuste pour estimer des RTD optimaux. Par contre, l'interférence influe sur la validité de la double ro-

ference. Using the algebra of partitioned regression, we explain the invalidity of double robustness of dWOLS when we consider interference, and propose two approaches to maintaining this important property.

bustesse des MCPd. En nous fondant sur l'aspect algébrique de la régression partitionnée, nous expliquons la non-validité de la double robustesse des MCPd sous l'angle de l'interférence et proposons deux approches pour conserver cette propriété importante.

[Monday May 27/lundi 27 mai, 16:00-16:15]

Alomgir Hossain (University of Ottawa/University of Ottawa Heart Institute) , **Benjamin Chow** (University of Ottawa Heart institute)

Long-Term Prognostic Value of Coronary CT Angiography

Valeur prognostique à long terme d'une coronarographie par tomodensitométrie (TDM)

Objective: We sought to determine the prognostic and incremental value of CAD severity, coronary atherosclerosis and left ventricular ejection fraction (LVEF) measured with CTA. Background: Computed tomographic coronary angiography (CTA) is an emerging tool used for the detection of obstructive coronary artery disease (CAD).However there is limited data supporting the prognostic value of 64-slice CTA and its ability to predict all-cause mortality and major adverse cardiac events (MACE) such as cardiac death and non-fatal myocardial infarction (MI). Methods: Between February 2006 and February 2018, 12,423 consecutive patients were prospectively enrolled and followed for a mean of 12 years. Each CTA was evaluated for CAD severity, total plaque score (TPS) and LVEF. Cox proportional hazard models were used to assess the independent prognostic value of CAD severity after adjusting the propensity score to take into account confounding factors.

Notre objectif : déterminer la valeur prognostique et incrémentielle de la gravité de la maladie coronarienne, de l'athérosclérose coronarienne et de la fraction d'éjection ventriculaire gauche (FEVG) mesurées par tomographie numérisée. Point de départ : la coronarographie par tomodensitométrie (TDM) est un outil d'usage récent pour le dépistage de la maladie coronarienne obstructive (MCO). Un nombre limité de données appuie cependant la valeur prognostique de la MCO de 64 barrettes et sa capacité de prédire la mortalité toutes causes confondues et les événements cardiaques majeurs indésirables (ECMI), tel que le décès de nature cardiaque et l'infarctus du myocarde (IM) non mortel. Méthodes : Entre février 2006 et février 2018, 12 423 patients consécutifs ont participé à une étude prospective et ont été suivis en moyenne 12 ans. Chaque TDM a été évaluée à l'égard de la gravité de la maladie coronarienne, du nombre total de plaques et de la FEVG. Des modèles à risque proportionnel de Cox ont été utilisés pour évaluer la valeur prognostique de la gravité de la maladie coronarienne après rajustement du score de propension afin de prendre en compte les facteurs de confusion.

[Monday May 27/lundi 27 mai, 16:15-16:30]

Dylan Spicker (University of Waterloo) , **Michael Wallace** (University of Waterloo)

Measurement Error in Precision Medicine and Dynamic Treatment Regimes

Erreur de mesure dans la médicine de précision et régimes thérapeutiques dynamiques

Precision medicine improves patient outcomes using patient-level covariates to tailor treatment. In longitudinal studies with time-varying covariates and sequential treatment decisions, precision medicine can be formalized with dynamic treatment regimes (DTRs). DTRs are sequences of covariate-dependent treatment rules that optimize expected outcomes. To date, the precision medicine literature has not addressed a ubiquitous concern in health research - measurement error - where observed data deviate from the truth. We will discuss the consequences of ignoring measurement error in the context of DTRs, with a focus on challenges unique to the personalized medicine setting. We show that relatively simple measurement error corrections provide substan-

La médecine de précision améliore la condition des patients en utilisant des covariables au niveau du patient pour adapter le traitement. Dans les études longitudinales avec des covariables qui varient dans le temps et des décisions de traitements séquentiels, la médecine de précision peut être formalisée avec des régimes thérapeutiques dynamiques (RTD). Les RTD sont des séquences de règles de traitement dépendantes de la covariable qui optimisent les résultats attendus. Jusqu'à ce jour, la littérature en médecine de précision n'a pas abordé une préoccupation omniprésente dans la recherche en santé – l'erreur de mesure – où les données observées s'écartent de la vérité. Nous discuterons des conséquences qui découlent du fait d'ignorer les erreurs de mesure dans le contexte des RTD avec un accent mis sur les défis propres au cadre de la médecine personnalisée. Nous démontrons que des

tial improvement over uncorrected analyses. Along with theoretical results, we demonstrate our findings through simulation and apply them to a study of depression (STAR*D) not previously analyzed for measurement error.

corrections relativement simples des erreurs de mesure permettent une amélioration considérable comparativement aux analyses non-corrigées. En plus des résultats théoriques, nous démontrons nos résultats par la simulation et nous les appliquons à une étude sur la dépression (STAR*D) qui n'a pas été précédemment analysée pour des erreurs de mesures.

[Monday May 27/lundi 27 mai, 16:30-16:45]

Katherine Daignault (University of Toronto) , **Keith A. Lawson** (Princess Margaret Cancer Centre, University Health Network) , **Antonio Finelli** (Princess Margaret Cancer Centre, University Health Network) , **Olli Saarela** (University of Toronto)

Implementation of Causal Mediation Analysis in Hospital Profiling

Exécution d'une analyse de médiation causale pour le profilage hospitalier

Process measures (e.g. procedures) are preferable to patient outcomes for targeting hospital quality improvement as interventions to improve care are easier to define. Causal mediation analysis allows decomposition of the total hospital effect on outcomes into an indirect effect acting through a specific process and a direct effect comprising all other pathways. The effect of a hypothetical intervention on the process can then be quantified and interventions targeted where greatest improvement in patient outcomes may occur. We present results of a mediation analysis assessing the impact of minimally invasive (MIS) vs. open surgery on length of hospital stay in surgical treatment of kidney cancer patients in Ontario. The intervention considered is to bring the MIS proportion to the provincial average. We discuss implementation of the methods in the presence of low volume hospitals and compare approaches for estimating the variability of the effect decomposition.

Les mesures d'un traitement (par ex. : les procédures) sont préférables aux résultats obtenus par les patients pour cibler l'amélioration de la qualité en milieu hospitalier puisque les interventions visant l'amélioration des soins sont plus définissables, même si l'objectif final est d'obtenir de meilleurs résultats. L'analyse de médiation causale permet de décomposer l'effet global de l'hôpital sur les résultats en un effet indirect agissant par un processus précis et un effet direct regroupant toutes les autres voies. L'effet d'une intervention hypothétique sur le traitement peut alors être quantifié et des interventions ciblées là où une plus grande amélioration des résultats obtenus par les patients peut se produire. Dans un cas de traitements chirurgicaux de patients atteints de cancer du rein en Ontario, nous présentons les résultats d'une analyse de médiation qui évalue l'impact sur la durée de séjour à l'hôpital d'une chirurgie mini-invasive (CMI) comparativement à une chirurgie ouverte. L'intervention envisagée est de rapprocher la proportion de CMI à la moyenne provinciale. Nous discutons de la mise en œuvre des méthodes pour des hôpitaux de faible volume et comparons les approches pour l'estimation de la variabilité de la décomposition de l'effet.

[Monday May 27/lundi 27 mai, 16:45-17:00]

Zayd Omar (McGill University) , **David Stephens** (McGill University) , **Alexandra M. Schmidt** (McGill University)

Estimating ICU Heart Rate Data Using a Bayesian State-Space Model with GARCH(1,1) Errors

Estimation des données relatives à la fréquence cardiaque dans une unité de soins intensifs à l'aide d'un modèle d'espace-état bayésien avec erreurs GARCH(1,1)

Heart rate data from the ICU often resembles a non-stationary time series and displays time varying volatility. The non-stationarity shown in the data makes modelling heart rate difficult using standard techniques for time series data. Gaussian State-Space models are a better approach since they allow us to model the non-stationarity observed using latent state vectors. Taking a Bayesian approach, we extend the Gaussian State-Space model by adding a GARCH(1,1) component at the observation level. The volatility clustering

Les données relatives à la fréquence cardiaque obtenues d'une unité de soins intensifs ressemblent souvent à une série temporelle non stationnaire et montrent une volatilité variable avec le temps. La non-stationnarité des données rend difficile la modélisation de la fréquence cardiaque à l'aide des techniques normatives utilisées pour les données de série temporelle. Les modèles d'espace-état gaussiens offrent une meilleure approche en nous permettant de modéliser la non-stationnarité observée à l'aide de vecteurs d'état latent. Avec une approche bayésienne, nous étendons le modèle d'espace-état gaussien en ajoutant une composante GARCH(1,1) à

seen in the observed data is then be modelled by the GARCH(1,1) component. We propose a Markov Chain Monte Carlo (MCMC) algorithm that allows us to estimate the GARCH(1,1) parameters, the latent state vectors and the variance of the state vectors. We compare the estimates from our model to the estimates from a standard Bayesian State-Space model.

l'observation. La grappe de la volatilité que présentent les données observées est ensuite modélisée par la composante GARCH(1,1). Nous proposons un algorithme de Monte-Carlo par chaînes de Markov qui nous permet d'estimer les paramètres GARCH(1,1), les vecteurs de l'état latent et la variance des vecteurs de l'état. Nous comparons les estimations tirées de notre modèle à celles découlant d'un modèle d'espace-état bayésien classique.

Graduate students in actuarial science
Étudiants de troisième cycle en science actuarielle

Chair/Président: Anne Mackay

Organizer/Responsable: Anne Mackay

Room/Salle: 113 (SS)

Abstract/Résumé

[Monday May 27/lundi 27 mai, 15:30-15:45]

Tsz Chai Fung (University of Toronto) , **Andrei Badescu** (University of Toronto) , **Sheldon Lin** (University of Toronto)

A Class of Mixture of Experts Models for General Insurance

Une classe de mélange de modèles experts pour l'assurance dommages

In the Property and Casualty (PC) ratemaking process, it is critical to understand the effect of policyholders' risk profile on the claim distributions and the dependence across business lines. This motivated us to propose a class of logit-weighted reduced mixture of experts (LRMoE) regression models for multivariate claim frequency or severity distributions. The LRMoE contains gating functions that classify policyholders into various latent sub-classes and expert functions that govern the claim distributions. Developing the denseness theory in regression setting, the LRMoE can be fully flexible to capture any distribution, dependence and regression structures. Deriving the marginalization, moment and identifiability properties, we show the LRMoE is mathematically and statistically tractable. Choosing Erlang Count expect functions and developing an efficient algorithm for model calibrations, the LRMoE shows an excellent fitting to a real automobile insurance frequencies dataset.

Dans le processus de tarification de l'assurance dommages, il est essentiel de comprendre l'effet du profil de risques des titulaires de police sur les distributions des sinistres et la dépendance entre les secteurs d'activité. Cela nous a incités à proposer une classe de modèle de régression logit pondéré réduit de mélange d'experts pour les distributions multivariées de fréquence ou la gravité des sinistres. Le modèle de régression logit pondéré réduit de mélange d'experts contient des fonctions de contrôle qui classent les titulaires de police en diverses sous-catégories latentes et des fonctions expertes qui régissent la répartition des sinistres. La création de la théorie de la densité dans un cadre de modèle de régression logit pondéré réduit de mélange d'experts permet d'être totalement souple pour capturer n'importe quelle structure de répartition, de dépendance et de régression. En dérivant les propriétés de marginalisation, de moment et d'identifiabilité, nous montrons que le modèle de régression logit pondérée réduit de mélange d'experts est mathématiquement et statistiquement fiable. En choisissant les fonctions de dénombrement d'Erlang et en élaborant un algorithme efficace pour le calage des modèles, le modèle de régression logit pondéré réduit de mélange d'experts s'adapte parfaitement à un ensemble de données réelles sur les fréquences d'assurance automobile.

[Monday May 27/lundi 27 mai, 15:45-16:00]

Francis Duval (Université du Québec à Montréal) , **Mathieu Pigeon** (Université du Québec à Montréal)

Gradient Boosting Techniques for Individual Loss Reserving in Non-Life Insurance

Techniques de gradient boosting pour la modélisation des réserves individuelles en assurance non-vie

Modeling based on data information is one of the most challenging research topics in actuarial science for loss reserving and risk valuation. Most of the analyzes are based on aggregate data but it is clear that this approach does not tell the whole story about a claim and does not describe precisely its development. Statistical learning approaches in general, and gradient boosting algorithms in particular, offer a set of tools that could help

La modélisation fondée sur des données est l'un des sujets de recherche qui pose le plus de défis dans la science actuarielle pour le provisionnement et l'évaluation du risque. La plupart des analyses sont basées sur des données agrégées, mais il est clair aujourd'hui que cette approche ne dit pas tout sur une réclamation et ne décrit pas précisément son évolution. Les approches d'apprentissage statistique en général, et les algorithmes de gradient boosting en particulier, offrent un ensemble d'outils qui pourraient aider à

to evaluate loss reserves in an individual framework. In this work we contrast some traditional aggregated techniques (portfolio-level) with individual models (claim-level) based on both parametric models and gradient boosting algorithms. These claim-level models use information about each of the payments made for each claim in the portfolio, as well as characteristics of the insured. We provide an explicit example based on a detailed dataset from a property and casualty insurance company and discuss points related to practical applications.

évaluer les réserves dans un cadre individuel. Dans ce mémoire, nous comparons certaines techniques agrégées traditionnelles (au niveau du portefeuille) avec des modèles individuels (au niveau de la réclamation) basés à la fois sur des modèles paramétriques et sur des algorithmes de gradient boosting. Ces modèles individuels utilisent de l'information sur chacun des paiements effectués pour chacune des réclamations du portefeuille, ainsi que sur les caractéristiques de l'assuré. Nous fournissons un exemple basé sur un ensemble de données détaillées provenant d'une compagnie d'assurance de biens et de riques et nous discutons de certains points liés aux applications pratiques.

[Monday May 27/lundi 27 mai, 16:00-16:15]

Jessica Ou Dang (University of Waterloo) , **Mingbin Feng** (University of Waterloo) , **Mary Hardy** (University of Waterloo)

Efficient Nested Simulation of Tail Risk Measures

Simulation emboitée efficace de mesures de risques extrêmes

Tail risk measures are of critical importance for enterprise risk management, especially for managing large portfolios of complex financial instruments. The computational burdens required to simulate these risk measures can be substantial or even infeasible, depending on the complexity of the underlying economic model and the risk management objective. This research proposes, analyzes, and tests an efficient nested simulation procedure for estimating tail risk measures. We propose a procedure that uses proxy models and their concomitants to quickly and accurately identify tail scenarios where the given computational budget can be concentrated. We demonstrate that the proposed procedure performs well in estimating tail risk measures. In particular, our numerical results show that, given a fixed computational budget, the proposed procedure can be an order of magnitude more accurate that a standard nested simulation procedure.

Les mesures de risques extrêmes sont d'une importance cruciale pour la gestion des risques d'entreprise, notamment pour la gestion de gros portefeuilles d'instruments financiers complexes. Le fardeau de calcul requis pour simuler ces mesures de risque peut être excessif, selon la complexité du modèle économique sous-jacent et de l'objectif de gestion de risques. Ce projet de recherche propose, analyse et teste une procédure de simulation emboitée efficace pour estimer les mesures de risques extrêmes. Nous proposons une procédure qui utilise des modèles substituts et leurs concomitants pour identifier rapidement et avec précision les scénarios de risques extrêmes sur lesquels le budget de calcul peut se concentrer. Nous montrons que la procédure proposée donne de bonnes estimations des risques extrêmes. En particulier, nos résultats numériques montrent que pour un budget de calcul fixe, cette procédure peut être d'un ordre de grandeur plus précise qu'une procédure de simulation emboitée standard.

[Monday May 27/lundi 27 mai, 16:15-16:30]

Carlos Andres Araiza Iturria (Concordia University) , **Mélina Mailhot** (Concordia University) , **Frédéric Godin** (Concordia University)

Modeling and Measuring Insurance Risks within the IFRS 17 Framework: A Hierarchical Copula Approach

Modélisation et mesure des risques d'assurance conformes à l'IFRS 17 : approche de copules hiérarchiques

A stochastic approach to insurance risk modeling and measurement that is compliant with IFRS 17 is proposed. The compliance is achieved through the use of a rank-based hierarchical copula which accounts for the dependence between the various lines of business of the Canadian auto insurance industry. A model for the marginal IBNR losses of each line of business based on double generalized linear models is also developed. De-

Nous proposons une approche stochastique de la modélisation et de la mesure des risques d'assurance qui respecte l'IFRS 17. Cette conformité est garantie par l'utilisation d'une copule hiérarchique axée sur le rang qui tient compte de la dépendance entre les divers secteurs d'activités de l'industrie canadienne de l'assurance automobile. Nous élaborons également un modèle pour les pertes survenues mais non déclarées (IBNR) marginales de chaque secteur d'activité, modèle fondé sur les modèles linéaires généralisés

velopment year and accident year effect factors along with an autoregressive feature for residuals enable modeling the dependence between the various entries of the IBNR loss triangles in a given line of business. Capital requirements calculations are then performed through simulation; numbers obtained with univariate and multivariate risk measures are compared. Moreover, a risk adjustment for non-financial risk required by IFRS 17 is also computed through a cost of capital approach.

doubles. Des facteurs d'effet de l'année de développement et de l'année de l'accident et une fonction autorégressive pour les résidus permettent de modéliser la dépendance entre les diverses inscriptions des triangles de pertes IBNR pour un secteur d'activité donné. Nous effectuons ensuite par simulation le calcul des exigences de capital ; puis nous comparons les résultats obtenus pour des mesures de risques à une et plusieurs variables. Enfin, nous calculons l'ajustement du risque non financier exigé par l'IFRS 17 par l'approche du coût du capital.

[Monday May 27/lundi 27 mai, 16:30-16:45]

Yunran Wei (University of Waterloo) , **Ruodu Wang** (University of Waterloo)

Risk Functionals with Convex Level Sets

Fonctions de risque et ensembles de niveaux convexes

We analyze the convex level sets (CxLS) property of risk functionals, which is a necessary condition for the notions of elicitability, popular in recent statistics and risk management literature. We put the CxLS property in the context of multi- dimensional risk functionals with a special focus on signed Choquet integrals. We obtain two main analytical results in dimension one and dimension two, by characterizing the CxLS property of all one-dimensional signed Choquet integrals, and that of all two-dimensional signed Choquet integrals with a quantile component. Using these results, we proceed to show that a co-monotonic-additive coherent risk measure is co-elicitable with a Value-at-Risk if and only if it is a convex combination of the mean and the corresponding expected shortfall. The new findings generalize several results in the recent literature and partially answer an open question on the characterization of multi-dimensional elicitability.

Nous analysons la propriété des ensembles de niveaux convexes des fonctions de risques, ce qui est une condition nécessaire pour les notions d'élicitabilité, populaires dans les statistiques et la littérature de gestion du risque récentes. Nous plaçons la propriété des ensembles de niveaux convexes dans le contexte des fonctions de risque multidimensionnels, tout en portant une attention particulière aux intégrales signées de Choquet. Nous obtenons deux résultats analytiques principaux en dimensions un et deux, en caractérisant la propriété ensembles de niveaux convexes de toutes les intégrales signées de Choquet unidimensionnelles, et celle de toutes les intégrales de Choquet bidimensionnelles avec une composante quantile. À partir de ces résultats, nous montrons qu'une mesure de risque cohérente, comonotone et additive est coélicitable avec une valeur à risque si et seulement si elle est une combinaison convexe de la moyenne et de l'écart prévu correspondant. Ces nouvelles découvertes généralisent plusieurs résultats décrits dans la littérature récente et répondent en partie à une question ouverte sur la caractérisation de l'élicitabilité multidimensionnelle.

[Monday May 27/lundi 27 mai, 16:45-17:00]

Ihsan Chaoubi (Laval University) , **Hélène Cossette** (Laval University) , **Étienne Marceau** (Laval University)

On Sums of Two Counter-Monotonic Risks

Les sommes de deux risques monotones contraires

In risk management, capital requirements are most often based on risk measurements of the aggregation of multiple risks treated as random variables (RVs). The dependence structure between such RVs has a massive impact on the aggregate loss behavior, which is why it is crucial to investigate. However, such a dependence is often not easily defined or modeled. Thus, in this paper, a boundary analysis of dependence is made for a bivariate portfolio. We focus on the counter-monotonic case, which has been less studied in the literature than the co-

En gestion du risque, les exigences en matière de capital sont le plus souvent basées sur les mesures du risque de l'agrégation des risques multiples traités comme des variables aléatoires (VAs). L'impact de la structure de dépendance entre de telles VAs est colossal sur le comportement de la perte agrégée, d'où la nécessité de l'examiner. Il est toutefois souvent peu facile de définir ou modéliser une telle dépendance. Nous présentons dans cet article une analyse des limites de la dépendance pour un portefeuille bivarié. Nous voyons plus précisément un cas de risque monotone contraire, moins bien documenté que la structure

monotonic dependence structure. We develop closed-form expressions for various risk measures, which allow us to quantify the diversification benefit under such a dependence structure. In addition, sub-exponentially distributed RVs will be used in conjunction with countermonotonic dependence to analyze the impact of the tail distribution on the diversification benefit.

de dépendance comonotone. Nous élaborons des expressions à forme fermée pour divers mesures du risque, ce qui nous permet de quantifier l'avantage de la diversification en présence d'une telle structure de dépendance. De plus, nous utilisons des VAs sous-exponentiellement distribuées en conjonction avec une dépendance monotone contraire pour analyser l'impact de la distribution de queue sur l'avantage de la diversification.

Designed experiments for complex engineered systems
Plans d'expériences pour systèmes techniques complexes

Chair/Président: Ryan Lekivetz

Organizer/Responsable: Ryan Lekivetz

Room/Salle: 122 (ICT)

Abstract/Résumé

[Monday May 27/lundi 27 mai, 15:30-16:00]

Joseph Morgan (SAS Institute)

Covering Arrays: A Tool for Testing Complex Engineered Systems

Tableaux de couverture : un outil pour tester des systèmes complexes en ingénierie

Covering arrays have been widely embraced by test engineers as a tool to derive test cases for complex engineered systems. The phrase "combinatorial testing" is now used to describe this testing approach and it has proven to be a cost-efficient way to determine test cases that are highly effective at identifying faults in such complex systems. In this talk we provide an overview of covering arrays, show how they are connected to several constructs used in experimental design, discuss the challenge that test engineers face in evaluating complex engineered systems, and illustrate how covering arrays may be used to address this challenge.

Les tableaux de couverture ont été largement utilisés par les ingénieurs chargés des tests comme outils pour obtenir des cas de tests pour des systèmes complexes. L'expression « test combinatoire » est maintenant utilisée pour décrire cette manière de tester qui s'est révélée être une des façons les plus économiques de déterminer des cas de tests qui sont très efficaces pour identifier les défauts dans des systèmes aussi complexes. Dans cet exposé, nous offrons une présentation des tableaux de couverture, nous démontrons comment ils sont liés à plusieurs notions en conception expérimentale, nous discutons des défis que les ingénieurs chargés des tests doivent affronter lors de l'évaluation des systèmes complexes en ingénierie et nous illustrons comment les tableaux de couverture peuvent être utilisés pour faire face à ces défis.

[Monday May 27/lundi 27 mai, 16:00-16:30]

Ryan Lekivetz (SAS Institute) , **Joseph Morgan** (SAS Institute)

Design and Analysis of Covering Arrays Using Prior Information

Conception et analyse de tableaux de couverture en utilisant de l'information préalable

Covering arrays are increasingly being used by test engineers to derive test cases to test complex engineered systems. This approach to testing is known as combinatorial testing and has proven to be a cost-efficient way to determine test cases that are highly effective at identifying faults in the system due to the combination of several inputs. However, when such faults are encountered, and failures occur, the test engineer is tasked with determining the inputs and associated values that triggered the failures. This talk addresses this issue by considering the prior knowledge about the system under test (SUT) that is often held by test engineers. We discuss how this prior knowledge can be used to construct covering arrays and how it can also be used to evaluate the effectiveness of covering arrays before any test cases are executed. We then discuss how this prior knowledge

Des tableaux de couverture sont de plus en plus utilisés par les ingénieurs chargés des tests pour obtenir des cas de tests pour des systèmes complexes en ingénierie. Cette façon de tester est connue sous le nom de test combinatoire et s'est avérée être un moyen économique pour déterminer les cas de test qui sont très efficaces pour identifier les défauts du système qui sont dus à la combinaison de plusieurs entrées. Par contre, quand de tels défauts sont rencontrés et que des défaillances se produisent, l'ingénieur chargé du test doit déterminer les entrées et les valeurs associées qui ont causé la défaillance. Cet exposé aborde ce problème en examinant la connaissance préalable du système testé (ST) qui est souvent détenue par les ingénieurs chargés des tests. Nous discutons de la façon dont cette connaissance préalable peut être utilisée pour construire des tableaux de couverture et comment elle peut aussi être utilisée pour évaluer l'efficacité de ces tableaux avant qu'un cas de test ne soit exécuté. Nous discutons ensuite de la

can be used to analyze the outcomes of a set of test cases executed on the SUT when failures occur.

façon dont cette connaissance peut être utilisée pour analyser les résultats d'un ensemble de cas de tests effectués sur le ST lorsqu'il y a une défaillance.

[Monday May 27/lundi 27 mai, 16:30-17:00]

Karen Meagher (University of Regina)

Covering Arrays and Pure Math

Tableaux de couverture et les mathématiques pures

In this talk I will focus on covering arrays from a pure math perspective. I will start with looking at some constructions of covering arrays and some conjectures about the structure of optimal covering arrays. This will include some recent new bounds on small covering arrays. I will also look a refinement of covering arrays that adds a graph structure to the array. I will give an overview of how this is related to some very interesting mathematics. I will also look at connections between covering arrays and extremal combinatorics.

Dans cet exposé, je me concentrerai sur les tableaux de couverture d'une perspective purement mathématique. Je commencerai par examiner quelques constructions de tableaux de couverture et quelques conjectures sur la structure des tableaux de couverture optimale. Ceci inclura quelques nouvelles limites récentes sur les petits tableaux de couverture. Je vais également examiner un raffinement des tableaux de couverture qui ajoute une structure graphique au tableau. Je donnerai un aperçu de la façon dont cette structure est liée à des mathématiques très intéressantes. J'examinerai également les liens entre les tableaux de couverture et la combinatoire extrémale.

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

Chair/Président: Brian Allen

Organizer/Responsable: Jack Gambino

Room/Salle: 148 (ST)

Abstract/Résumé

[Tuesday May 28/mardi 28 mai, 08:30-09:50]

Douglas Wiens (University of Alberta)

Robustness of Design: A Survey

Robustesse du plan : une étude

When an experiment is conducted for purposes that include fitting a particular model to the data, then the 'optimal' experimental design is highly dependent upon the model assumptions, including linearity of the response function, independence and homoscedasticity of the errors. When these assumptions are violated the design can be far from optimal, and so a more robust approach is called for. We should seek a design which behaves reasonably well over a large class of plausible models. I will review the progress which has been made on such problems, in a variety of experimental and modelling scenarios, including prediction, extrapolation, discrimination, survey sampling, dose-response, and machine learning.

Lorsqu'une expérience est menée à des fins qui comprennent l'ajustement d'un modèle particulier aux données, le plan expérimental ≪ optimal ≫ dépend fortement des hypothèses du modèle, ainsi que de la linéarité de la fonction de réponse, de l'indépendance et l'homoscédasticité des erreurs. Lorsque ces hypothèses ne sont pas respectées, le plan peut être loin d'être optimal et une approche plus robuste s'impose. Nous devrions rechercher un plan qui se comporte raisonnablement bien sur une grande classe de modèles plausibles. Je passerai en revue les progrès qui ont été réalisés à l'égard de ces problèmes, dans divers scénarios expérimentaux et de modélisation, ainsi que la prédiction, l'extrapolation, la discrimination, l'échantillonnage d'enquête, la relation dose-réponse et l'apprentissage machine.

Advances in spatial epidemiology and ecology
Avancées en épidémiologie et écologie spatiales

Chair/Président: Rob Deardon

Organizer/Responsable: Rob Deardon

Room/Salle: 102 (ICT)

Abstract/Résumé

[Tuesday May 28/mardi 28 mai, 10:20-10:50]

Andrew Lawson (Medical University of South Carolina)

Bayesian Spatial Modeling for Prospective ID Surveillance: With Application to Seasonal Influenza

Modélisation spatiale bayésienne pour la surveillance prospective des maladies infectieuses, avec application à la grippe saisonnière

Bayesian modeling of infectious disease behavior can be carried out at a variety of scales. Commonly, aggregation leads to counts of infectives available in small areas over time periods. In that case, modeling of the disease dynamic must rely on count models. SIR or SEIR models are often assumed, and dependence on previous counts or counts in neighborhoods is often assumed. In prospective surveillance the focus is on prediction of future events so that detection of change is carried out as early as possible. A Bayesian space-time (ST) model can be formulated to allow for detection of changes. In this talk there is two foci: First, I propose the modeling of an ST probability surface of spread potential that is estimated prospectively with the underlying Bayesian SIR model. This surface can be used to direct interventions. Second I will address the issue of endemic and epidemic model components and whether single component model are better in prospective surveillance.

La modélisation bayésienne du comportement des maladies infectieuses peut être effectuée à des échelles diverses. Couramment, l'agrégation permet de disposer de nombres d'infectants dans des régions restreintes pendant certains laps de temps. Dans ce cas, la modélisation de la dynamique d'une maladie doit faire appel à des modèles de comptage. À cette fin, on se fie souvent aux modèles SIR ou SEIR, tout en dépendant de comptages antérieurs ou en cours dans les voisinages. La surveillance prospective met l'accent sur la prévision d'événements futurs de façon à détecter le plus tôt possible les changements. Un modèle espace-temps bayésien peut être formulé pour permettre la détection des changements. Notre allocution porte sur deux points. Je propose tout d'abord la modélisation d'une probabilité de surface ST du potentiel de dissémination dont l'estimation prospective s'appuie sur le modèle SIR bayésien sous-jacent. Cette surface peut servir à des interventions directes. J'aborde ensuite le problème des composantes endémiques et épidémiques du modèle, tout en demandant s'il vaut mieux utiliser un modèle à une seule composante pour la surveillance prospective.

[Tuesday May 28/mardi 28 mai, 10:50-11:20]

Joanna Elizabeth Mills Flemming (Dalhousie University)

New Approaches for Estimating Population Size for Marine Species

Nouvelles approches pour l'estimation de la taille des populations d'espèces marines

The problem of estimating population size has a long history in statistical ecology. However, the abundance and productivity of highly valuable, severely depleted species remain difficult to assess with standard models. By taking advantage of modern genetics, a new way to estimate abundance (and other key parameters such as mortality rates), the close-kin mark-recapture (CKMR) method, has recently been proposed. It only requires small pieces of tissue, taken from either live or dead animals, and generalizes the standard mark-recapture (MR)

Le problème de l'estimation de la taille de la population a une longue histoire en écologie statistique. Cependant, l'abondance et la productivité d'espèces d'une grande valeur et en grand déclin restent difficiles à évaluer avec les modèles standards. En tirant parti de la génétique moderne, on a récemment proposé une nouvelle méthode d'estimation de l'abondance (et d'autres paramètres clés, comme les taux de mortalité), la méthode de marquage-recapture de parents proches. Elle ne nécessite que de petits morceaux de tissus, prélevés sur des animaux vivants ou morts, et généralise l'approche standard de marquage-recapture pour utili-

approach to use the resulting DNA marks to obtain information about relatedness among individuals in the sample. Here we compare CKMR and MR estimates of population size for brook trout populations and speak to the potential for CKMR going forward.

ser les marques d'ADN obtenues afin d'obtenir de l'information sur la parenté entre les individus dans l'échantillon. Nous comparons les estimations de la taille des populations d'omble de fontaine par marquage-recapture de parents proches et par marquage-recapture et nous parlons du potentiel du marquage-recapture de parents proches pour l'avenir.

[Tuesday May 28/mardi 28 mai, 11:20-11:50]

Md Mahsin (University of Calgary) , **Rob Deardon** (University of Calgary) , **Patrick Brown** (University of Toronto)

A New Class of Spatiotemporal Individual-Level Models for Infectious Diseases Transmission

Nouvelle catégorie de modèles spatio-temporels au niveau de l'individu en matière de transmission de maladies infectieuses

Modeling of infectious diseases has been increasingly used to evaluate the potential impact of different control measures and to guide public health policy decisions. In recent years, individual-level models (ILMs) have been effectively used to model infectious disease transmission. These models are well developed but assume the probability of disease transmission between two individuals depends only on their spatial (or network-based) separation. In this study, we extend ILMs to geographically-dependent ILMs (GD-ILMs) that allow the evaluation of the effect of spatially varying risk factors, environmental factors, as well as unobserved spatial structure, upon the transmission of infectious disease. We consider a conditional autoregressive model to capture the effects of unobserved spatially structured latent covariates or measurement error. We show how GD-ILMs can be fitted to data on both simulation and Alberta influenza outbreaks epidemic within a Bayesian statistical framework.

La modélisation des maladies infectieuses est de plus en plus utilisée pour évaluer l'impact potentiel de diverses mesures de contrôle et orienter les décisions en matière de politique de santé publique. Depuis quelques années, des modèles au niveau de l'individu (MNIs) ont été efficacement utilisés pour modéliser la transmission des maladies infectieuses. Même s'il sont bien élaborés, ces modèles présument que la probabilité d'une transmission de la maladie entre deux individus dépend seulement de leur séparation dans l'espace (ou basée sur le réseau). Cette étude étend les MNIs à des versions MNIs géographiquement dépendantes (MNI-GDs) qui permettent l'évaluation de facteurs de risque variables dans l'espace, de facteurs environnementaux et d'une structure spatiale non observée, au moment de la transmission d'une maladie infectieuse. Nous examinons un modèle autorégressif conditionnel pour capturer les effets de covariables latentes spatialement structurées ou de l'erreur de mesure. Nous montrons comment les MNI-GDs peuvent être ajustés à des données de simulation ou relatives aux épidémies de grippe en Alberta dans un cadre statistique bayésien.

Modern methods in functional data analysis
Méthodes modernes pour l'analyse de données fonctionnelles

Chair/Président: Joel A. Dubin

Organizer/Responsable: Joel A. Dubin

Room/Salle: 116 (ICT)

Abstract/Résumé

[Tuesday May 28/mardi 28 mai, 10:20-10:50]

Hans-Georg Müller (University of California, Davis) , **Xiongtao Dai** (Iowa State University)

Nonparametric Modeling of Longitudinal Compositional and Functional Data on Riemannian Manifolds

Modélisation non paramétrique des données fonctionnelles et compositionnelles longitudinales sur les variétés riemanniennes

Longitudinal compositional data are commonly encountered in longitudinal behavioral studies or metabolomics. Such data exhibit dependencies over time as well as among the p components of the compositional vectors, which are constrained to be non-negative and to sum to one, and can be represented as trajectories on the positive quadrant of a (p-1)-dimensional sphere. This motivates the study of functional data with trajectories on smooth Riemannian manifolds and of Riemannian functional principal component analysis. Theory and illustrations will be presented for an approach where one maps the trajectories at each fixed time on a suitable tangent space.

On trouve souvent des données longitudinales compositionnelles dans les études comportementales longitudinales ou dans la métabolomique. Ces données présentent des dépendances dans le temps, ainsi qu'entre les composantes p des vecteurs de composition, qui sont contraintes d'être non négatives et d'avoir pour somme 1, et peuvent être représentées comme des trajectoires sur le quadrant positif d'une sphère dimensionnelle (p-1). Ceci motive l'étude des données fonctionnelles avec des trajectoires sur des variétés riemanniennes lisses et l'analyse des principales composantes fonctionnelles de Riemann. Nous présenterons une théorie et des illustrations concernant une approche où l'on cartographie les trajectoires à chaque temps fixe sur un espace tangent approprié.

[Tuesday May 28/mardi 28 mai, 10:50-11:20]

Peijun Sang (University of Waterloo) , **Jiguo Cao** (Simon Fraser University)

Distance-Weighted Discrimination for Functional Data

Discrimination pondérée par la distance pour données fonctionnelles

Due to the curse of dimensionality, it is difficult to apply conventional classifiers for multivariate data to classify functional data. Dimension reduction is imperative for addressing this problem. With the aid of functional principal component analysis (FPCA), various Bayes classifiers built on finite-dimensional FPC scores have been proposed to classify functional data. However, some restrictive assumptions in terms of mean and/or covariance functions are imposed on functional data to construct these Bayes classifiers. We propose a margin-based classifier called distance-weighted discrimination (DWD) by employing FPC scores. The proposed classifier is free of those restrictive assumptions associated with the Bayes classifiers. Moreover, we theoretically establish the Bayes risk consistency of the DWD classifier under the scenario when functional data are ob-

À cause de la malédiction de la dimension, il est difficile d'appliquer les classificateurs traditionnels aux données multivariées pour classer les données fonctionnelles. La réduction de la dimension est essentielle pour résoudre ce problème. À l'aide de l'analyse de composante principale fonctionnelle (ACPF), différents classificateurs bayésiens fondés sur les scores CPF de dimension fini ont été proposés pour classer les données fonctionnelles. Par contre, certaines hypothèses restrictives en termes de fonctions de covariance et/ou de moyenne sont imposées aux données fonctionnelles pour construire ces classificateurs bayésiens. Nous proposons un classificateur fondé sur la marge nommé discrimination pondérée par la distance (DPD) en utilisant les scores CPF. Le classificateur proposé n'a aucune de ces hypothèses restrictives associées aux classificateurs bayésiens. De plus, nous avons établi théoriquement la convergence du risque bayésien du classificateur DPD dans le cas où les données fonctionnelles sont observées sur une grille de

served at a dense grid of points and measured with random noises.

points dense et mesurées avec du bruit aléatoire.

[Tuesday May 28/mardi 28 mai, 11:20-11:50]

Jane-Ling Wang (University of California, Davis) , **Xiaoke Zhang** (George Washington University)

Varying-Coefficient Additive Models: Two Birds with One Stone?

Modèles additifs à coefficients variables : d'une pierre deux coups ?

Both varying-coefficient and additive models have been widely adopted as non-parametric modeling approaches that enjoy flexibility and parsimony. An intriguing question is how to choose between these two models in practice. In this talk, we show how to extend both models into the varying-coefficient additive model (VCAM) that allows for sparsely observed functional responses, a.k.a. longitudinal data, and longitudinal covariates, in addition to vector covariates. A new algorithm is proposed and its performance is demonstrated through simulations and data applications. The algorithm involves non-convex maximization so the choice of the initial estimates plays a crucial role and we discuss several options and their empirical performance. Theoretical results are established for the nonparametric component functions of the model, including rates of convergence. Future directions will be discussed.

Les modèles à coefficients variables et les modèles additifs sont des approches de modélisation populaires souples et parcimonieux. Une question qui intrigue est celle de savoir choisir en pratique entre ces deux modèles. Dans cette présentation, nous montrons comment les étendre en un modèle additif à coefficients variables (MACV) adapté aux réponses fonctionnelles éparses, ou données longitudinales, et aux covariables longitudinales, ainsi qu'aux covariables vecteurs. Nous proposons un nouvel algorithme et en montrons la performance par des simulations et des applications sur des données. L'algorithme implique une maximisation non convexe, et donc le choix des estimations initiale joue un rôle essentiel ; nous discutons de plusieurs options et de leur performance empirique. Nous déterminons des résultats théoriques pour les fonctions de composantes non paramétriques du modèle, y compris les taux de convergence, et discutons des orientations futures.

A Showcase of Student Research from the CANSSI CRT 'Joint Analysis of Neuroimaging Data: High-Dimensional Problems, Spatiotemporal Models and Computation'
Recherches d'étudiants du PRC de l'INCASS "Analyse conjointe de données de la neuroimagerie : problèmes en grande dimension, modèles spatiotemporels et calculs"

Chair/Président: Farouk Nathoo

Organizer/Responsable: Farouk Nathoo

Room/Salle: 146 (SB)

Abstract/Résumé

[Tuesday May 28/mardi 28 mai, 10:20-10:50]

Yin Song (University of Victoria) , **Farouk Nathoo** (University of Victoria) , **Arif Babul** (University of Victoria)

A Potts-Mixture Spatiotemporal Joint Model for Combined MEG and EEG Data

Un modèle de mélange Potts spatio-temporel conjoint pour données combinées MEG et EEG

We develop a new methodology for determining the location and dynamics of brain activity from combined MEG and EEG data. The resulting inverse problem is ill-posed and is one of the most difficult problems in neuroimaging data analysis. In our development we propose a solution that combines the data from three different modalities: MRI, MEG, and EEG. We propose a new Bayesian spatial finite mixture model that builds on the mesostate-space model developed by Daunizeau and Friston (2007). Our new model incorporates two major extensions: (i) We combine EEG and MEG data together and formulate a joint model for dealing with the two modalities simultaneously; (ii) we incorporate the Potts model to represent the spatial dependence in an allocation process that partitions the cortical surface into a small number of latent states termed mesostates. We formulate the new spatio-temporal model and derive an efficient procedure for simultaneous point estimation and model selection.

Nous développons une nouvelle méthodologie pour déterminer l'emplacement et la dynamique de l'activité cérébrale à partir de données combinées MEG et EEG. Le problème inverse qui en résulte est mal posé et un des problèmes les plus compliqués dans l'analyse de données en neuroimagerie. Nous proposons une solution qui combine les données de trois différentes modalités : IRM, MEG et EEG. Nous proposons un nouveau modèle spatial bayésien de mélange fini qui tire partie du modèle spatial de méso-états développé par Daunizeau et Friston (2007). Notre nouveau modèle comprend deux extensions majeures : (i) nous combinons les données EEG et MEG et nous élaborons un modèle conjoint pour traiter les deux modalités simultanément ; (ii) nous intégrons le modèle Potts pour représenter la dépendance spatiale dans un processus de répartition qui partage la surface corticale en un petit nombre d'états latents nommés méso-états. Nous élaborons le nouveau modèle spatio-temporel et nous obtenons une procédure efficace pour une estimation ponctuelle simultanée et une sélection du modèle.

[Tuesday May 28/mardi 28 mai, 10:50-11:20]

Farouk Nathoo (University of Victoria), **Yunlong Nie** (Simon Fraser University)

Spectral Dynamic Causal Modelling of Resting-State fMRI: Relating Effective Brain Connectivity in the Default Mode Network to Genetics

Modélisation causale à dynamique spectrale de l'IRMf au repos : faire le pont de la génétique à la connectivité cérébrale efficace dans le réseau du mode par défaut

We conduct a novel imaging genetics study of the Alzheimer's Disease Neuroimaging Initiative based on resting-state fMRI (rs-fMRI) and genetic data obtained from 112 subjects, where each subject is classified as cognitively normal (CN), as having mild cognitive impairment (MCI), or as having Alzheimer's Disease (AD). A Dynamic Causal Model (DCM) is fit to the rs-fMRI time series in order to estimate a directed network representing effective brain connectivity within the de-

Nous menons une nouvelle étude d'imagerie génétique portant sur l'Alzheimer's Disease Neuroimaging Initiative (ADNI) basée sur des IRMf au repos et des données génétiques recueillies de 112 sujets, chacun ayant été classifié comme ≪cognitivement normal≫ (CN), souffrant d'un déficit cognitif léger (DCL) ou de la maladie d'Alzheimer. Un modèle causal dynamique (MCD) est adapté aux séries chronologiques de l'IRMf au repos dans le but d'estimer un réseau dirigé représentant la connectivité cérébrale efficace dans le réseau du mode par défaut (RMPD), un réseau

fault mode network (DMN), a key network commonly known to be active when the brain is at rest. These networks are then related to genetic data and Alzheimer's disease in the first genome-wide association study to use DCM as a neuroimaging phenotype. Our proposed pipeline is comprised of four sequential analyses linked together with the objective of shedding light on the relationship between brain connectivity and genetics in relation to disease.

essentiel bien connu pour s'activer lorsque le cerveau est en état de repos. Ces réseaux sont ensuite reliés aux données génétiques et à la maladie d'Alzheimer dans la première étude d'association pangénomique à employer un MCD en guise de phénotype de neuroimagerie. Le pipeline que nous proposons est composé de quatre analyses séquentielles reliées ensemble avec comme objectif d'éclairer la relation entre la connectivité cérébrale et la génétique en lien avec la maladie.

[Tuesday May 28/mardi 28 mai, 11:20-11:50]

Eugene Opoku (University of Victoria) , **Farouk Nathoo** (University of Victoria) , **Ejaz Syed Ahmed** (Brock University)

Ant Colony System Optimization for Estimation in Spatial Hidden Markov Models

Optimisation du système de colonie de fourmis pour l'estimation dans un modèle de Markov caché spatial

Hidden Markov models incorporating the Potts model for the labelling process are an important class of models in spatial statistics that have been applied widely. Jointly estimating the model parameters and the pixel labels by maximizing the posterior distribution is recognized as a difficult combinatorial optimization problem that is commonly approached using the simple iterated conditional modes algorithm (ICM; Besag, 1986). We consider this estimation problem within the context of a Gaussian mixture model incorporating labels based on the Potts model to represent spatial dependence among neighboring pixels in a 2-dimensional image. Our primary goal is to introduce the ACS algorithm, an algorithm which has seen application in areas outside of statistics but is relatively unknown within the statistical computing literature where ICM or MCMC algorithms are most commonly used for estimation in the hidden Potts model.

Les modèles de Markov cachés qui intègrent le modèle de Potts pour le processus d'étiquetage sont une classe de modèles importants en statistique spatiale, avec de nombreuses applications. L'estimation conjointe des paramètres du modèle et des étiquettes de pixel par une maximisation de la distribution postérieure est reconnue comme un problème d'optimisation combinatoire difficile qui est communément abordé à l'aide de l'algorithme simple du mode conditionnel itéré (MCI ; Besag, 1986). Nous étudions ce problème d'estimation dans le contexte d'un modèle de mélange gaussien qui inclut des étiquettes sur la base du modèle de Potts pour représenter la dépendance spatiale entre pixels voisins dans une image 2D. Notre objectif principal est de présenter l'algorithme ≪ système de colonie de fourmis ≫, un algorithme qui s'applique à des domaines non statistiques mais relativement peu connu en informatique statistique où les algorithmes MCI ou MCCM sont plus communément utilisés pour l'estimation dans le modèle de Potts caché.

Effective Implementation of Statistics Capstone Courses
Mise en place effective de cours finaux en statistique

Chair/Président: Asokan Mulayath Variyath

Organizer/Responsable: Asokan Mulayath Variyath

Room/Salle: 109 (SS)

Abstract/Résumé

[Tuesday May 28/mardi 28 mai, 10:20-10:50]

Gemai Chen (University of Calgary)

Capstone Course: What For?

Cours de synthèse : pour quoi ?

At University of Calgary, we have been offering a capstone course called Practice of Statistics for more than a dozen years. In this talk, we want to share our experiences and encourage more departments to offer a similar course.

À l'Université de Calgary, nous offrons un cours de synthèse appelé « Pratique de la statistique » depuis plus d'une douzaine d'années. Dans cet exposé, nous voulons partager notre expérience et encourager plus de départements à offrir un cours similaire.

[Tuesday May 28/mardi 28 mai, 10:50-11:20]

Karen Buro (MacEwan University)

A Course Engaging Undergraduates in Statistical Consultation

Un cours qui incite les étudiants de premier cycle à la consultation statistique

Students in the Applied Statistics program at MacEwan University are required to complete a course on statistical consulting where they gain experience as consultants in ongoing research projects working with real data. In seminars, students discuss topics like the scientific method, effective communication, ethical conduct as a consultant, and relevant statistical methods, while tuning their consultation skills via case studies. The centerpiece of the program is the consultation projects students engage in. Students consult with researchers from various disciplines in research projects, conduct appropriate statistical analyses, and communicate their findings in oral and written reports. We will share our approach and challenges in teaching this course.

Les étudiants du programme de statistique appliquée de l'université MacEwan doivent suivre un cours de consultation statistique dans le cadre duquel ils acquièrent de l'expérience à titre de consultants dans des projets de recherche en cours avec des données réelles. Lors de séminaires, les étudiants discutent de sujets, comme les méthodes scientifiques, la communication efficace, la conduite éthique en tant que consultant et les méthodes statistiques pertinentes, tout en affinant leurs compétences en consultation au moyen d'études de cas. Les éléments centraux du programme sont les projets de consultation dans lesquels les étudiants s'engagent. Les étudiants consultent des chercheurs de diverses disciplines dans le cadre de projets de recherche, effectuent des analyses statistiques appropriées et communiquent leurs conclusions dans des rapports oraux et écrits. Nous vous ferons part de notre approche et des défis concernant l'enseignement de ce cours.

[Tuesday May 28/mardi 28 mai, 11:20-11:50]

Gabriela Cohen Freue (University of British Columbia)

Case Studies and Consulting in Statistics

Études de cas et consultation en statistique

Knowing a collection of methods does not totally reflect the actual needs that data analysts face to date. As data science grows and more complex datasets are available, collaborative and interdisciplinary work becomes

Connaître une foule de méthodes ne règle pas complètement les problèmes que les analystes de données doivent résoudre. Plus les sciences des données se développent et plus les jeux de données sont complexes. Il devient donc impératif de travailler

Effective Implementation of Statistics Capstone Courses
Mise en place effective de cours finaux en statistique

imperative. Thus, I have developed a new model between an undergraduate course ("Case Studies in Statistics") and a graduate course ("Techniques of Statistical Consulting") at UBC to address these needs. Students work collaboratively on a real case study brought by researchers from different disciplines. Undergraduate students analyze the data under the supervision of graduate students. Students, instructors, and data owners share a GitHub repository that hosts computational codes, related papers, reports, and discussions, thus promoting reproducibility. Overall, this model stimulates relevance of statistical concepts, effective communication, productive collaborative work, and dynamic learning. In this talk, I will discuss the benefits and challenges of this model.

en collaboration et de façon interdisciplinaire. Par conséquent, j'ai mis au point un nouveau modèle croisant un cours de premier cycle (≪études de cas en statistiques≫) et un cours de cycle supérieur (≪techniques de consultation en statistiques≫) à l'Université de la Colombie-Britannique pour combler ces besoins. Les étudiants travaillent en collaboration sur une étude de cas réelle fournie par des chercheurs provenant de différentes disciplines. Les étudiants de premier cycle analysent les données sous la supervision d'étudiants de cycle supérieur. Les étudiants, les professeurs et les propriétaires des données partagent un dépôt GitHub qui contient des codes informatiques, des documents pertinents, des rapports et des discussions, ce qui encourage la reproductibilité. En général, ce modèle souligne l'importance des concepts statistiques, d'une communication efficace, d'un travail collaboratif productif et de l'apprentissage dynamique. Au cours de cet exposé, j'examinerai les avantages et les défis relatifs à ce modèle.

Recent advances in statistical inference for complex data structures
Dernières avancées en inférence statistique pour les structures de données complexes (commandité par le chapitre canadien de l'ICSA)

Chair/Président: Liqun Wang

Organizer/Responsable: Liqun Wang

Room/Salle: 142 (AD)

Abstract/Résumé

[Tuesday May 28/mardi 28 mai, 10:20-10:50]

Christopher M. Manuel (Texas AM University)

Matched Case-Control Data with a Misclassified Exposure: What Can Be Done with Instrumental Variables?

Données de comparaison avec les témoins appariés accompagnées d'une classification erronée de l'exposition : que faire avec des variables instrumentales ?

I will describe how to reduce the bias in the log-odds ratio parameter due to misclassification of an exposure variable in a matched case-control study. Particular attention will be given to the situation when there is no validation dataset to assess the misclassification probability. I will discuss two approaches to reduce the bias; both approaches use instrumental variables to extract information on the misclassification probabilities. The usefulness of the methodologies will be demonstrated through simulation studies and analyses of a real dataset.

Dans le cadre d'une étude de comparaison avec les témoins appariés, je décrirai comment réduire le biais du paramètre log-rapport de cote en raison d'une classification erronée d'une variable d'exposition. Une attention toute particulière est accordée aux situations où il n'y a pas de données de validation pour évaluer la probabilité de classification erronée. J'aborderai deux approches pour réduire le biais ; chacune se sert de variables instrumentales pour tirer de l'information relative aux probabilités de classification erronée. Je démontrerai l'utilité des méthodologies à partir d'études de simulation et d'analyses de données réelles.

[Tuesday May 28/mardi 28 mai, 10:50-11:20]

Mahmoud Torabi (University of Manitoba) , **Vahid Tadayon** (Higher Education Center of Eghlid, Iran)

Measurement Error in Spatial Models with Fat Tails and Skewed Errors

Erreur de mesure dans les modèles spatiaux avec queues épaisses et erreurs asymétriques

Spatial models have been widely used in public health research. In the case of continuous outcomes, the traditional approaches to model spatial data are based on the Gaussian distribution. The real data could be highly non-Gaussian and may show features like heavy tails and/or skewness. In spatial data modeling, it is also commonly assumed that the covariates are observed without errors, but for various reasons, such as measurement techniques or instruments used, uncertainty is inherent in spatial (especially geostatistics) data, and so, these data are susceptible to measurement errors in the covariates of interest. In this paper, using a likelihood approach, we introduce a general class of spatial models with covariate measurement error that can account for heavy tails, skewness, and uncertainty of the covariates. The proposed approach is evaluated through a simulation study and also by a real data application.

Les modèles spatiaux sont communément utilisés en santé publique. Dans le cas de résultats continus, les approches traditionnelles de la modélisation se basent sur la distribution gaussienne. Or les données réelles sont parfois non gaussiennes ou présentent des caractéristiques comme des queues épaisses et/ou une asymétrie. En modélisation de données spatiales, il est également souvent supposé que les covariables sont observées sans erreurs ; or pour diverses raisons liées aux techniques ou aux outils de mesure employés, l'incertitude est inhérente aux données spatiales (notamment géostatisques), si bien que ces données sont sujettes à des erreurs de mesure dans les covariables d'intérêt. Dans cette présentation, avec une approche de vraisemblance, nous introduisons une classe générale de modèles spatiaux avec erreur de mesure des covariables qui tient compte des queues épaisses, de l'asymétrie et de l'incertitude des covariables. Nous évaluons l'approche proposée via une étude de simulation et par une application sur données réelles.

[Tuesday May 28/mardi 28 mai, 11:20-11:50]

Recent advances in statistical inference for complex data structures
Dernières avancées en inférence statistique pour les structures de données complexes (commandité par le chapitre canadien de l'ICSA)

Zhou Zhou (University of Toronto) , **Weichi Wu** (Tsinghua University)

MACE: Multiscale Abrupt Change Estimation under Complex Temporal Dynamics

EMCA : Estimation multiéchelle de changement abrupt en fonction d'une dynamique temporelle complexe

We consider the problem of detecting abrupt changes in trend whereas the covariance and higher-order structures of the system can experience both smooth and abrupt changes in time. The number of jump points is allowed to diverge to infinity with the jump sizes possibly shrinking to zero. The method is based on a multiscale application of an optimal jump-pass filter to the time series, where the scales are dense between admissible lower and upper bounds. The MACE method is shown to be able to detect all possible jump points within an optimal range with a prescribed probability asymptotically. Here the probability can be sample size dependent and converges to 1. Simulations and data analysis show that, under complex dynamics, MACE performs favorably compared with some of the state-of-the-art change point detection methods.

Nous examinons le problème de détection de changements abrupts dans les tendances où la covariance et les structures d'ordre supérieur du système peuvent rencontrer des changements à la fois abrupts et réguliers dans le temps. Le nombre de points de saut peut diverger vers l'infini avec la taille des sauts pouvant être diminuée à zéro. La méthode est basée sur une application multiéchelle d'un filtre passe-saut optimal aux séries temporelles, où les échelles sont denses entre les limites supérieures et inférieures acceptables. La méthode EMCA est reconnue pour détecter tous les points de saut possibles dans une étendue optimale avec une probabilité prescrite asymptotiquement. Dans ce cas-ci, la probabilité peut dépendre de la taille de l'échantillon et converger à 1. Les analyses de simulation et de données démontrent que, dans une dynamique complexe, EMCA fonctionne mieux par rapport à certaines méthodes de pointe pour détecter les points de changements.

Decisions in Finance and Economics
Décisions en matière de finance et d'économie

Chair/Président: François Bellavance

Room/Salle: 143 (ST)

Abstract/Résumé

[Tuesday May 28/mardi 28 mai, 10:20-10:35]

Jean-François Bégin (Simon Fraser University)

Economic Scenario Generator and Parameter Uncertainty: A Bayesian Approach

Générateur de scénarios économiques et incertitude des paramètres : une approache bayésienne

In this presentation, we study parameter uncertainty and its actuarial implications in the context of economic scenario generators. To account for this additional source of uncertainty in a consistent manner, we cast Wilkie's four-factor framework into a Bayesian model. The posterior distribution of the model parameters is estimated using MCMC methods and is used to perform Bayesian predictions on the future values of the inflation rate, the dividend yield, the dividend index return, and the long-term interest rate. According to US data, parameter uncertainty has a significant impact on the dispersion of the four economic variables of Wilkie's framework. The impact of such parameter uncertainty is then assessed for a portfolio of annuities: the right tail of the loss distribution is significantly heavier when parameters are assumed random and when this uncertainty is estimated in a consistent manner.

Dans cette présentation, nous étudions l'incertitude des paramètres et ses implications actuarielles dans le contexte des générateurs de scénarios économiques. Pour tenir compte de cette source supplémentaire d'incertitude de manière cohérente, nous avons converti le modèle de Wilkie, une structure avec quatre facteurs, en un modèle bayésien. La distribution a posteriori des paramètres du modèle est estimée à l'aide des méthodes MCCM et est utilisée pour effectuer des prévisions sur les valeurs futures du taux d'inflation, du taux de dividendes, du rendement de l'indice du dividende et du taux d'intérêt à long terme. En utilisant des données américaines, l'incertitude des paramètres a un impact significatif sur la dispersion des variables économiques. L'impact d'une telle incertitude sur les paramètres est ensuite évalué pour un portefeuille de rentes : l'aile droite de la distribution des pertes est nettement plus lourde lorsque les paramètres sont supposés aléatoires et lorsque cette incertitude est estimée de manière cohérente.

[Tuesday May 28/mardi 28 mai, 10:35-10:50]

Yifan Li (Western University) , **Reg Kulperger** (Western University) , **Hao Yu** (Western University)

A Test of Knightian Uncertainty under the G-Expectation Framework

Un test sur l'incertitude de Knight dans le cadre de la G-espérance

Since the notion of Knightian Uncertainty or Ambiguity was proposed, the importance of uncertainty and the danger of ignoring it have been discussed thoroughly in various areas. Currently, the use of uncertainty mostly comes from subjective belief or experience which is not always justifiable at the start of analyzing a new dataset. However, can one use a data-driven method to decide if uncertainty is a significant feature of the data? We propose a scheme to visualize and test the uncertainty under the G-Expectation Framework, a generalization of the classical probabilistic system allowing us to cover both certain and uncertain scenarios. We theoretically study the scheme to detect mean and variance uncertainty under the lack of information on the underlying data structure. We apply the method to both simulated

Depuis que la notion d'incertitude ou d'ambiguïté de Knight a été présentée, l'importance de l'incertitude ainsi que le danger si elle est ignorée ont fait l'objet de discussions approfondies dans différents domaines. L'utilisation de l'incertitude provient actuellement principalement d'expériences ou de croyances subjectives qui ne sont pas toujours justifiables au début de l'analyse d'un nouvel ensemble de données. Par contre, est-il possible d'utiliser une méthode fondée sur les données pour décider si l'incertitude est une caractéristique importante des données ? Nous présentons un plan pour visualiser et tester l'incertitude dans le cadre de la G-espérance qui est une généralisation du système probabiliste classique qui nous permet de traiter à la fois les scénarios certains et incertains. Nous étudions théoriquement le plan pour détecter l'incertitude de la moyenne et de la variance lorsqu'il y a un manque d'information sur la structure des données sous-jacente. Nous ap-

and real data (stock data) to illustrate its feasibility and capability. Our vision is to provide a preliminary data analysis tool to detect uncertainties for precautions.

pliquons la méthode à des données simulées et réelles (données en stock) pour illustrer sa faisabilité et sa capacité. Notre vision est d'offrir un outil préliminaire d'analyse de données pour détecter certaines incertitudes par précautions.

[Tuesday May 28/mardi 28 mai, 10:50-11:05]

Zhenxian Gong (Western University) , **Marcos Escobar-Anel** (Western University)

The Mean-Reverting 4/2 Stochastic Volatility Model: Properties and Financial Applications.

Modèle de volatilité stochastique 4/2 avec retour à la moyenne : propriétés et applications financières

In this paper, we devise and study properties of a new stochastic process that combines a mean-reverting model with an advance stochastic volatility process: the 4/2 process. This is inspired by the modeling of at least two financial asset classes: commodities and volatility indices. We provide an analytical expression for the generalized conditional characteristic function and study feasible changes of measures, with an aim to price financial products. The empirical analysis and the estimation methodology confirm the need of such a model in several examples from the targeted asset classes. In comparison to particular cases, applications to option pricing corroborate the substantial impact on implied volatility surfaces of this rich model.

Dans cette présentation, nous élaborons et étudions les propriétés d'un nouveau processus stochastique qui combine un modèle avec retour à la moyenne et un processus de volatilité stochastique avancée : le processus 4/2. Ces travaux s'inspirent de la modélisation d'au moins deux classes d'actifs financiers : les produits de base et les indices de volatilité. Nous donnons une expression analytique de la fonction caractéristique conditionnelle généralisée et étudions les changements de mesures faisables, en vue d'une évaluation du prix de produits financiers. L'analyse empirique et la méthode d'estimation confirment le besoin d'un tel modèle pour plusieurs exemples des classes d'actifs ciblées. Pour des cas particuliers, les applications à l'évaluation du prix des options corroborent l'impact notable sur les surfaces de volatilité implicite de ce modèle très riche.

[Tuesday May 28/mardi 28 mai, 11:05-11:20]

Wei-Hsiang Lin (Simon Fraser University) , **Shih-Kuei Lin** (National Chengchi University) , **Cary Chi-Liang Tsai** (Simon Fraser University)

Impact of Interest Rate, Surrender, and Liquidity Risks on the Surplus of a Portfolio of Endowment Policies Using Optimal Portfolio Selection Techniques

Analyse de l'impact du taux d'intérêt, du rachat et des risques d'illiquidité sur l'excédent d'un portefeuille de polices de dotation de fonds à l'aide de techniques de sélection de portefeuille optimal

A life insurer charges an endowment policyholder high premiums from which the policyholder's cash value is built at an interest rate. The life insurer invests the collected premiums in financial securities to meet or exceed the interest rate, and a policyholder can surrender his policy before maturity and get his cash value back subject to a surrender charge. When lots of policyholders surrender their policies, the life insurer needs to liquidate some securities in a short time, which exposes the insurer to liquidity risk. In this paper, we propose a framework to analyse the impact of interest rate, surrender, and liquidity risks on the surplus of a portfolio of endowment policies. Under the framework, we formulate the fair premium and risk-based reserves calculations. In addition, we adopt optimal portfolio selection methods for maximizing utilities. A series of sensitivity analyses are conducted to illustrate the surplus distribu-

Un assureur-vie impose à un titulaire de police de dotation de fonds des primes élevées sur lesquelles est établie la valeur de rachat à un certain taux d'intérêt. L'assureur-vie investit les primes collectées dans des titres financiers pour égaler ou dépasser le taux d'intérêt et l'assuré peut racheter sa police avant l'échéance et en obtenir la valeur de rachat assujettie à des frais de rachat. Lorsque de nombreux assurés rachètent leur police, l'assureur doit liquider certains titres en peu de temps, ce qui l'expose à un risque d'illiquidité. Nous proposons un cadre d'analyse de l'impact du taux d'intérêt, du rachat et des risques d'illiquidité sur l'excédent d'un portefeuille de polices de dotation de fonds. Dans ce cadre de travail, nous formulons des calculs de réserves basées sur la prime équitable et sur le risque. De plus, nous recourons à des techniques de sélection de portefeuille optimal pour maximiser les utilités. Une série d'analyses de sensibilité est menée pour illustrer les répartitions de l'excédent et les utilités correspondantes après l'adoption d'un portefeuille.

tions and corresponding utilities after the adoption.

[Tuesday May 28/mardi 28 mai, 11:20-11:35]

Xing Gu (University of Western Ontario) , **Rogemar Mamon** (Western University) , **Matt Davison** (Western University) , **Hao Yu** (Western University)

An Analysis and Forecasting of Financial Market Liquidity Regimes

Analyse et prévision des régimes de liquidité des marchés financiers

A multivariate hidden Markov model (HMM)-based approach is developed to capture simultaneously the regime-switching dynamics of four financial market indices: the Treasury bill yield-Eurodollar spread (TED), the US Dollar Index (DXY), the VIX and the SP 500 bid-ask spread. These indices are deemed to drive the main characteristics of market liquidity risk. The Ornstein–Uhlenbeck (OU) process and geometric Brownian motion (GBM) are integrated to capture the evolutions of four time series, which mirror liquidity levels in the financial markets. All model parameters are modulated by a discrete-time HMM, and they randomly switch between different liquidity regimes which correspond to states produced by the resultant forces acting on the financial markets. Adaptive multivariate filters are derived via the change of probability measure method. An early-warning signal extraction system is put forward in order to generate alerts for occurrence of illiquidity episodes.

Une approche fondée sur un modèle de Markov caché multivarié est mise au point pour saisir simultanément la dynamique de changement de régime de quatre indices des marchés financiers : le rendement des bons du Trésor et l'écart de l'eurodollar, l'indice du dollar américain, le VIX et l'écart acheteur-vendeur du SP 500. Ces indices sont censés déterminer les principales caractéristiques du risque de liquidité du marché. Le processus d'Ornstein-Uhlenbeck et le mouvement brownien géométrique sont intégrés pour saisir les évolutions de quatre séries temporelles, qui reflètent les niveaux de liquidité sur les marchés financiers. Tous les paramètres du modèle sont modulés par un modèle de Markov caché multivarié à temps discret et basculent aléatoirement entre différents régimes de liquidité qui correspondent aux états produits par les forces résultantes agissant sur les marchés financiers. On obtient des filtres adaptatifs multivariés par la méthode de mesure du changement de probabilité. On propose un système d'extraction de signaux d'alerte précoce afin de générer des alertes pour l'apparition d'épisodes d'illiquidité.

[Tuesday May 28/mardi 28 mai, 11:35-11:50]

Javad Rastegari (Western University) , **Lars Stentoft** (Western University) , **Marcos Escobar-Anel** (Western University)

Option Pricing with Conditional GARCH Models

Établir le prix des options avec les modèles GARCH conditionnels

The Heston-Nandi GARCH model is well-known for providing a closed-form option pricing formula in discrete time setting with Gaussian shocks. We generalize this model to a class of conditional GARCH models under which asset returns over a single period follow a normal variance-mean mixture distribution. We show that the conditional moment generating function is obtained in closed form via a set of recursive equations, and when combined with Monte Carlo simulation, gives an efficient method of option pricing. This class of conditional GARCH models also admits a variance-dependent Radon-Nikodym derivative which allows for pricing the variance risk. As our main example, we examine the case where the mixing variable is driven by a two-state Markov-chain representing the normal and crisis periods in financial markets, and we illustrate the importance of accommodating state dependency and

Le modèle GARCH Heston-Nandi est bien connu pour sa formule à forme fermée pour établir le prix des options dans un cadre de temps discret avec des chocs gaussiens. Nous généralisons ce modèle à une classe de modèles GARCH conditionnels dans lesquels les rendements des actifs d'une seule période suivent une distribution de mélange variance-moyenne normale. Nous démontrons que la fonction génératrice de moments conditionnels est obtenue sous une forme fermée par un ensemble d'équations récursives et que lorsqu'elle est combinée avec une simulation Monte Carlo elle offre une méthode efficace pour fixer le prix des options. La classe des modèles GARCH conditionnels admet aussi une dérivée Radon-Nikodym dépendante de la variance, ce qui permet d'établir le prix du risque de la variance. Dans notre principal exemple nous examinons le cas où la variable de mélange est guidée par une chaîne de Markov à deux états qui représente les périodes de crise et les périodes normales du marché financier et nous illustrons l'importance de tenir compte de la dépendance de

variance risk in option pricing and hedging.

l'état et du risque de la variance dans l'établissement du prix des options et de la couverture de risque.

Methods for High-Dimensional and Large Data I
Méthodes pour traiter les données volumineuses et de grande dimension I

Chair/Président: Whitney K. Huang

Room/Salle: 105 (SB)

Abstract/Résumé

[Tuesday May 28/mardi 28 mai, 10:20-10:35]

Zheng Gao (University of Michigan) , **Stilian Stoev** (University of Michigan)

Fundamental Limits of Exact Support Recovery in High Dimensions

Limites fondamentales du support de redressement exact en haute dimension

We study the estimation of the support (set of non-zero components) of a high-dimensional signal observed with additive and dependent noise. With suitable parameterization of the signal sparsity and magnitude, we characterize a phase-transition phenomenon akin to the signal detection boundary. Namely, when the signal is above the so-called strong classification boundary, thresholding estimators achieve asymptotically perfect support recovery. This is under arbitrary error dependence assumptions, provided that the marginal error distribution has rapidly varying tails. Conversely, under mild dependence conditions on the noise, we show that no thresholding estimators can achieve perfect support recovery when the signal is below the boundary. The proofs of these results exploit a certain concentration of maxima phenomenon known as relative stability. We provide a complete characterization of the relative stability phenomenon for Gaussian arrays in terms their correlation structure.

Nous étudions l'estimation du support (ensemble de composantes non nulles) d'un signal de haute dimension observé avec un bruit additif et dépendant. Grâce à un paramétrage approprié de la faiblesse et de l'amplitude du signal, nous caractérisons un phénomène de transition de phase qui s'apparente à la limite de détection du signal. En d'autres termes, lorsque le signal est au-dessus de la limite dite forte de la classification, les estimateurs de seuillage permettent d'obtenir un redressement asymptotique parfait du support. Cette opération se fait dans le cadre d'hypothèses arbitraires de dépendance à l'égard des erreurs, à condition que la distribution marginale de l'erreur a des queues qui varient rapidement. À l'inverse, dans des conditions d'une dépendance modérée au bruit, nous montrons qu'aucun estimateur de seuillage ne peut atteindre un redressement parfait du support lorsque le signal est en dessous de la limite. Les preuves de ces résultats exploitent une certaine concentration de phénomènes de maxima appelés stabilité relative. Nous fournissons une caractérisation complète du phénomène de stabilité relative des réseaux gaussiens en termes de structure de corrélation.

[Tuesday May 28/mardi 28 mai, 10:35-10:50]

Grace Guan Hsu (Simon Fraser University) , **Derek Bingham** (Simon Fraser University)

Super Fast Emulation and Calibration of Large Computer Experiments

Émulation et étalonnage ultra-rapides de grandes expériences informatiques

Scientific investigations are often expensive and the ability to quickly perform analysis of data on-location at experimental facilities can save valuable resources. Further, computer models that leverage scientific knowledge can be used to gain insight into complex processes and reduce the need for costly physical experiments, but in turn may be computationally expensive to run. We compare multiple statistical surrogates or emulators based on Gaussian processes for expensive computer models, with the goal of producing predictions quickly given large training sets. We then present a modularised approach for finding the values of inputs that allow for the surrogate model to match reality, or field observa-

Pour des recherches scientifiques souvent coûteuses, la possibilité d'analyser rapidement des données sur le site d'expérimentation peut faire économiser des ressources précieuses. Par ailleurs, des modèles informatiques qui exploitent le savoir scientifique peuvent aider à mieux comprendre des processus complexes et réduire le besoin d'expériences physiques coûteuses, mais ces modèles sont souvent coûteux en ressources informatiques. Nous comparons divers substituts ou émulateurs statistiques à processus gaussiens qui permettent d'éviter le recours à ces modèles informatiques coûteux et de produire des prévisions rapidement à partir de grands jeux de données de formation. Nous présentons ensuite une approche modulaire de l'identification de la valeur des intrants qui permette au modèle substitut de rejoindre la réalité ou

tions. This process is model calibration. We then extend the emulator of choice and calibration procedure for use with high-dimensional responses and demonstrate their speed and efficacy on datasets from a series of transmission impact experiments.

les observations de terrain. On nomme ce processus l'étalonnage du modèle. Nous élargissons ensuite l'émulateur choisi et la procédure d'étalonnage à des réponses en haute dimension, et démontrons la vitesse et l'efficacité sur des jeux de données tirés d'une série d'expériences sur l'impact de la transmission.

[Tuesday May 28/mardi 28 mai, 10:50-11:05]

Shubhadeep Chakraborty (Texas AM University, USA) , **Xianyang Zhang** (Texas AM University)

A New Framework for Distance and Kernel-Based Metrics in High Dimensions

Un nouveau cadre pour les mesures fondées sur des distances et des noyaux en haute dimension

We present new metrics to quantify and test for the equality of two high-dimensional distributions. We show that the energy distance based on the usual Euclidean distance cannot completely characterize the homogeneity of two high-dimensional distributions in the sense that it only detects the equality of means and a specific covariance structure in the high-dimensional setup. To overcome such a limitation, we propose a new class of metrics which inherit some nice properties of energy distance and maximum mean discrepancy in the low-dimensional setting and is capable of detecting the homogeneity of the low-dimensional marginal distributions in the high dimensional setup. We construct a high-dimensional two sample t-test based on the U-statistic type estimator of the proposed metric and study the asymptotic behavior of the t-test under HDLSS and HDMSS setups. We demonstrate the superior power behavior of the proposed tests via both simulated and real datasets.

Nous présentons de nouvelles mesures pour quantifier et tester l'égalité de deux distributions de hautes dimensions. Nous démontrons que la distance énergie fondée sur la distance euclidienne habituelle ne peut pas complètement décrire l'homogénéité de deux distributions de hautes dimensions en ce sens qu'elle détecte seulement l'égalité des moyennes et une structure de covariance spécifique dans la configuration de haute dimension. Pour surmonter une telle limitation, nous proposons une nouvelle classe de mesures qui héritent de certaines propriétés intéressantes de la distance énergie et du maximum de l'écart moyen dans le cadre de faible dimension et qui est capable de détecter l'homogénéité des distributions marginales dans le cadre de faible dimension. Nous construisons un test t sur deux échantillons de hautes dimensions fondé sur un estimateur de type U-statistiques de la mesure proposée et nous étudions le comportement asymptotique du test t sous des configurations HDLSS et HDMSS. Nous démontrons la puissance supérieure du comportement des tests proposés en utilisant des ensembles de données réelles et simulées.

[Tuesday May 28/mardi 28 mai, 11:05-11:20]

Carolyn Augusta (University of Guelph) , **Graham W. Taylor** (University of Guelph) , **Rob Deardon** (University of Calgary)

Conditional Variational Recurrent Graph Autoencoders

Auto-encodeurs de graphes conditionnels de variations récurrentes

In this work, we consider an extension of graph neural networks for learning representations of dynamic graphs, focused on the task of dynamic link prediction. We investigate the combination of conditional variational recurrent autoencoders and graph convolutional networks to learn useful representations of dynamic, directed graphs for the task of link prediction on the whole graph. This task requires a more sophisticated network model than for link prediction considered as a more typical missing data problem, as we are constructing all of the graph's links in the output. We compare our method to previous techniques for link prediction, including autoregressive and persistence baselines, us-

Dans cet exposé, nous examinons une extension des graphes de réseaux de neurones pour apprendre les représentations de graphes dynamiques, axé sur la prédiction du lien dynamique. Nous étudions la combinaison des auto-encodeurs conditionnelles de variations récurrentes et de graphes de réseaux de convolution pour découvrir des représentations utiles des graphes dirigés et dynamiques pour la tâche de prédire le lien dans le graphe en entier. Cette tâche nécessite un modèle de réseau plus sophistiqué que pour la prédiction de lien considéré comme un problème typique de données manquantes puisque nous construisons tous les liens du graphe dans les résultats. Nous comparons notre méthode avec les techniques antérieures pour la prédiction de lien, y compris l'autorégression et les bases de persistance, en utilisant un réseau

ing a real-world dynamic network of shipping among swine farms in Manitoba, Canada, as well as the popular GitHub network.

dynamique réel de livraison entre fermes porcines au Manitoba, Canada, ainsi que le réseau populaire GitHub.

[Tuesday May 28/mardi 28 mai, 11:20-11:35]

Anh Nam Tran (University of Manitoba) , **Saumen Mandal** (University of Manitoba)

Construction of Bayesian Optimal Designs for Nonlinear Models

Construction de plans bayésiens optimaux pour modèles non linéaires

There are a variety of problems in statistics that demand the calculation of some optimizing probability distributions or measures. Optimal design is a particular example. Motivated by the fact that Bayesian methods are ideally suited to contribute to experimental design for nonlinear models, we construct Bayesian optimal designs by incorporating prior information and uncertainties regarding the statistical model. In our Bayesian framework, we consider a discretization of the parameter space to efficiently represent the posterior distribution. We construct designs using the D-optimality criterion. We construct such designs for some logistic models using a clustering approach and a group sequential multiplicative algorithm. The idea is that, at an appropriate iteration, the single distribution is replaced by conditional distributions within clusters and a marginal distribution across the clusters. Finally, we discuss some ideas on reducing computational time using this approach.

Il existe en statistique une variété de problèmes qui exigent le calcul de distributions des probabilités ou de mesures optimales. Le plan optimal en est un exemple particulier. Motivés par le fait que les méthodes bayésiennes conviennent parfaitement au plan d'expérience d'un modèle non linéaire, nous construisons des plans bayésiens optimaux en intégrant des informations préalables et des incertitudes concernant le modèle statistique. Dans notre cadre bayésien, nous considérons une discrétisation de l'espace des paramètres pour représenter la distribution postérieure de manière efficace. Nous construisons des plans à l'aide du critère d'optimalité D. Nous en construisons pour certains modèles logistiques par une approche de regroupement et un algorithme multiplicatif séquentiel de groupe. L'idée est, pour une itérée appropriée, de remplacer la distribution simple par des distributions conditionnelles dans les grappes et par une distribution marginale entre les grappes. Enfin, nous discutons des manières de réduire le temps de calcul par cette approche.

[Tuesday May 28/mardi 28 mai, 11:35-11:50]

Klaus Herrmann (University of Waterloo) , **Maximilian Coblenz** (Karlsruhe Institute of Technology) , **Oliver Grothe** (Karlsruhe Institute of Technology) , **Marius Hofert** (University of Waterloo)

Smooth Bootstrapping of Copula Functionals

Procédure de bootstrap lissé pour les fonctions de copule

This presentation is concerned with results on multivariate kernel distribution estimation and smooth bootstrapping of copula functionals. We discuss how to implement the smooth bootstrap for a functional that depends solely on the underlying dependence structure. Particular attention is paid to the distortion of the dependence structure due to the smooth bootstrap in this setting. While dependence distortion is present in most circumstances, we work out specific cases in which resampling does not affect the dependence structure or the functional in question. A crucial part of multivariate kernel distribution estimation and hence our algorithm is the selection of a suitable bandwidth matrix. While the literature on bandwidth selection for multivariate kernel distribution function estimation has so far focused

Cette présentation porte sur des résultats de l'estimation d'une distribution par noyaux multivariés et d'une procédure de bootstrap lissé pour les fonctions de copule. Nous voyons de quelle façon mettre en œuvre le bootstrap lissé pour une fonction qui dépend uniquement de la structure de dépendance sous-jacente. Nous nous attardons en particulier à la distorsion de la structure de dépendance en raison du bootstrap lissé dans cette configuration. Même si la distorsion de la dépendance est présente dans la plupart des cas, nous mettons au point des cas précis dans lesquels le rééchantillonnage n'affecte pas la structure de dépendance ou la fonction en cause. Un élément indispensable de l'estimation d'une distribution par noyaux multivariés et par conséquent, de notre algorithme, est le choix d'une matrice de paramètres de lissage qui soit convenable. Même si jusqu'à maintenant la littérature relative au choix de paramètres de lissage pour l'estimation de la fonc-

on special cases, we present a cross-validation based approach that is valid for general bandwidth matrices, thus overcoming previous restrictions in the literature.

tion de distribution par noyaux multivariés est axée sur des cas particuliers, nous présentons une approche de validation croisée valable pour des matrices de paramètres de lissage générales, ce qui élimine les restrictions antérieures dans la littérature.

Classification and Learning
Classification et apprentissage

Chair/Président: Wei Liu

Room/Salle: 201 (ENA)

Abstract/Résumé

[Tuesday May 28/mardi 28 mai, 10:20-10:35]

Pramoda Sachinthana Jayasinghe (University of Manitoba) , **Mohammad Jafari Jozani** (University of Manitoba) , **Behzad Kordi** (University of Manitoba)

Developing New Statistical Pattern Recognition and System Identification Techniques for Partial Discharge Analysis

Élaboration de nouvelles techniques de reconnaissance des formes statistique et d'identification de systèmes pour l'analyse des décharges partielles

Partial discharge (PD) in power transmission systems can be harmful to insulators, the equipment that they are connected to. Therefore, identifying partial discharges at an earlier stage is important. We analyze the data obtained in an experimental setting where three partial discharge sources are connected to a high voltage using an oscilloscope. The PD signal waveform obtained for each pulse is used as the predictor and the corresponding source is used as the response class. We use Laguerre function coefficients as the features for training the classification model. A proportion of signals are left out for the purpose of testing the models. We trained linear discriminant analysis (LDA), quadratic discriminant analysis (QDA) and support vector machine (SVM) classifiers. The fitted models performed very well in classifying the signals into their respective sources. The QDA model almost always provided a perfect classification.

Les décharges partielles dans les réseaux de transport d'énergie peuvent être nocives pour les isolateurs, l'équipement auquel ils sont connectés. Par conséquent, il est important d'identifier les décharges partielles à un stade plus précoce. Nous analysons les données obtenues dans un cadre expérimental où trois sources de décharges partielles sont reliées à une haute tension à l'aide d'un oscilloscope. La forme d'onde du signal des décharges partielles obtenue pour chaque impulsion est utilisée comme prédicteur et la source correspondante est utilisée comme classe de réponse. Nous utilisons les coefficients de fonction de Laguerre comme caractéristiques pour proposer un modèle de classification. Une partie des signaux n'est pas prise en compte pour tester les modèles. Nous avons créé des classificateurs d'analyse discriminante linéaire, d'analyse discriminante quadratique et de machine à vecteurs de support. Les modèles ajustés ont très bien réussi à classer les signaux dans leurs sources respectives. Le modèle d'analyse discriminante quadratique fournissait presque toujours une classification parfaite.

[Tuesday May 28/mardi 28 mai, 10:35-10:50]

Pingzhao Hu (University of Manitoba) , **Jiaying You** (University of Manitoba) , **Bob McLeod** (University of Manitoba)

New Machine Learning Approaches for Drug-Target Interaction Network Prediction and Drug Repurposing

Nouvelles approches dans l'apprentissage machine pour la prédiction du réseau d'interaction médicament-cible et la reconversion de médicaments

Traditional methods for drug discovery are time-consuming and expensive, so new efforts have been made to perform drug repurposing, which can develop new uses from already approved drugs. In order to explore new ways for this innovation, many computational approaches have been proposed to predict drug-target interactions (DTIs). However, due to the high-dimensional nature of the data sets extracted from drugs and targets, traditional machine learning approaches cannot analyze these data efficiently. In this study, we collected drug descriptors, protein sequence data from

Les méthodes traditionnelles pour la découverte de médicaments étant longues et coûteuses, de nouveaux efforts ont été consacrés pour la reconversion de médicaments, ce qui pourrait conduire à de nouvelles applications pour des médicaments déjà approuvés. Afin d'explorer de nouvelles façons d'innover, plusieurs approches computationnelles ont été proposées pour prédire les interactions de médicament-cible (IMC). Par contre, à cause de la grande dimension des ensembles de données provenant des médicaments et des cibles, les approches d'apprentissage machine traditionnelles ne peuvent pas analyser ces données efficacement. Dans cet exposé, les données sur les descripteurs des médicaments et sur la

Drugbank and protein domain information from the National Center for Biotechnology Information. Validated DTIs were downloaded from Drugbank. We propose a new deep neural network (DNN) model to predict DTIs and a bipartite clustering analysis was performed using the new predictions for exploring potential drugs in enriched clusters, which can be potentially repurposed for breast cancer.

séquence de protéine proviennent de la Drugbank et l'information sur le domaine protéique provient du National Center for Biotechnology Information. Les IMCs validées sont téléchargées à partir de la Drugbank. Nous proposons un nouveau modèle de réseau de neurones profondes (RNP) pour prédire les IMCs et une analyse de regroupement bipartie a été accomplie en utilisant les nouvelles prédictions pour l'exploration de médicaments potentiels dans des regroupements enrichis, qui pourraient potentiellement être reconvertis pour le cancer du sein.

[Tuesday May 28/mardi 28 mai, 10:50-11:05]

Hanning Chen (University of Calgary) , **Jingjing Wu** (University of Calgary)

Two-Class Classification Problem of Rare and Weak Signal on Variances

Problème de classification en deux classes d'un signal rare et faible des variances

The past tens of years has been an era of data explosion. As a result, in many applications, the data dimension "p" is far larger than the sample size "n". The classical classification methods are no longer applicable. Recently, many efforts have been made to tackle this issue. In my research, I study a two-class classification problem with the same mean but different variances under the "p¿¿n" regime. My interest lies in the possible classification region based on the data features and the performance of quadratic discriminant analysis trained by the higher criticism. Other researchers have studied the cases of the same variances but different means. Combined with my research, one can generalize it to a more common question.

Une explosion de données s'est produite depuis quelques dizaines d'années. Conséquemment, la dimension « p » est beaucoup plus étendue que la taille d'échantillon « n » dans de nombreuses applications. Les méthodes classiques de classifications ne sont plus applicables. Beaucoup d'efforts ont été faits récemment pour venir à bout de ce problème. Mon travail de recherche a porté sur un problème de classification en deux classes de même moyenne mais avec des variances différentes sous un régime « p¿¿n ». Je m'intéresse à une région possible de classification basée sur les caractéristiques des données et la performance de l'analyse discriminante quadratique entrainée par la critique supérieure. D'autres chercheurs ont étudié des cas de mêmes variances mais avec les moyennes différentes. En associant leurs recherches à la mienne, on peut la généraliser à une question plus commune.

[Tuesday May 28/mardi 28 mai, 11:05-11:20]

Jiaxin Zhang (University of Alberta) , **Adam Kashlak** (University of Alberta)

High Dimensional Classification Using Sparse Covariance Matrices

Classification de grande dimension à l'aide de matrices de covariance éparses

High dimensional classification has drawn massive attention due to its increasing application to areas covering genetic diagnosis, image or speech recognition, and financial analysis. Traditional methods such as LDA and QDA, which are optimal Bayes classifiers under normality, fail in high dimensional space where the number of parameters is considerably greater than the sample size, because the conventional empirical estimator for the covariance matrix does not perform well. However, it is reasonable to assume a sparse covariance structure in high dimensions so that we can take advantage of the sparsity to get a better covariance estimator. In our work, we improved the conventional LDA and QDA methods by replacing the empirical estimator with a sparse estimator based on previous research. We

La classification de grande dimension retient de plus en plus l'attention en raison de son application accrue dans des domaines s'étendant du diagnostic génétique, en passant par la reconnaissance visuelle ou vocale jusqu'à l'analyse financière. Les méthodes traditionnelles, comme l'analyse discriminante linéaire (ADL) et l'analyse quantitative descriptive (AQD), qui sont des classificateurs bayésiens optimaux sous la normalité, ne sont pas adéquates dans des espaces de grande dimension où le nombre de paramètres est largement plus élevé que la taille de l'échantillon, car la performance de l'estimateur empirique conventionnel de la matrice de covariance est médiocre. Il est cependant raisonnable de présumer une structure de covariance éparse dans de grandes dimensions afin de tirer parti de l'éparsité pour obtenir un meilleur estimateur de covariance. Notre travail a permis d'améliorer les méthodes conventionnelles ADL et AQD en remplaçant l'estima-

compared our approach with other high-dimension classifiers such as Naïve Bayes and SVM on both simulated and real data.

teur empirique par un estimateur épars en nous basant sur une recherche préalable. Nous comparons notre approche à d'autres classificateurs de grande dimension comme la classification naïve bayésienne et les machines à vecteurs de support (MVS) à l'aide de données de simulées et réelles.

[Tuesday May 28/mardi 28 mai, 11:20-11:35]

Rachid Kharoubi (UQAM) , **Karim Oualkacha** (UQAM)

The Cluster Correlation-Network Support Vector Machine for High-Dimensional Binary Classification

La machine à vecteurs de support avec réseau de corrélation par regroupement pour la classification binaire de haute dimension

Identifying homogeneous subsets of predictors in classification can be challenging in the presence of high-dimensional data with highly correlated variables. We propose a new method called cluster correlation-network support vector machine (CCNSVM) that simultaneously estimates clusters of predictors that are relevant for classification and coefficients of penalized SVM. The new CCN penalty is a function of the well-known Topological Overlap Matrix whose entries measure the strength of connectivity between predictors. CCNSVM implements an efficient algorithm that alternates between searching for predictors' clusters and optimizing a penalized SVM loss function using Majorization–Minimization tricks and a coordinate descent algorithm. This combining of clustering and sparsity into a single procedure provides additional insights into the power of exploring dimension reduction structure in high-dimensional binary classification. Simulation studies are considered to compare methods.

Il peut être ardu de reconnaître les sous-ensembles homogènes des prédicteurs dans une classification en présence de données de haute dimension avec des variables hautement corrélées. Nous proposons une nouvelle méthode nommée machine à vecteurs de support avec réseau de corrélation par regroupement (MVSRCR) qui estime simultanément les regroupements de prédicteurs qui sont pertinents pour la classification et les coefficients des machines à vecteurs de support (MVS) pénalisés. La nouvelle pénalité du réseau de corrélation par regroupement (RCR) est une fonction de la célèbre matrice de chevauchement topologique, dont les termes mesurent la solidité de la connectivité entre les prédicteurs. MVSRCR met en place un algorithme efficace qui alterne entre la recherche de regroupements de prédicteurs et l'optimisation des fonctions de perte des MVS pénalisés grâce à une méthode de maximisation-minimisation et d'un algorithme de descente par coordonnée. Cette combinaison de regroupements et d'éparsité en une seule procédure permet d'apporter davantage de perspectives pour contribuer à l'étude des structures de réduction de dimension dans une classification binaire de haute dimension. Nous examinons des études par simulations pour comparer les méthodes.

[Tuesday May 28/mardi 28 mai, 11:35-11:50]

Zihang Lu (DLSPH, University of Toronto) , **Wendy Lou** (University of Toronto)

Bayesian Growth Mixture Model with Variable Selection for Clustering Longitudinal Trajectories

Modèle de mélange bayésien de croissance avec sélection de variables pour le regroupement des trajectoires longitudinales

In longitudinal studies, growth mixture models are commonly used in clustering the growth trajectories for identifying trajectory patterns. Despite its importance in facilitating medical findings, little work has been done in selecting the predictors related to class membership in the context of growth mixture models. Motivated by novel lung function measures from a Canadian birth cohort study, we propose a unified Bayesian growth mixture model for clustering longitudinal growth trajectories. Our approach provides simultaneous imputation of missing mixed-type (e.g. continuous, categorical) co-

Dans le cadre des études longitudinales, on utilise souvent les modèles de mélange de croissance pour regrouper les trajectoires de croissance afin de déterminer les modèles de trajectoire. Malgré leur importance pour faciliter les découvertes médicales, peu de travaux ont été effectués pour sélectionner les prédicteurs liés à l'appartenance à une classe dans le contexte des modèles de mélange de croissance. Motivés par de nouvelles mesures de la fonction pulmonaire tirées d'une étude de cohorte de naissances canadienne, nous proposons un modèle de mélange bayésien unifié de croissance pour regrouper les trajectoires de croissance longitudinale. Notre approche permet d'imputer simul-

variates and variables selection in the growth mixture models context, in which longitudinal growth trajectories are modelled and subjects are clustered into subgroups conditional on their covariates. Bayesian inference via Monte Carlo Markov Chain algorithm is implemented to estimate the parameters of interest. Results from analyzing real and simulated data will be presented and discussed.

tanément des covariables mélangées (continues et catégorielles) et la sélection des variables manquantes dans le contexte des modèles de mélange de croissance, dans lesquels les trajectoires longitudinales de croissance sont modélisées et les sujets sont regroupés en sous-groupes en fonction de leurs covariables. On met en œuvre l'inférence bayésienne au moyen l'algorithme de la chaîne de Markov de Monte Carlo pour estimer les paramètres d'intérêt. Nous présenterons les résultats de l'analyse des données réelles et simulées et en discuterons.

Poster Session
Session d'affiches

Room/Salle: 103Z (ST)

Abstract/Résumé

[Tuesday May 28/mardi 28 mai, 12:00-17:00]

Dongmeng Liu (Simon Fraser University) , **Jinko Graham** (Simon Fraser University)

Sampling Partial Genealogies Using Sequential Importance Sampling

Échantillonnage de généalogies partielles à l'aide de l'échantillonnage séquentiel par importance

A gene genealogy traces the ancestry of segments of DNA sequence back in time to their common ancestor. We cannot observe the genealogies but the DNA sequence data give us information which can be used to sample from the posterior distribution of the underlying genealogies. Partial genealogies trace the ancestry to a fixed point back in time only and can dramatically improve the efficiency of some commonly-used sampling methods. The posterior distribution of partial genealogies back to a fixed time can be approximated using the empirical distribution of ancestral haplotypes at the stopping time. We use SLiM, a forward-simulation framework, to approximate this empirical distribution. We introduce an algorithm for sampling the partial genealogies of a set of DNA sequences from their posterior distribution. Our algorithm uses sequential importance sampling and accommodates coalescence, mutation and recombination events in the ancestral history of the sequences.

Une généalogie génétique retrace l'ascendance des segments de la séquence d'ADN dans le temps jusqu'à leur ancêtre commun. Nous ne pouvons pas observer les généalogies, mais les données de séquence d'ADN nous donnent de l'information qui peut être utilisée pour échantillonner la loi a posteriori des généalogies sous-jacentes. Les généalogies partielles retracent l'ascendance à un point fixe dans le temps seulement et peuvent améliorer considérablement l'efficacité de certaines méthodes d'échantillonnage couramment utilisées. On peut faire une approximation de la loi a posteriori des généalogies partielles remontant à un temps fixe au moyen de la loi empirique des haplotypes ancestraux au temps d'arrêt. Nous utilisons SLiM, un cadre de simulation prévisionnelle, pour faire une approximation de cette loi empirique. Nous présentons un algorithme permettant d'échantillonner les généalogies partielles d'un ensemble de séquences d'ADN à partir de leur loi a posteriori. Notre algorithme utilise l'échantillonnage séquentiel par importance et tient compte des événements de coalescence, de mutation et de recombinaison dans l'histoire ancestrale des séquences.

[Tuesday May 28/mardi 28 mai, 12:00-17:00]

Mehdi Rostamiforooshani (University of Toronto) , **Nancy Reid** (University of Toronto) , **Olli Saarela** (University of Toronto)

Non-Parametric Feature Selection with False Discovery Rate Control

Sélection de caractéristiques non paramétriques avec contrôle du taux de fausses découvertes

Many variable selection methodologies rely on distributional assumptions and they do not necessarily control the type I error rate. We propose a novel and scalable non-parametric feature selection method, referred to as Data Splitting Selection (DSS). DSS involves randomly splitting the data in half and applying identical search through each portion. We define a feature to be "important" if it shows strong associations in two independent sets. This method controls the False Discovery Rate (FDR) and through simulations, it shows high power in detecting relatively large signals. A variant of DSS is also introduced with the advantage of both controlling FDR and having higher power in case of weak signals. No assumptions are made on the distribution of response

De nombreuses méthodes de sélection des variables reposent sur des hypothèses de distribution et ne contrôlent pas nécessairement le taux d'erreur de type I. Nous proposons une méthode nouvelle et évolutive de sélection de caractéristiques non paramétriques, appelée sélection par division des données. La sélection par division des données consiste à diviser au hasard des données en deux et à effectuer une recherche identique dans chaque partie. Nous définissons une caractéristique comme étant « importante » si elle présente de solides associations dans deux ensembles indépendants. Cette méthode contrôle le taux de fausses découvertes et, grâce à des simulations, montre une puissance élevée de détection de signaux relativement forts. Nous présentons une variante de la sélection par division des données : son avantage est de contrôler le taux de fausses découvertes et d'avoir

or joint distribution of features. We apply the method to an undirected graphical model to remove unimportant connections.

une puissance plus élevée en cas de signaux faibles. Aucune hypothèse n'est faite sur la distribution de la réponse ou la distribution conjointe des caractéristiques. Nous appliquons la méthode à un modèle graphique non dirigé pour supprimer les connexions sans importance.

[Tuesday May 28/mardi 28 mai, 12:00-17:00]

Menglin Zhou (University of British Columbia) , **Natalia Nolde** (University of British Columbia) , **Chen Zhou** (Erasmus University)

An Extreme Value Approach to CoVaR Estimation

Approche des valeurs extrêmes de l'estimation de la CoVaR

An effective measurement of systemic risk should be able to capture the co-movements between a financial system and individual financial institutions. One popular measure of systemic risk is CoVaR. In this talk, a methodology is proposed to compute dynamic forecasts of CoVaR. CoVaR can be viewed as a high quantile of a conditional distribution where the conditioning event corresponds to large losses of an institution. The idea of our methodology is to relate this conditional distribution to the tail dependence function. As a first step, we develop an EVT-based framework to estimate static CoVaR by combining parametric modelling of the tail dependence function to address the issue of data sparsity in the joint tail regions and semi-parametric univariate tail estimation techniques. In the second step, we extend our EVT-based method to the dynamic setting using a bivariate GARCH process. The performance of the methodology is illustrated via simulation studies and real data examples.

Une mesure efficace d'un risque systémique devrait pouvoir saisir les covariations entre un système financier et des institutions financières données. La CoVaR est une mesure populaire du risque systémique. Dans cette présentation, nous proposons une méthode pour calculer des prévisions dynamiques de la CoVaR. La CoVaR peut s'analyser comme un quantile élevé d'une distribution conditionnelle où l'évènement conditionnant correspond aux pertes lourdes d'une institution. L'idée de notre méthodologie est de relier cette distribution conditionnelle à la fonction de dépendance de queue. Dans un premier temps, nous développons un cadre fondé sur la théorie de la valeur extrême (TVE) pour estimer la CoVaR statiquement. Pour cela, nous combinons une modélisation paramétrique de la fonction de dépendance de queue pour répondre au problème de l'insuffisance de données insuffisantes dans les régions communes de queue et des techniques d'estimation de queue univariées semi-paramétriques. Dans un second temps, nous étendons notre méthode fondée sur la TVE à la situation dynamique à l'aide d'un processus de GARCH bivarié. Nous illustrons la performance de la méthode par des études de simulation et des exemples sur données réelles.

[Tuesday May 28/mardi 28 mai, 12:00-17:00]

Li-Pang Chen (University of Waterloo) , **Grace Yi** (University of Waterloo)

Analysis of Graphical Models with Error-Prone Variables

Analyse de modèles graphiques avec variables sujettes à erreur

The graphical model is a useful tool in characterizing the dependence structure of variables. It has been commonly used for analysis of high-dimensional data, including genetic data and data with network structures. Many estimation procedures have been developed under various graphical models with a stringent assumption that the associated variables must be measured precisely. However, in applications, this assumption is often unrealistic and mismeasurement in variables is usually presented in the collected data. We consider analysis of error-prone data under graphical models and propose valid estimation procedures to account for mea-

Le modèle graphique est un outil utile pour caractériser la structure de dépendance des variables. Il est communément utilisé pour l'analyse de données en haute dimension, notamment de données génétiques ou de données avec des structures en réseau. De nombreuses procédures d'estimation ont été développées pour divers modèles graphiques avec la stricte hypothèse que les variables associées doivent être mesurées avec précision. Or cette hypothèse est souvent irréaliste dans les applications, les données collectées présentant généralement des erreurs de mesures des variables. Dans cette présentation, nous étudions l'analyse de données sujettes à erreur dans des modèles graphiques et proposons des procédures d'estimation valides qui tiennent compte des effets

surement error effects. Theoretical results are established for the proposed methods and numerical studies are reported to assess the performance of our proposed methods.

d'erreur de mesure. Nous déterminons des résultats théoriques pour les méthodes proposées et présentons des études numériques qui permettent d'en évaluer la performance.

[Tuesday May 28/mardi 28 mai, 12:00-17:00]

Anqi Chen (Simon Fraser University) , **Boxin Tang** (Simon Fraser University)

Selecting Baseline Two-Level Designs Using Optimality and K-Aberration Criteria When Some Two-Factor Interactions Are Important

Sélection de plans de référence à deux niveaux à l'aide de critères d'optimalité et d'aberration K lorsque certaines interactions entre deux facteurs sont importantes

The baseline parametrization is less commonly used in two-level fractional factorial designs than the orthogonal parametrization. However, it is more natural than its alternative when there exists a default or preferred setting for each factor of a design. The current method selects optimal baseline two-level designs for the main effect model under the assumption that all other effects are negligible. In this presentation, we consider the problem of selecting optimal baseline designs when both main effects and some two-factor interactions are of primary interest. In this case, any nonnegligible effect not in the fitted model causes bias in the estimation of main effects and the important two-factor interactions. Thus, the minimum K-aberration criterion is used to minimize the contamination of the nonnegligible effects. A- and D-optimality criterion are also used to minimize the variance of the estimates. We present a collection of optimal designs of 16 runs and 20 runs.

Dans les plans factoriels fractionnés à deux niveaux, on utilise moins souvent le paramétrage de référence que le paramétrage orthogonal. Cependant, ce paramétrage est plus naturel s'il existe un paramètre défaut ou préférentiel pour chaque facteur d'un plan. Dans la méthode actuelle, des plans de référence à deux niveaux optimaux sont sélectionnés pour le modèle de l'effet principal en partant du principe que tous les autres effets sont négligeables. Dans cette présentation, nous étudions le problème de sélection de plans de référence optimaux lorsque les effets principaux et certaines interactions entre deux facteurs sont d'intérêt primordial. Dans ce cas, tout effet non négligeable qui n'est pas pris en considération dans le modèle ajusté cause un biais dans l'estimation des effets principaux et des interactions importantes entre deux facteurs. Nous utilisons le critère d'aberration K minimum pour minimiser la contamination des effets non négligeables, ainsi que le critère d'optimalité A et D pour minimiser la variance des estimations. Enfin, nous présentons une collection de plans optimaux de 16 et 20 essais.

[Tuesday May 28/mardi 28 mai, 12:00-17:00]

Jeffrey Negrea (University of Toronto)

Optimal Scaling, Optimal Shaping and Speed Limits of Random Walk Metropolis via Diffusion Limits of Block-I.I.D. Targets

Échelle optimale, formation optimale et limites de vitesse d'une marche aléatoire Metropolis par des limites de diffusion des cibles blocs I.I.D

We extend the results of Roberts et al. (1997) by considering limits of Random Walk Metropolis (RWM) applied to block I.I.D. targets with corresponding block-independent proposals. We verify their recommendation to "tune the scaling to achieve an acceptance rate of 0.234", for any fixed choice of proposal shaping under within-block dependence. Using the block structure, we upgrade the form of convergence from weak convergence of the path of a single component to weak convergence of the infinite dimensional process. We also optimise the limiting speed over the choice of covariance structure. We derive the optimal proposal shaping (in terms of the decay of certain autocorrelations) for gen-

Nous étendons les résultats de Roberts et al. (1997) en considérant les limites de la marche aléatoire Metropolis (MAM) appliquée aux cibles blocs I.I.D. avec les propositions correspondantes par blocs indépendants. Nous vérifions leur recommandation de « ajuster l'échelle pour obtenir un taux d'acceptation de 0,234 » pour n'importe quel choix fixe de proposition de formation lors de dépendance entre les blocs. En utilisant la structure par blocs, nous améliorons la forme de la convergence de faible convergence du chemin d'une composante unique à une faible convergence du processus à dimension infinie. Nous optimisons aussi la vitesse limitante sur le choix de la structure de la covariance. Nous dérivons la formation optimale (en termes de la décroissance de certaines autocorrélations) pour des cibles générales. Nous dérivons aussi

eral targets. We also derive the optimal shaping in terms of spectral gaps in special cases. Lastly, we show that RWM performance degrades when certain functionals of the dependence scale infavourably with dimension. In such cases, no tuning of RWM will give performance on par with an I.I.D. target.

la formation optimale en termes d'écarts spectraux dans des cas particuliers. Finalement, nous démontrons que la performance de la MAM se détériore quand certaines fonctions de la dépendance s'ajustent défavorablement avec la dimension. Dans ce cas, aucun ajustement de la MAM ne donnera une performance au niveau de la cible I.I.D.

[Tuesday May 28/mardi 28 mai, 12:00-17:00]

Zhiyang Zhou (Simon Fraser University) , **Peijun Sang** (University of Waterloo)

Continuum Centroid Classifier for Functional Data

Classificateur centroïde continu pour les données fonctionnelles

Aiming at classification of functional data, we propose the continuum centroid classifier (CCC), which is built upon projections of functional data onto one specific direction. This direction is obtained via bridging regression and classification. This dimension reduction technique falls between the unsupervised learning and fully supervised one. Thanks to the intrinsic infinite dimension of functional data, the CCC possibly enjoys the (asymptotic) zero misclassification rate without specifying the distribution of data. We propose an effective algorithm to implement this classifier. Simulation studies and two real examples all demonstrate that CCC compares favorably with some existing methods in terms of the misclassification rate.

Ayant pour but la classification des données fonctionnelles, nous proposons le classificateur centroïde continu (CCC), établi selon les projections vers un point précis de données fonctionnelles. Cette orientation est obtenue en rapprochant régression et classification. Cette technique de réduction dimensionnelle se situe entre l'apprentissage non supervisé et entièrement supervisé. Grâce à la dimension infinie intrinsèque des données fonctionnelles, le CCC profite possiblement d'un taux nul de classification erronée (asymptotique) sans nécessiter de préciser la distribution des données. Nous proposons un algorithme efficace pour la mise en oeuvre de ce classificateur. Des études en simulation et deux exemples tirés de la réalité montrent que relativement au taux de classification erronée, le CCC se compare avantageusement à certaines méthodes en usage.

[Tuesday May 28/mardi 28 mai, 12:00-17:00]

Myriam Brossard (Sinai Health System) , **Andrew Paterson** (Hospital for Sick Children University of Toronto) , **Osvaldo Espin-Garcia** (Sinai Health System University of Toronto) , **Radu Craiu** (University of Toronto) , **Shelley Bull** (Sinai Health System University of Toronto)

Joint Modelling of Multivariate Longitudinal and Time-To-Event Phenotypes in Genetic Association Studies of Complex Traits

Modélisation conjointe de phénotypes longitudinaux et de durées de vie multivariés dans les études sur l'association génétique de caractéristiques complexes

Type 1 diabetes complications (T1DC) are characterized by complex genetic architecture where variants (SNPs) are associated directly with risk of T1DC and/or indirectly through longitudinal risk factors. We propose a joint model (JM) of multiple longitudinal time-to-event traits to infer SNP effects that characterize T1DC architecture. The JM is formulated with (a) a mixed model for longitudinal traits as a function of time SNP effects; (b) a frailty model for survival traits depending on SNP effects longitudinal trajectories from (a). We estimate parameters via a 2-stage approach and the covariance matrix by the bootstrap, and develop hypothesis tests for direct /or indirect SNP association with multiple traits. We establish a data simulation algorithm that mimics complex genetic architecture in a T1DC study

Les complications du diabète de type 1 (CDT1) se caractérisent par une architecture génétique complexe dans laquelle les variations (PSN) sont associées directement au risque de CDT1 ou indirectement à des facteurs de risque longitudinaux. Nous proposons un modèle conjoint (MC) de caractéristiques longitudinales et de durées de vie multiples afin d'inférer les effets PSN caractéristiques de l'architecture des CDT1. Le MC est formulé avec (a) un modèle de mélange pour les caractéristiques longitudinales comme fonction de temps et d'effets PSN ; (b) un modèle de fragilité pour les caractéristiques de survie dépendantes des effets PSN et des trajectoires longitudinales de (a). Nous estimons les paramètres en utilisant une approche en deux temps et la matrice de covariables par bootstrap et mettons au point des tests d'hypothèse pour une association PSN directe ou indirecte avec caractéristiques multiples. Nous établissons un algorithme de simu-

and demonstrate the feasibility, validity power of the JM to correctly retrieve various direct and/or indirect SNP associations.

lation de données imitant l'architecture génétique complexe d'une étude sur les CDT1 et montrons la faisabilité, la validité et la puissance du MC pour extraire diverses associations PSN directes ou indirectes.

[Tuesday May 28/mardi 28 mai, 12:00-17:00]

Michela Panarella (University of Toronto DLSPH) , **Razvan Romanescu** (Lunenfeld-Tanenbaum Research Institute) , **Gessica Gos** (Lunenfeld-Tanenbaum Research Institute) , **Irene Andrulis** (Lunenfeld-Tanenbaum Research Institute) , **Shelley Bull** (Lunenfeld-Tanenbaum Research Institute)

Extending Rare Variant Association Tests in Affected Siblings to Account for Background Genetic Risk

Prolongement des tests d'association de variantes rares chez des frères-sœurs atteints pour prendre en compte le risque génétique

Genetic diseases are thought to be caused by both common and rare variants. Family studies are often used to detect inherited rare variants associated with disease while population studies are used to detect common variants. Population studies have led to development of individual-level genetic risk scores (GRS), an aggregate of putative susceptibility variants derived from genome-wide measures of disease risk. Currently, rare variant tests for affected sibling pairs (ASPs) do not include the GRS to account for background disease risk from common variants. We demonstrate how to extend existing rare variant ASP tests to include the GRS and report simulations that evaluate: (1) validity of the methods and (2) relative efficacy of considering risk from common variants when testing for rare variant association in family settings. We apply the methods to whole-exome sequence data of sisters ascertained on early-onset breast cancer and incorporate recently reported population GRS.

On estime que les maladies génétiques sont causées par des variantes à la fois répandues et rares. On utilise souvent des études familiales pour dépister des variantes rares héréditaires associées à la maladie, tandis que les études basées sur une population sont utilisées pour dépister les variantes communes. Les études basées sur une population ont permis d'en arriver à des scores de risque génétique (SRG), une agrégation de variantes de susceptibilité putative dérivées des mesures pangénomiques du risque de maladie. Pour l'instant, les tests de dépistage de variantes rares chez des paires frères-sœurs atteints (PFSA) n'englobent pas les SRG pour la prise en compte du risque lié aux antécédents de maladie découlant de variantes communes. Nous montrons comment étendre les tests de dépistage de variantes rares dans les paires FSA pour y intégrer les SRG et des simulations de rapport afin d'évaluer : (1) la validité des méthodes et (2) l'efficacité relative de prendre en compte le risque lié aux variantes communes pour tester l'association de variantes rares selon des paramètres familiaux. Nous appliquons les méthodes aux données de séquençage de l'exome complet chez des sœurs chez qui on a constaté une apparition précoce de cancer du sein et incorporons des SRG populationnels rapportés récemment.

[Tuesday May 28/mardi 28 mai, 12:00-17:00]

Faezeh Yazdi (Simon Fraser University) , **Derek Bingham** (Simon Fraser University) , **Daniel Williamson** (University of Exeter)

Emulation of Computer Models Using Deep Gaussian Processes

Émulation de modèles informatisés à l'aide de processus gaussiens profonds

Deterministic computer models (or simulators) are used to explore physical systems in many scientific disciplines. Simulators are often computationally demanding to evaluate and in other cases, they are fast to evaluate but are not readily available to all scientists. In such settings, a statistical emulator can serve as a surrogate, making predictions of computer model output at unsampled input values with estimates of uncertainty. The traditional approach for emulator building is using a Gaus-

Les modèles informatisés déterministes (ou simulateurs) sont utilisés pour l'exploration de systèmes physiques dans plusieurs disciplines scientifiques. Pour l'évaluation, les simulateurs sont souvent exigeant sur le plan computationnel, tandis que dans d'autres cas, ils permettent des évaluations rapides, mais ne sont pas à la portée de tous les scientifiques. Dans ce contexte, un émulateur statistique peut servir de substitut en prédisant les données de sortie d'un modèle informatique pour des valeurs de données d'entrée non échantillonnées avec une estimation de l'incertitude. L'ap-

sian process prior, where stationarity is a common simplifying assumption. This assumption is often realistic in practice. In this work, we propose a non-stationary model to recreate the desired predictive behavior of the emulator. We show that deep Gaussian processes with hierarchical convolution constructions can capture non-standard features of computer models, thereby giving good predictive performance. We illustrate our method with some simulation examples.

proche traditionnelle pour la construction d'un émulateur est l'utilisation d'un apriori de processus gaussien, lorsque la stationnarité est une hypothèse simplificatrice commune. En pratique, cette hypothèse est souvent réaliste. Nous proposons ici un modèle non stationnaire pour recréer le comportement prédictif désiré de l'émulateur. Nous montrons aussi que les processus gaussiens profonds avec des constructions convolutives hiérarchiques peuvent capturer les caractéristiques non normatives des modèles informatiques, offrant ainsi une bonne performance prédictive. Nous illustrons notre méthode à l'aide de quelques exemples de simulations.

[Tuesday May 28/mardi 28 mai, 12:00-17:00]

Amirhossein Alvandi (Memorial University of Newfoundland) , **Armin Hatefi** (Memorial University of Newfoundland)

Population Proportion Estimation Using Rank-Based Sampling

Estimation proportionnelle d'une population à l'aide d'un échantillonnage basé sur les rangs

Rank-based sampling (RBS) designs as cost-effective sampling schemes have found a wide range of applications such as medical research. These statistical techniques are commonly used in situations where measuring the variable of interest is expensive, but a small number of sampling units can be easily ranked before taking the final measurements on them using auxiliary information such as inexpensive concomitant variables. When multiple concomitants are available, it is important to incorporate all the available information into the analysis. Various nonparametric and likelihood-based estimation procedures have been proposed in the literature for population proportion estimation using RBS data. We compare the performance of these estimators with multiple rankers-based estimations. The estimation procedures are then applied to a breast cancer study to estimate disease prevalence.

Comme techniques économiques pour l'échantillonnage, les concepts d'échantillonnage basé sur les rangs (EBR) servent à une multitude d'applications variées, notamment la recherche médicale. Ces techniques statistiques sont couramment utilisées dans des cas où la mesure de la variable d'intérêt est coûteuse, mais où il est facile de classer en rangs un petit nombre d'unités d'échantillonnage avant d'en prendre les mesures finales à l'aide de renseignements complémentaires, telles que des variables concomitantes peu coûteuses. Lorsque des variables concomitantes multiples sont disponibles, il est important d'intégrer à l'analyse tous les renseignements disponibles. La documentation propose diverses procédures d'estimation non paramétrique et basée sur la vraisemblance pour l'estimation proportionnelle d'une population à l'aide de données EBR. Nous comparons la performance de ces estimateurs avec des estimations basées sur de multiples facteurs de rang . Les procédures d'estimation sont ensuite appliquées à une étude sur la cancer du sein visant à estimer la prévalence de la maladie.

[Tuesday May 28/mardi 28 mai, 12:00-17:00]

Yunjing Li (University of Toronto) , **Nicholas Mitsakakis** (University of Toronto)

Categorical State Sequence Analysis to Describe and Analyze Pathways of Health State Transitions

Analyse de séquence d'états catégorique pour décrire et analyser les chemins de transitions d'états de santé

In medical research, transitions between mutually exclusive health states of a disease are often described with state transition diagrams. However, data following this graphical representation are not conducive to further analysis and pattern investigation. Inspired by state sequence analysis methods developed and used in social science and bioinformatics, we consider describing patient pathways across different health states as sequences of categorical data. We investigate the use of three dissimilarity measures in sequence analy-

En recherche médicale, les transitions entre les états de santé mutuellement exclusifs d'une maladie sont souvent illustrées à l'aide d'un diagramme d'états-transitions. Toutefois, les données suivant cette représentation graphique ne sont pas prédisposées à approfondir l'analyse et l'étude de tendances. En nous inspirant des méthodes d'analyse de séquence d'états, que les sciences sociales et la bio-informatique ont conçues puis adoptées, nous examinons la description des chemins du patient à travers différents états de santé en guise de séquences de données catégoriques. Nous étudions aussi l'utilisation de trois mesures de dissemblance en

sis: optimal matching, order-based similarity measure and longest common subsequence. We use simulated and real data from prostate cancer patients, while we appropriately consider and discuss issues that death as an absorbing state, and censoring pose in sequence analysis. Furthermore, we utilize the dissimilarity scores and we conduct cluster analysis and apply sequence visualization methods, aiming to explore patterns that these pathway data exhibit.

analyse de séquence : l'appariement optimal, la mesure de similitude basée sur l'ordre et la sous-suite commune la plus longue. Nous nous servons de données simulées et réelles provenant de patients atteints du cancer de la prostate pour examiner de façon pertinente les problèmes que le décès, en tant qu'état absorbant, et la censure posent dans les analyses de séquence. En outre, nous utilisons les scores de dissemblances, menons des analyses de grappes et appliquons des méthodes de visualisation de séquences dans le but d'étudier les tendances présentes dans ces données des chemins.

[Tuesday May 28/mardi 28 mai, 12:00-17:00]

Yunqi Ji (Alberta Health Services) , **Farrell Cahill** (Memorial University of Newfoundland) , **Yanqing Yi** (Memorial University of Newfoundland) , **Edward Randell** (Memorial University of Newfoundland) , **Guang Sun** (Memorial University of Newfoundland)

New Sex Specific Predictive Equation for Evaluating Body Fat Percentage

Une nouvelle équation prédictive spécifique au sexe pour déterminer le pourcentage de masse grasse

The high prevalence of obesity and obesity-related commodities are critically important global health issues. The Inexpensive and accurate assessment of adiposity are required for large studies on obesity and related diseases, especially in rural communities. Percent of body fat (%BF) measured utilizing a dual-energy X-ray absorptiometry (DXA) is one of the most reliable measurements of adiposity, but expensive and inconvenient for large populations. The widely-used Body Mass Index (BMI) and recently-developed Body Adiposity Index (BAI) ignore sex differences in weight, adiposity and fat distribution and may misclassify obesity status. We applied Lin's concordance correlation coefficient which represents the strength-of-agreement between two continuous variables to develop a new sex-specific equation to predict body fat percentage based on a population-based study in Newfoundland. The new body fat index shows higher power to predict %BF compared to BMI and BAI.

Le caractère hautement généralisé de l'obésité et des produits qui en sont dérivés représente un important problème de santé à l'échelle globale. Dans le cadre de grandes études portant sur l'obésité et sur les maladies qui y sont associées, il est nécessaire de pouvoir déterminer précisément l'adiposité de façon abordable, tout particulièrement dans les régions rurales. Mesurer le pourcentage de masse grasse (% MG) au moyen d'une absorptiométrie biénergétique à rayons X (DXA) est le moyen le plus fiable pour mesurer l'adiposité, mais il est dispendieux et peu pratique pour les grandes populations. L'indice de masse corporelle (IMC), largement utilisé, et l'indice de masse adipeuse (IMA), tout récemment conçu, ne tiennent pas compte de la différence de poids, d'adiposité et de distribution des graisses selon le sexe, ce qui pourrait entraîner des erreurs de classification de l'obésité. Nous avons appliqué le coefficient de corrélation de concordance de Lin, représentant le degré de concordance entre deux variables continues pour créer une nouvelle équation spécifique au sexe pour prédire le pourcentage de masse grasse à partir d'une étude basée sur une population à Terre-Neuve. La performance du nouvel indice de masse grasse pour prédire le % MG est supérieure par rapport à l'IMC et à l'IMA.

[Tuesday May 28/mardi 28 mai, 12:00-17:00]

Xiaohua Liu (University of Regina) , **Taehan Bae** (University of Regina)

Predictive Modelling of Extreme Values in Stock Return Data

Modélisation prédictive des valeurs extrêmes dans les données sur le rendement des actions

This study aims at predictive modelling of a two-sided, heavy-tailed data under a mixture model setting. A robust regression method is used to fit the main body, while the peaks-over-threshold method is employed to model the tails. For the estimation of the tail parts, the

Cette étude vise à modéliser de façon prédictive des données bilatérales et à queue lourde dans le cadre d'un modèle de mélange. On utilise une méthode de régression robuste pour ajuster la partie principale, tandis que la méthode de dépassement de seuil est utilisée pour modéliser les queues. Pour l'estimation des par-

Bayesian maximum a posterior estimation with conjugate priors is used to smooth the maximum likelihood estimates (MLEs). This filter tuning process provides stability and efficiency in computation and prediction. Several constrained, non-convex optimization problems have been converted to unconstrained, convex problems by quadratic approximation and variable changes. The approach is applied to a large, multi-period, unbalanced data set of daily returns of global stocks, containing nearly 100,000 records. Out-of-sample prediction results show the out-performance of the smoothed estimates over the regular MLEs.

ties de la queue, le maximum a posteriori bayésien avec des a priori conjugués est utilisé pour lisser les estimations du maximum de vraisemblance. Ce processus de réglage du filtre assure la stabilité et l'efficacité du calcul et de la prédiction. Plusieurs problèmes d'optimisation non convexes et limités ont été convertis en problèmes convexes et non contraints par approximation quadratique et changements de variables. L'approche est appliquée à un vaste ensemble de données multipériodiques et non équilibrées sur les rendements quotidiens des actions mondiales, qui contient près de 100 000 enregistrements. Les résultats des prédictions hors échantillon montrent la surefficacité des estimations lissées par rapport aux estimations du maximum de vraisemblance classiques.

[Tuesday May 28/mardi 28 mai, 12:00-17:00]

Xuewen Lu (University of Calgary) , **Rutong Cai** (University of Calgary) , **Beijia Hu** (University of Calgary) , **Alexander Liu** (University of Calgary) , **Liping Luo** (Hangyang Normal University) , **Connie Sze** (University of Calgary) , **Yao Yao** (University of Calgary)

Group Selection for Accelerated Failure Time Models with Random Effects

Sélection de groupes pour les modèles de temps de défaillance accéléré avec effets aléatoires

The Accelerated failure time (AFT) model is a parametric regression model which can effectively analyze the relationship between survival time and explanatory variables. When survival times are correlated, for example, in data from multi-center trials or health studies with multilevel structures, we include random effects to describe the dependence. Variable selection procedure of fixed effects at individual levels in the AFT random-effect models has been studied in the literature. However, in practice, there are often group structures in the covariates. Upenalized h-likelihood, we introduce several group selection methods to select variables at group levels. To implement these methods, we propose an efficient optimization algorithm for nonconvex group penalties by combining the concave convex procedure (CCCP) and the group LASSO algorithm. We compare these methods via simulation studies, then apply the method to analyze two real data sets.

Le modèle du temps de défaillance accéléré est un modèle de régression paramétrique qui peut analyser efficacement la relation entre le temps de survie et les variables explicatives. Lorsque les temps de survie sont corrélés, par exemple, dans les données d'essais multicentriques ou d'études sur la santé à structures multiniveaux, nous incluons des effets aléatoires pour décrire la dépendance. Dans la littérature, on a étudié la procédure de sélection des variables à effets fixes aux niveaux individuels dans les modèles à effets aléatoires du temps de défaillance accéléré. Toutefois, dans la pratique, il existe souvent des structures de groupe dans les covariables. Nous présentons plusieurs méthodes de sélection de groupes pour sélectionner des variables au niveau des groupes. Pour mettre en œuvre ces méthodes, nous proposons un algorithme d'optimisation efficace des pénalités de groupes non convexes en combinant la procédure convexe concave et l'algorithme de lasso groupé. Nous comparons ces méthodes avec des études de simulation, puis nous appliquons la méthode à l'analyse de deux ensembles de données réelles.

Recent Advances in Actuarial and Quantitative Finance
Dernières avancées en finance actuarielle et quantitative

Chair/Président: Jean-François Bégin

Organizer/Responsable: Jean-François Bégin

Room/Salle: 146 (SB)

Abstract/Résumé

[Tuesday May 28/mardi 28 mai, 13:30-14:00]

Jean-François Renaud (Université du Québec à Montréal)

How Are Dividend Payments Affected by Parisian Ruin?

Comment la ruine parisienne influence-t-elle les paiements de dividende ?

The quest for the dividend policy maximizing the expected present value of the dividend payments made by an insurance company dates back at least to the pioneering work of Bruno de Finetti (1957). More recently, the idea of implementation delays in the recognition of the insurer's default has generated a lot of research attention. In this talk, we will look at the impact of Parisian ruin on dividend payments in a model where the underlying surplus process is a spectrally negative Lévy process and the controlled surplus process is allowed to spend time under the default level. The set of horizontal barrier strategies will be discussed. In particular, we will compare the optimal barrier levels when classical ruin or Parisian ruin is implemented.

La quête d'une politique de dividende maximisant la valeur actuelle attendue des paiements de dividende effectués par un assureur remonte aux premiers travaux de Bruno de Finetti (1957). Plus récemment, la notion de délais de mise en œuvre relatifs à la reconnaissance des obligations de l'assureur a grandement attiré l'attention des chercheurs. Lors de cet exposé, nous examinerons quels sont les effets d'une ruine parisienne sur les paiements de dividende dans un modèle où le processus d'excédent sous-jacent est un processus de Lévy à spectre négatif, et le processus d'excédent contrôlé est autorisé à passer du temps sous le niveau par défaut. Nous aborderons aussi les stratégies concernant l'ensemble de barrière horizontale. Tout particulièrement, nous comparerons les niveaux de barrière optimaux selon la mise en œuvre d'une ruine classique ou parisienne.

[Tuesday May 28/mardi 28 mai, 14:00-14:30]

Patrice Gaillardetz (Concordia University) , **Saeb Hachem** (Concordia University)

Risk-Control Strategies

Stratégies de contrôle des risques

We consider the pricing of derivative products that involve dynamic hedging strategies and payments within the planning horizon. Equity-indexed annuities (EIAs), Guaranteed Investment Certificates (GICs), and American and Barrier options are typical examples of these products. Our exploration involves evaluation under crossovers of assumptions related to the portfolio composition and to the risk tolerated by the issuer. The unified constrained discrete stochastic dynamic programming framework in this presentation makes use of sequential local minimizing strategies related to stochastic transitions. This sequential minimization takes into account all intermediate requirements and involves either dynamic risk measures or riskless modeling. To demonstrate the flexibility of this framework, we present numerical examples featuring GICs.

Nous étudions l'évaluation des produits dérivés avec des stratégies de couverture dynamique et des paiements dans l'horizon de planification. Les produits de rente indexée, certificats de placement, options américaines et à barrière en sont des exemples typiques. Notre exploration inclut une évaluation avec hypothèses croisées concernant la composition du portefeuille et le risque toléré par l'émetteur. Le cadre de programmation dynamique stochastique discret limité unifié présenté ici exploite des stratégies de minimisation locale séquentielle liées aux transitions stochastiques. Ces minimisations séquentielles tiennent compte de toutes les exigences intermédiaires et impliquent soit des mesures de risque dynamiques, soit une modélisation sans risque. Pour démontrer la souplesse de ce cadre, nous présentons des exemples numériques de certificats de placement.

[Tuesday May 28/mardi 28 mai, 14:30-15:00]

Adam Metzler (Wilfrid Laurier University) , **Mark Reesor** (Wilfrid Laurier University) , **Wisdom S. Avusuglo** (Western University)

A General Framework for Modelling PD-LGD Correlation in Loan Portfolios: Some Interesting Observations

Un cadre général pour la modélisation de corrélations PD-LGD dans des portefeuilles de prêts, avec quelques observations intéressantes

We show that a number of PD-LGD correlation models that have been proposed in the literature, are special cases of a more general framework. We explore this framework in detail and make several interesting observations. For instance, (i) we identify a common modelling error that tends to overestimate economic capital, and (ii) we find that the relationship between account-level correlations (modelling inputs) and portfolio-level correlations (modelling outputs) is surprisingly weak.

Nous montrons qu'un certain nombre de modèles de corrélations PD-LGD (probabilité de défaut de paiement-perte encourue en cas de défaut) que propose la documentation sont des cas particuliers d'un cadre plus général. Nous voyons de près ce cadre qui donne lieu à plusieurs observations intéressantes. Par exemple, (i) nous identifions une erreur répandue de modélisation qui tend à surestimer le capital économique, et (ii) nous constatons la faiblesse étonnante de la relation entre les corrélations de comptes (données d'entrée de la modélisation) et celles de portefeuilles (données de sortie de la modélisation).

Modeling, Imputation and non response
Modélisation, imputation et non-réponse

Chair/Président: David Haziza

Organizer/Responsable: Susie Fortier

Room/Salle: 201 (ENA)

Abstract/Résumé

[Tuesday May 28/mardi 28 mai, 13:30-14:00]

Katherine Jenny Thompson (U.S. Census Bureau) , **Nicole Czaplicki** (U.S. Census Bureau) , **Brian Dumbacher** (U.S. Census Bureau) , **Stephen Kaputa** (U.S. Census Bureau)

Developing Imputation Models for the Advanced Monthly Retail Trade Survey: A Subsampled Survey with Seasonal Data, Low Unit Response, and High Profile

Élaboration de modèles d'imputation pour l'Advance Monthly Retail Trade Survey : étude par sous-échantillonnage avec données saisonnières, faible réponse par unité et haut profil

The Advanced Monthly Retail Trade Survey (MARTS) publishes early sales estimates of retail and food service industries approximately nine working days after the reference month. One month later, the MARTS sales estimate is superseded by the preliminary estimate from the Monthly Retail Trade Survey (MRTS). MARTS is a subsample of MRTS, and MARTS nonrespondents may provide a valid response to MRTS for the same reference period. Large revisions between corresponding estimates are highly scrutinized as MARTS is an economic indicator. Consequently, the U.S. Census Bureau is investigating alternative missing data treatments for MARTS, focusing primarily on item imputation. The methods considered include regression trees, RegARIMA models, and hierarchical Bayesian regression models. We present the design and preliminary results from a simulation study that assesses statistical properties of the proposed imputation methods in concert with the currently-used synthetic MARTS estimation procedure

L'Advance Monthly Retail Trade Survey (MARTS) américain publie des estimations des ventes des secteurs du détail et de la restauration environ neuf jours ouvrables après le mois de référence. Un mois plus tard, les estimations du MARTS sont supplantées par les estimations préliminaires du Monthly Retail Trade Survey (MRTS). Le MARTS est un sous-échantillon du MRTS et les non-répondants au MARTS peuvent fournir une réponse valide au MRTS pour la même période de référence. Les révisions importantes entre estimations correspondantes sont analysées minutieusement car le MARTS est un indicateur économique. Par conséquent, le Census Bureau américain explore d'autres façons de traiter les données manquantes pour le MARTS, et notamment des méthodes d'imputation des réponses. Parmi les méthodes étudiées, notons les arbres de régression, les modèles RegARIMA et les modèles de régression bayésienne hiérarchique. Nous présentons le plan et les résultats préliminaires d'une étude par simulation qui évalue les propriétés statistiques des méthodes d'imputation proposées de concert avec la procédure d'estimation synthétique qu'emploie actuellement le MARTS.

[Tuesday May 28/mardi 28 mai, 14:00-14:30]

Geneviève Vézina (Statistics Canada) , **Andrew Brennan** (Statistics Canada) , **Catherine Deshaies-Moreault** (Statistics Canada)

Measuring Cannabis Prevalence, Consumption and Price: The Statistics Canada Experience

Mesure de la prévalence, de la consommation et du prix du cannabis : l'expérience de Statistique Canada

With the recent legalization of cannabis use in Canada, Statistics Canada has implemented a number of new programs to monitor and measure the cannabis market. In this paper we describe three of those initiatives. The first is a description of the new National Cannabis Survey which is part of Statistics Canada's Rapid Stats

Avec la légalisation récente de la consommation de cannabis au Canada, Statistique Canada a mis en place plusieurs programmes pour surveiller et mesurer le marché du cannabis. Dans cette présentation, nous décrivons trois de ces initiatives. La première est une description de l'Enquête nationale sur le cannabis, qui s'intègre dans le programme Rapidonnées de Statistique Canada.

program. We describe this program and how it has allowed Statistics Canada to produce timely cannabis related statistics. The second is a pilot project that aims to estimate how much cannabis is consumed by humans from what can be measured in wastewater. We describe the objective and scope of the pilot project, the algorithms to back-calculate consumption and sources of uncertainty in the parameters of the algorithm. Finally, we describe a crowdsourcing project that allowed Statistics Canada to produce cannabis related statistics on the price, quantity, quality, city and purpose of consumption. Strengths and limitations of this approach are discussed.

Nous décrivons le programme et comment il a permis à Statistique Canada de produire des statistiques opportunes sur le cannabis. La deuxième est un projet pilote qui vise à estimer la consommation humaine de cannabis à partir de mesures des eaux usées. Nous en décrivons l'objectif et la portée, les algorithmes qui permettent le rétrocalcul de la consommation et les sources d'incertitude dans les paramètres de l'algorithme. Enfin, nous décrivons un projet de ≪ crowdsourcing ≫ qui a permis à Statistique Canada de produire des statistiques sur le prix, la quantité, la qualité, le lieu et la raison de la consommation de cannabis. Nous discutons des forces et des limites de cette démarche.

[Tuesday May 28/mardi 28 mai, 14:30-15:00]

Valéry Dongmo Jiongo (Canada Mortgage and Housing Corporation)

Predicting Rental Prices for Canadian Rural Centres

Prévision des prix de location des centres ruraux canadiens

The Rental Market Survey (RMS) is conducted annually in urban centres and every five years in rural areas. The last rural RMS was conducted in 2015. However, there is a demand for annual rural rental data to meet the need of the Core Housing Need Income analysis. In the absence of an actual rural RMS, the panel data modeling of the rural rentals is studied to impute the missing data. We adopt a sequential approach consisting of an out-off-sample followed by an in-sample prediction using historical survey data, administrative data and the current urban RMS data. We compare the predicted rental data with both the special rural RMS from 2018, which was conducted in 66 rural areas, and the rural RMS from 2015, which was conducted in 218 rural centres. The results show that the modeling performs satisfactorily well.

L'enquête sur le marché locatif est menée chaque année dans les centres urbains et tous les cinq ans dans les régions rurales. La dernière enquête dans les régions rurales a été réalisée en 2015. Toutefois, les données annuelles sur les loyers en milieu rural sont nécessaires pour répondre aux besoins de l'analyse des besoins impérieux en matière de logement selon le revenu. En l'absence d'une enquête réelle sur le marché locatif dans les régions rurales, on étudie la modélisation des données de panel des loyers en milieu rural pour imputer les données manquantes. Nous adoptons une approche séquentielle consistant en une prévision hors échantillon suivie d'une prévision dans l'échantillon à l'aide de données d'enquêtes historiques, de données administratives et des données actuelles de l'enquête sur le marché locatif en milieu urbain. Nous comparons les données sur les loyers prévus avec celles de l'enquête spéciale sur le marché locatif de 2018, qui a été réalisée dans 66 zones rurales, et l'enquête sur le marché locatif de 2015, qui a été réalisée dans 218 centres ruraux. Les résultats montrent que la modélisation donne de bons résultats.

Models and Applications for Functional Data Analysis
Modèles et applications pour l'analyse de données fonctionnelles

Chair/Président: Haocheng Li

Organizer/Responsable: Haocheng Li

Room/Salle: 109 (SS)

Abstract/Résumé

[Tuesday May 28/mardi 28 mai, 13:30-14:00]

David A. Campbell (Simon Fraser University) , **Subhash Lele** (University of Alberta) , **Peter Solymos** (University of Alberta)

Functional Data Analysis for Assessing Convergence of Sampled Densities

Analyse de données fonctionnelles pour évaluer la convergence de densités échantillonnées

KL divergence is a well known asymmetric measure of distance between distributions, but when densities are sampled instead of computed, distances and differences may be obfuscated by sampling variability and lack of certainty about a reference density. This work showcases a model for assessing whether or not densities are statistically distinguishable using tools from functional data analysis with application to assessing if two or more MCMC runs have converged to the same limiting distribution. The problem is formulated using functional generalized linear models to classify if samples came from a particular MCMC run. The model is interpretable as a symmetric measure of disagreement between sampled densities but rather than an integrated measure, it showcases specific regions and directions of distributional dispute.

La divergence KL est une mesure asymétrique bien connue de la distance entre distributions mais quand les densités sont échantillonnées au lieu d'être calculées, les distances et différences peuvent être embrouillées par la variabilité de l'échantillon et par l'absence de certitude à propos d'une densité de référence. Cet exposé présente un modèle pour évaluer si les densités sont statistiquement différentiables en utilisant des outils de l'analyse des données fonctionnelles avec une application pour évaluer si deux itérations MCCM ou plus ont convergé vers la même distribution limite. Le problème est formulé en utilisant des modèles linéaires fonctionnels généralisés pour classifier si des échantillons proviennent d'une itération MCCM particulière. Le modèle s'interprète comme une mesure symétrique du désaccord entre les densités échantillonnées mais plutôt qu'une mesure intégrée, il indique spécifiquement les régions et les directions des conflits de distributions.

[Tuesday May 28/mardi 28 mai, 14:00-14:30]

Greg Rice (University of Waterloo) , **Piotr Kokoszka** (Colorado State University) , **Han Lin Shang** (Australian National University) , **Yuqian Zhao** (University of Waterloo) , **Tony Wirjanto** (University of Waterloo)

Inference for the Autocovariance of a Functional Time Series and Goodness-Of-Fit Tests for fGARCH Models

Inférence pour l'autocovariance d'une série chronologique fonctionnelle et tests de qualité de l'ajustement de modèles fGARCH

Most methods for analyzing functional time series rely on the estimation of lagged autocovariance operators or surfaces. Testing whether or not such operators are zero is an important diagnostic step that is well understood when the data, or model residuals, form a strong white noise. When functional data are constructed from dense records of, for example, asset prices or returns, a weak white noise model allowing for conditional heteroscedasticity is often more realistic. Applying inferential procedures for the autocovariance based on a strong white noise to such data often leads to the er-

La plupart des méthodes d'analyse de séries chronologiques fonctionnelles reposent sur l'estimation d'opérateurs ou de surfaces d'autocovariance décalés. Il est important pour le diagnostic de tester si ces opérateurs sont nuls ou non, étape bien comprise lorsque les données ou les résidus du modèle forment un bruit blanc fort. Mais lorsque les données fonctionnelles sont construites à partir d'enregistrements denses, par exemple de prix ou de rendements d'actifs, un modèle à bruit blanc faible permettant une hétéroscédasticité conditionnelle est souvent plus réaliste. L'application à de telles données de procédures inférentielles pour l'autocovariance basées sur un bruit blanc fort mène souvent à la

roneous conclusion that the data exhibit significant autocorrelation. We develop methods for performing inference for the lagged autocovariance operators of stationary functional time series that are valid under general conditional heteroscedasticity conditions, and apply these to conduct goodness-of-fit tests for fGARCH models.

conclusion erronée que les données présentent une autocorrélation importante. Nous développons des méthodes d'inférence pour les opérateurs d'autocovariance décalés de séries chronologiques fonctionnelles stationnaires qui sont valables pour des conditions d'hétéroscédasticité conditionnelle générale, et les appliquons pour vérifier la qualité de l'ajustement de modèles fGARCH.

[Tuesday May 28/mardi 28 mai, 14:30-15:00]

Jiguo Cao (Simon Fraser University) , **Fei Jiang** (University of Hong Kong) , **Seungchul Baek** (University of South Carolina) , **Yanyuan Ma** (Penn State University)

Functional Single Index Model

Modèle à indice fonctionnel simple

We propose a semiparametric functional single index model to study the relationship between a univariate response and multiple functional covariates. The parametric part of the model integrates the functional linear regression model and the sufficient dimension reduction structure. The nonparametric part of the model allows the response-index dependence or the link function to be unspecified. The B-spline method is used to approximate the coefficient function, which leads to a dimension folding type model. A new kernel regression method is developed to handle the dimension folding model, which allows the efficient estimation of both the index vector and the B-spline coefficients. We also establish the asymptotic properties and semiparametric optimality for the estimators.

Nous proposons un modèle à indice fonctionnel simple semiparamétrique pour étudier la relation entre une réponse univariée et des covariables fonctionnelles multiples. La partie paramétrique du modèle intègre le modèle de régression linéaire fonctionnelle et la structure de réduction de dimension suffisante. La partie non paramétrique du modèle permet que la réponse et l'indice dépendent l'un de l'autre ou que la fonction de liaison reste indéterminée. Nous utilisons la méthode B-spline pour estimer la fonction de coefficient, ce qui produit un modèle de type de pli de dimension. Nous développons une nouvelle méthode de régression de type kernel pour traiter le modèle de pli de dimension, qui permet une estimation efficace du vecteur d'indice et des coefficients B-spline. Nous déterminons les propriétés asymptotiques et l'optimalité semiparamétrique des estimateurs.

New statistical techniques in exploring modern data with complex structure
Nouvelles techniques statistiques pour l'exploration de données modernes à structure complexe

Chair/Président: Linglong Kong

Organizer/Responsable: Linglong Kong

Room/Salle: 102 (ICT)

Abstract/Résumé

[Tuesday May 28/mardi 28 mai, 13:30-14:00]

Kai Zhang (University of North Carolina at Chapel Hill)

Binary Expansion Testing (BET) on Independence

Tests d'expansion binaire sur l'indépendance

We study the problem of nonparametric dependence detection. Many existing methods may suffer severe power loss due to non-uniform consistency, which we illustrate with a paradox. To avoid such power loss, we approach the nonparametric test of independence through the new framework of binary expansion statistics (BEStat) and binary expansion testing (BET), which examine dependence through a novel binary expansion filtration approximation of the copula. Through a Hadamard transform, we find that the symmetry statistics in the filtration are complete sufficient statistics for dependence. These statistics are also uncorrelated under the null. By utilizing symmetry statistics, the BET avoids the problem of non-uniform consistency and improves upon a wide class of commonly used methods (a) by achieving the minimax rate in sample size requirement for reliable power and (b) by providing clear interpretations of global relationships upon rejection of independence.

Nous étudions le problème de la détection non paramétrique des dépendances. De nombreuses méthodes actuelles peuvent subir de graves pertes de puissance en raison d'une convergence non uniforme, ce que nous illustrons au moyen d'un paradoxe. Pour éviter une telle perte de puissance, nous abordons le test non paramétrique d'indépendance par le nouveau cadre de statistiques d'expansion binaire et de tests d'expansion binaire, qui examinent la dépendance au moyen d'une nouvelle approximation par filtration à expansion binaire de la copule. Grâce à une transformée d'Hadamard, nous constatons que les statistiques de symétrie dans la filtration sont complètes et exhaustives pour la dépendance. De plus, ces statistiques ne sont pas corrélées sous l'hypothèse nulle. En utilisant des statistiques de symétrie, le test d'expansion binaire permet d'éviter le problème de convergence non uniforme et d'améliorer de nombreuses méthodes couramment utilisées a) en atteignant le taux minimum requis en matière de taille d'échantillon pour une puissance fiable et b) en fournissant des interprétations claires des relations globales lors du rejet de l'indépendance.

[Tuesday May 28/mardi 28 mai, 14:00-14:30]

Ruoqing Zhu (University of Illinois, Urbana-Champaign) , **Wenzhuo Zhou** (University of Illinois Urbana-Champaign)

Semiparametric Models for Personalized Dose Finding

Modèles semi-paramétriques pour la détermination de la posologie personnalisée

Learning an individualized dose rule (IDR) in personalized medicine is a challenging statistical problem. Existing methods for estimating the optimal IDR often suffer from the curse of dimensionality, especially when the IDR is learned nonparametrically using machine learning approaches. To tackle this problem, we propose a semiparametric framework. We postulate that the IDR can be reduced to a nonparametric function which relies on only a few linear combinations of the original covariates, hence leading to a more parsimonious model. Based on this framework, we propose two ap-

L'apprentissage d'une règle de posologie personnalisée en médecine personnalisée est un problème statistique complexe. On reproche souvent la dimensionnalité des méthodes existantes d'estimation de la règle de posologie personnalisée optimale, en particulier lorsque la règle de posologie personnalisée est apprise de manière non paramétrique au moyen d'approches d'apprentissage machine. Pour résoudre ce problème, nous proposons un cadre semi-paramétrique. Nous supposons que la règle de posologie personnalisée peut être réduite à une fonction non paramétrique qui ne repose que sur quelques combinaisons linéaires des covariables d'origine, ce qui conduit à un modèle plus parcimonieux.

proaches, a direct learning approach that yields the IDR, and a pseudo-direct learning approach that sets a connection with the partial dimension reduction space. Our estimator is solved using an orthogonality constrained optimization approach on the Stiefel manifold. The performances of the proposed methods are evaluated through simulation studies and a warfarin pharmacogenetic dataset.

En fonction de ce cadre, nous proposons deux approches : une approche d'apprentissage direct qui donne la règle de posologie personnalisée, et une approche d'apprentissage pseudo-direct qui établit un lien avec l'espace de réduction dimensionnelle partielle. Notre estimateur est résolu à l'aide d'une approche d'optimisation à contrainte d'orthogonalité sur la variété de Stiefel. Nous évaluons l'efficacité des méthodes proposées par des études de simulation et un ensemble de données pharmacogénétiques sur la warfarine.

[Tuesday May 28/mardi 28 mai, 14:30-15:00]

Zhengwu Zhang (University of Rochester) , **Xiao Wang** (Purdue University) , **Hongtu Zhu** (University of North Carolina) , **Linglong Kong** (University of Alberta)

High-Dimensional Spatial Quantile Function-On-Scalar Regression

Régression quantile spatiale fonctions-scalaires en haute dimension

We develop a novel spatial quantile function-on-scalar regression model, which is used to describe the conditional spatial distribution of a high-dimensional functional response given scalar predictors. With the strength of both quantile regression and copula modeling, we are able to explicitly characterize the conditional distribution of the functional or image response on the whole spatial domain. Our method provides a comprehensive understanding of the effect of scalar covariates at different quantile levels and also gives a practical way to generate new images for given covariate values. Theoretically, we establish the minimax rates of convergence for estimating coefficient functions under both fixed and random designs. We further develop an efficient primal-dual algorithm to handle high-dimensional image data. Simulations and real data analysis are conducted to examine the finite-sample performance.

Nous développons un nouveau modèle de régression quantile spatiale fonctions-scalaires, qui étudie la distribution spatiale conditionnelle d'une réponse fonctionnelle en haute dimension avec des prédicteurs scalaires. Grâce à la force de la régression quantile et de la modélisation en copules, nous pouvons explicitement caractériser la distribution conditionnelle de la réponse fonctionnelle ou d'image sur l'ensemble du domaine spatial. Notre méthode permet de bien comprendre l'effet des covariables scalaires à divers niveaux de quantiles et permet également, en pratique, de générer de nouvelles images pour des valeurs de covariables données. De manière théorique, nous déterminons les taux de convergence minimax pour l'estimation des fonctions de coefficient dans des plans fixes ou aléatoires. Nous développons ensuite un algorithme primal-dual efficace pour traiter les données d'imagerie de haute dimension. Nous effectuons des simulations et une analyse de données réelles pour examiner la performance sur échantillon fini.

Recent developments in quantitative psychology/psychometrics
Récentes évolutions en psychologie quantitative / psychométrie

Chair/Président: Heungsun Hwang

Organizer/Responsable: Heungsun Hwang

Room/Salle: 116 (ICT)

Abstract/Résumé

[Tuesday May 28/mardi 28 mai, 13:30-14:00]

Heungsun Hwang (McGill University)

Imaging Genetics Structural Equation Modeling for Examining Gene-Brain-Behavioural/Cognitive Relationships

Modélisation d'équations structurelles d'imagerie génétique pour l'étude des relations gène-cerveau-comportement/cognition

With advances in neuroimaging and genetics, imaging genetics is a naturally emerging field that integrates genetic and imaging data with behavioural or cognitive outcomes to investigate genetic influence on altered brain activities, which are in turn associated with behavioural or cognitive variation. This field is faced with an ever-increasing need for statistical tools to examine such gene-brain-behaviour/cognition (G-B-B/C) relationships in a more unified manner, while taking into account biological complexities (e.g., genetic networks, gene-gene or gene-environment interactions) and methodological issues (e.g., multicollinearity). Thus, we develop a general statistical approach, named Imaging Genetics Structural Equation Modeling (IGSEM), which allows specifying and testing various biologically plausible G-B-B/C relationships based on previous theories or knowledge. We will present the conceptual underpinnings of IGSEM and its illustrative applications.

Avec les progrès actuels en neuroimagerie et génétique, l'imagerie génétique est un domaine émergent qui intègre les données génétiques et d'imagerie avec des résultats comportementaux ou cognitifs pour étudier les facteurs génétiques qui influencent une modification des activités du cerveau, à leur tour associées à des variations comportementales ou cognitives. Ce domaine fait face à un besoin toujours croissant d'outils statistiques pour examiner ces relations gène-cerveau-comportement/cognition (G-C-C/C) de manière plus uniforme, tout en tenant compte de complexités biologiques (p. ex., réseaux génétiques, interactions gène-gène ou gène-environnement, etc.) et de problèmes méthodologiques (p. ex., multicolinéarité). Nous développons une approche statistique générale, la modélisation d'équations structurelles d'imagerie génétique (MESIG), qui permet de spécifier et de tester diverses relations G-C-C/C biologiquement plausibles selon nos théories ou connaissances préalables. Nous présenterons les fondements conceptuels de la MESIG et l'illustrerons par des applications.

[Tuesday May 28/mardi 28 mai, 14:00-14:30]

James O. Ramsay (McGill University) , **Marie Wiberg** (Umea University) , **Juan Li** (Ottawa Hospital Research Institute)

Efficient Scoring of Test Data

Notation efficace des données d'essai

When responses to tests or scales are quantified by summing pre-allocated fixed option weights, an important source of information is ignored. This is the variation from item to item in the shape of option and item characteristic curves. That is, sum-scoring uses row information but ignores column information in the response matrix. Sum-scoring also ignores the information available in wrong option choices. Efficient estimates of an examinee's ability have been available since with early 1950s, but have yet to see widespread and routine use

Lorsque les réponses aux tests ou aux échelles sont quantifiées en faisant une somme des poids pré-alloués à option fixe, une source importante d'information est ignorée, c'est-à-dire la variation entre items sous forme de courbes d'option et de courbes caractéristiques des items. En effet, la sommation des scores utilise l'information en lignes mais ignore l'information en colonnes de la matrice des réponses. La sommation des scores ignore aussi l'information disponible dans les mauvais choix d'options. Des estimations efficaces des compétences des candidats sont disponibles depuis le début des années 50 mais ne sont toujours

in educational measurement, even by large-scaling testing agencies, where the number right or sum score still prevails. Improvements in root-mean-square error for examinees near the median score are about 25% for designed achievement tests, and substantially more for typical classroom tests. Moreover, improvements of efficiency for the extremely important cohort at the highest performance level are far higher.

pas utilisées de façon généralisée et routinière dans le milieu de l'éducation, même par des agences de contrôle à grande échelle, où la somme des scores continue de prévaloir. L'amélioration de l'erreur quadratique moyenne pour les candidats près de la médiane est d'environ 25% pour les tests de compétences et est considérablement plus élevée pour les tests typiques en classe. De plus, l'amélioration de l'efficacité pour la cohorte très importante au plus haut niveau de performance est beaucoup plus élevée.

[Tuesday May 28/mardi 28 mai, 14:30-15:00]

Carl F. Falk (McGill University) , **Leah M. Feuerstahler** (Fordham University)

On the Performance of Semi- and Non-Parametric Item Response Functions in Computer Adaptive Tests

De la performance des fonctions semiparamétriques et non paramétriques de réponse aux items dans les tests adaptatifs informatisés

Large scale assessments often use a computer adaptive test (CAT) for selection of items and for scoring respondents. Such tests often assume a parametric form for the relationship between item responses and the underlying construct. Although semi- and non-parametric response functions could be used, there is scant research on their performance in a CAT. In this work, we compare parametric response functions versus those estimated using kernel smoothing (KS) and a logistic function of a monotonic polynomial (LMP). We argue that LMP items can be used with traditional CAT item selection algorithms that use derivatives whereas KS items require a different approach. We compared these approaches in CAT simulations with a variety of item selection algorithms. Our simulations varied sample size, the presence of missing data, and the percentage of non-standard items. In general, results support the use of semi- and non-parametric item response functions in a CAT.

Les évaluations à grande échelle font souvent appel à un test adaptatif informatisé (TAI) pour le choix des items et la notation des répondants. Ces tests prennent souvent une forme paramétrique pour la relation entre les réponses aux items et le concept sous-jacent. Même s'il est possible d'utiliser des fonctions semiparamétriques et non paramétriques de réponse aux items, rares sont les recherches qui ont porté sur leur performance dans un TAI. Notre travail permet la comparaison des fonctions de réponse paramétriques à celles estimées par lissage de noyau (LN) ainsi qu'à une fonction logistique d'un polynôme monotone (LPM). Nous soutenons qu'il est possible d'utiliser les items LPM avec des algorithmes de sélection d'items d'un TAI traditionnel qui fait appel à des dérivés, tandis qu'il faut adopter une approche différente pour des items LN. Nous avons comparé ces approches à l'aide de simulations TAI et divers algorithmes de sélection d'items. Nos simulations comportaient diverses tailles d'échantillonnage, des données manquantes et le pourcentage d'items non normatifs. Les résultats appuyaient généralement l'utilisation de fonctions semiparamétriques et non paramétriques de réponse aux items dans un test adaptatif informatisé.

Advanced statistical methods for the integration of omic data
Méthodes statistiques avancées pour l'intégration de données -omiques

Chair/Président: Thierry Chekouo

Organizer/Responsable: Thierry Chekouo

Room/Salle: 201 (ENG)

Abstract/Résumé

[Tuesday May 28/mardi 28 mai, 13:30-14:00]

Sandra Safo (University of Minnesota) , **Thierry Chekouo** (University of Calgary)

Bayesian Integrative Analysis Method with Incorporation of Grouping Information

Méthode bayésienne intégrative d'analyse avec incorporation d'information de regroupement

Advances in data collection and processing in biomedical research allow different data types to be measured on the same subjects, each representing different sets of characteristics, but collectively helping to explain underlying complex mechanisms. In some instances, phenotypic data are available. The main goal is to study the overall dependency structure among the data types, and to develop a model for predicting future phenotypes. We present a Bayesian factor analysis method that simultaneously models the overall association between data types using only relevant variables, while also predicting future outcomes using factor loadings. Through prior distributions, we incorporate structural information (e.g., biological networks) in our model that allows us to select functionally meaningful networks involved in the determination of factor loadings. We demonstrate the effectiveness of the proposed approach using simulations and observed data.

Les progrès dans la collection et le traitement des données en recherche biomédicale permettent de mesurer divers types de données relatives aux mêmes sujets. Chaque type de données mesure différents ensembles de caractéristiques, mais tous contribuent collectivement à expliquer des mécanismes complexes sous-jacents. Dans certains cas, des données phénotypiques sont aussi disponibles. Avec ces problèmes, nous visons principalement à étudier la structure de dépendance globale parmi les types de données et à élaborer un modèle prédictif de futurs phénotypes. Nous présentons une méthode d'analyse bayésienne de facteurs qui modélise simultanément l'association globale entre les types de données, en utilisant seulement des variables pertinentes, tout en prédisant des résultats ultérieurs à l'aide de coefficients de saturation. À l'aide de distributions a priori, nous incorporons à l'information structurelle de notre modèle (par ex. : les réseaux biologiques) qui permet la sélection de réseaux participant à la détermination des coefficients de saturation. Nous illustrons l'efficacité de l'approche proposée à l'aide de simulations et de données observées.

[Tuesday May 28/mardi 28 mai, 14:00-14:30]

Francesco Claudio Stingo (University of Florence)

Bayesian Data Integration in Cancer Genomics

Intégration de données bayésiennes dans le domaine de la génomique du cancer

Identifying patient-specific prognostic biomarkers is of critical importance in developing personalized treatment for clinically and molecularly heterogeneous diseases such as cancer. We propose a novel regression framework, Bayesian hierarchical varying-sparsity regression (BEHAVIOR) models, to select clinically relevant disease markers by integrating proteogenomic (proteomic+genomic) and clinical data. Our methods allow flexible modeling of protein-gene relationships and induce sparsity in both protein-gene and protein-survival relationships to select genomically driven prognostic

L'identification de biomarqueurs pronostiques propres à chaque patient est cruciale dans la mise au point d'un traitement personnalisé pour des maladies cliniquement et moléculairement hétérogènes, comme le cancer. Nous proposons un nouveau cadre de régression, les modèles de régression hiérarchiques bayésiens à parcimonie variable, pour sélectionner des marqueurs de maladies cliniquement pertinents en intégrant des données protéogénomiques (protéomiques et génomiques) et cliniques. Nos méthodes permettent une modélisation souple des relations protéines-gènes et entraînent une parcimonie des relations protéines-gènes et protéines-survie pour sélectionner des mar-

protein markers at the patient-level. Simulation studies demonstrate the superior performance of BEHAVIOR against a competing method in terms of both protein marker selection and survival prediction.

queurs de protéines pronostiques génomiques au niveau du patient. Les études de simulation démontrent l'efficacité supérieure des modèles de régression hiérarchiques bayésiens à parcimonie variable par rapport à une méthode concurrente en ce qui concerne la sélection des marqueurs protéiques et la prédiction de la survie.

[Tuesday May 28/mardi 28 mai, 14:30-15:00]

Discussion

Modeling Risk
Risque de la modélisation

Chair/Président: Shu Li

Room/Salle: 105 (SB)

Abstract/Résumé

[Tuesday May 28/mardi 28 mai, 13:30-13:45]

Étienne Marceau (Université de Laval) , **Hélène Cossette** (Université Laval) , **Julien Trufin** (Université Libre de Bruxelles) , **Pierre Zuyderhoff** (Université Libre de Bruxelles)

Ruin-Based Risk Measures: Properties and Capital Allocation

Mesures de risque basées sur la ruine : propriétés et allocation de capital

We study the properties of ruin-based risk measures, which are defined within the general class of discrete-time risk models for an insurance portfolio. Desirable properties of the ruin-based risk measures, such as homogeneity, subadditivity, convexity, consistency to the multivariate usual stochastic order, consistency to the multivariate increasing convex order, and consistency to the supermodular order, are examined. Specific risk models are considered. We also apply some of our ruin-based risk measures to identify the contributions of the components to riskiness of an insurance portfolio.

Dans cet article, nous étudions les propriétés des mesures de risque basées sur la ruine, qui sont définies dans la classe générale des modèles de risque à temps discret pour un portefeuille d'assurance. Les propriétés souhaitables des mesures de risque basées sur la ruine, telles que l'homogénéité, la sous-additivité, la convexité, la cohérence selon l'ordre en dominance stochastique multivarié, la cohérence selon l'ordre convexe croissant multivarié, et la cohérence selon l'ordre supermodulaire, sont examinées. Des modèles spécifiques de risque sont considérés. Nous appliquons également certaines de nos mesures de risque basées sur la ruine pour identifier les contributions des composantes d'un portefeuille au risque global d'un portefeuille d'assurance.

[Tuesday May 28/mardi 28 mai, 13:45-14:00]

Shanoja Naik (Laurentian University) , **Peter Adamic** (Laurentian University)

Stochastic Cause-Deleted Life Expectancy for Multiple Risks

Espérance de vie stochastique avec cause de mortalité éliminée pour risques multiples

The established approach to calculating the cause-deleted life expectancy in a competing risks framework is to simply remove the deaths caused by the risks in question from the life table. However, this is clearly untenable, since the lack of presence of a cause of death will take time to be fully realized. To correct for this oversimplification, we propose a multivariate cure-of-cause model that can incorporate probability distributions for the cures of multiple causes of death over time as a means to more accurately predict the overall increase in life expectancy that would ensue at each age. Note that our models are decidedly distinct from what is commonly understood as "cure modeling" in the survival literature.The theoretical results are applied to a real data set involving both diabetes and HIV-related deaths. Future work might include generalizing the model even further to account for dependence between the various cure-of-cause distributions.

L'approche actuelle pour calculer l'espérance de vie avec cause de mortalité éliminée dans un cadre de risques concurrents est simplement d'éliminer de la table de survie les décès causés par les risques en question. Par contre, ceci est clairement intenable, car l'absence de cause de décès va prendre du temps avant d'être pleinement réalisée. Pour corriger cette simplification excessive, nous proposons un modèle multivarié guérison-cause qui intègre les distributions de probabilité pour les guérisons de multiples causes de décès dans le temps et prédit ainsi plus précisément l'augmentation générale de l'espérance de vie qui en résulterait à chaque âge. Il faut noter que nos modèles sont nettement distincts de ce qui est généralement considéré comme ≪ modélisation de guérison ≫ dans la littérature de survie. Les résultats théoriques sont appliqués à des données réelles sur les décès liés au diabète et au VIH. Les travaux futurs pourraient inclure une plus grande généralisation du modèle pour tenir compte de la dépendance entre les différentes distributions guérison-cause.

[Tuesday May 28/mardi 28 mai, 14:00-14:15]

Modeling Risk
Risque de la modélisation

Daniel Hadley (University of British Columbia) , **Natalia Nolde** (University of British Columbia) , **Harry Joe** (University of British Columbia)

The Selection of Loss Severity Distributions to Model Operational Risk

Choix des distributions de la gravité des pertes pour modéliser le risque opérationnel

Accurate modeling of operational risk is important for financial institutions to prepare for potentially catastrophic losses. The loss distribution approach requires losses to be grouped into risk categories and loss frequency and loss severity distributions selected for each category. The annual operational loss distribution is estimated as the compound sum of losses over each risk category, and regulatory capital equal to the 0.999-quantile of this distribution is set aside. In practice, this approach can produce unstable regulatory capital year-to-year as the selected loss severity distribution family changes. We promote using truncation probability estimates and a consistent quantile scoring function on annual loss data as criteria for selecting severity distributions. Additionally, the Sinh-arcSinh distribution provides a flexible family for modeling loss severities. Finally, we investigate the effect of collecting and analyzing all loss severities.

Une modélisation précise du risque opérationnel est importante pour les institutions financières afin de se préparer à des pertes potentiellement catastrophiques. L'approche de distribution des pertes exige que les pertes soient regroupées en catégories de risques et que les distributions de fréquence et de gravité des pertes soient choisies pour chaque catégorie. On estime la distribution annuelle des pertes opérationnelles comme la valeur acquise des pertes sur chaque catégorie de risque, et on met de côté un capital réglementaire égal à quantile 0,999 de cette distribution. Dans la pratique, cette approche peut produire un capital réglementaire instable d'une année à l'autre à mesure que la famille de distribution de la gravité des sinistres choisie change. Nous encourageons l'utilisation d'estimations de probabilités de troncature et d'une fonction de notation de quantile convergente sur les données de pertes annuelles comme critères de sélection des distributions de gravité. De plus, la distribution Sinh-arcSinh fournit une famille flexible permettant de modéliser la gravité des pertes. Enfin, nous étudions l'effet de la collecte et de l'analyse de toutes les gravités de pertes.

[Tuesday May 28/mardi 28 mai, 14:15-14:30]

Anne Mackay (Université du Québec à Montréal) , **Michael Kouritzin** (University of Alberta)

Simulating the Heston Model Using Explicit Weak Solutions

Simuler le modèle de Heston en utilisant des solutions faibles explicites

In this presentation, I discuss new simulation algorithms for the Heston model, which are based on recent results which show that the Heston model presents explicit weak solutions that can be used for simulating volatilities and prices. Most often, efficient simulation is done under an artificial reference probability and then converted to the real probability with the appropriate likelihood. The resulting simulation algorithm can therefore be seen as the analog of a weighted particle filter. It is then natural to introduce some type of resampling to improve the performance of the simulation algorithm. Here we focus on recently developed branching algorithms, which have the advantage of preserving the historical property of the particle system. Through numerical results, we illustrate the increased performance and accuracy due to branching. We also compare the resulting simulation algorithm to popular Heston simulation methods.

Dans cette présentation, j'explore de nouveaux algorithmes de simulations pour le modèle de Heston. Ces algorithmes sont basés sur de récents résultats qui montrent que le modèle de Heston présente des solutions faibles explicites qui peuvent être utilisées pour simuler le prix d'une action et sa volatilité. Ces simulations sont généralement effectuées sous une mesure de probabilité artificielle, pour ensuite être ramenées à la vraie probabilité à l'aide de la vraisemblance appropriée. L'algorithme de simulation qui en résulte pouvant être vu comme un filtre à particules pondéré, il est alors naturel d'y introduire du ré-échantillonnage afin d'en améliorer la performance. Dans cette présentation, nous utilisons des algorithmes de branchement développés récemment, ayant l'avantage de conserver la propriété historique du système de particules simulé. Des résultats numériques illustrent l'amélioration de la performance possible grâce aux algorithmes de branchement. Nous comparons également l'algorithme de simulation qui en résulte avec les méthodes plus connues de simulation de Heston.

[Tuesday May 28/mardi 28 mai, 14:30-14:45]

Modeling Risk
Risque de la modélisation

Jose Garrido (Concordia University) , **Deive Ciro de Oliveira** (ICSA - UNIFAL-MG: Federal University of Alfenas)

Hidden Markov Over-Dispersed Poisson Models Applied to Highways Accident Counts

Modèles de Poisson surdispersés par Markov caché appliqués au dénombrement d'accidents routiers

High accident rates are observed on Brazilian roads. Among other factors, the number of accidents is highly correlated to space and location. Hidden Markov Models (HMMs) are useful to find and measure differences between dangerous stretches along highways. Here we use 2-state HMMs with Poisson, Borel-Tanner and Lagrange Poisson distributions, fitted to Brazilian highway accident data on route BR-381. This dataset lists 1,379 accidents occurred along a 449.1-kilometer segment of BR-381, stretching from Belo Horizonte to Extrema, both being cities in Minas Gerais state, in southeast Brazil. The data is over-dispersed and gives accident counts in sections of 0.1-kilometer granularity, so there are 4,491 spots with accident counts. MLE estimation of the parameters in HMMs with these 3 distributions is obtained by the EM Algorithm. HMMs with Lagrange-Poisson outperform other models. The dangerous sections (hot spots) are identified and the profile of safe locations is provided.

On observe un taux élevé d'accidents routiers sur les routes brésiliennes. On remarque une corrélation élevée entre le nombre d'accidents, l'emplacement et l'espace, pour ne nommer que certains facteurs. Les modèles de Markov cachés (MMC) sont pratiques pour trouver et mesurer les différences entre les points dangereux sur les routes. Nous adoptons ici des MMC à deux états avec les lois de Poisson, de Borel-Tanner et de Lagrange-Poisson ajustés aux données des accidents routiers au Brésil sur la route BR-381. Ce jeu de données compte 1379 accidents survenus sur une section de 449,1 km de la route BR-381 qui s'étend de Belo Horizonte jusqu'à Extrema ; deux villes dans l'État du Minas Gerais, au sud-est du Brésil. Les données sont surdispersées et dénombrent les accidents en sections de granularité de 0,1 km, nous dénombrons donc des accidents à 4491 points. Au moyen de l'algorithme espérance-maximisation (EM), on obtient l'estimation de l'estimateur du maximum de vraisemblance (EMV) des paramètres dans les MMC avec ces trois lois. Les MMC avec lois de Lagrange-Poisson surpassent les autres modèles. Nous identifions les points dangereux (zones à risque) et établissons le profil des zones sécuritaires.

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

Chair/Président: Alomgir Hossain

Room/Salle: 142 (AD)

Abstract/Résumé

[Tuesday May 28/mardi 28 mai, 13:30-13:45]

Yunqi Ji (Alberta Health Services) , **Jerry Ren** (Alberta Health Services) ,

Using Administrative Data for Health Services Planning: Healthcare Utilization Projection
L'emploi de données administratives pour planifier les services de santé : une projection de l'utilisation des soins de santé

In healthcare planning, administrative health data are an important source to analyze healthcare utilization in sectors of inpatient, outpatient and continuing care. The historical data from administrative databases combined with population projection data can be used to project future healthcare utilization, and advise on future investments in healthcare to meet the demands of a growing and aging population. We will introduce how to use administrative data through a base case model developed to consider age, sex, and geographic areas into the projection in the acute care scenario. The base case model is extended further to accommodate potential hospital visits and utilization trends for different programs and clinical groups to obtain more realistic projections. In addition, we also introduce the concept of using simulation modeling methods to consider avoidable utilization in future projections.

Dans le cadre de la planification des soins de santé, les données administratives sur la santé représentent une source importante pour analyser l'utilisation des soins de santé concernant les malades hospitalisés, les patients externes et les soins continus. Les données historiques tirées des bases de données administratives combinées aux données de projections démographiques peuvent servir à prévoir l'utilisation des soins de santé à venir et à guider les recommandations relatives aux futurs investissements dans les soins de la santé pour combler la demande dans le contexte d'un vieillissement démographique grandissant. Nous présenterons une façon de se servir des données administratives par l'entremise d'un scénario de base qui tient compte de l'âge, du sexe et des zones géographiques pour prévoir un scénario de soins actifs. Dans le but d'obtenir des projections plus réalistes, nous avons élargi l'étendue du scénario de référence pour qu'il s'adapte aux visites potentielles dans les hôpitaux et aux tendances d'utilisation selon différents programmes et groupes cliniques. De plus, nous présenterons aussi le concept qui consiste à employer des méthodes de modélisation par simulation pour tenir compte des utilisations évitables dans les projections à venir.

[Tuesday May 28/mardi 28 mai, 13:45-14:00]

Leif Erik Lovblom (University of Toronto) , **Nicholas Mitsakakis** (University of Toronto)

Exponential Dispersion Models for Healthcare Cost Data

Modèles de dispersion exponentielle pour les données relatives au coût des soins de santé

The structure of healthcare cost data presents many analytical challenges. It is difficult to separate patients with large costs due to frequent utilization from those with infrequent but very costly utilization. The presence of zero-cost values for non-users can be high. A potential statistical modelling approach for these data uses the Tweedie case of exponential dispersion models; these models have been used for actuarial studies and survival analysis, but have seen limited use for healthcare cost data. Tweedie distributions are a sum of n independent gamma random variables where n follows a Poisson distribution; they can be analyzed using a GLM approach.

La structure des données relatives au coût des soins de santé comprend de nombreux défis analytiques. Il est difficile de distinguer les patients assumant de grandes dépenses en raison d'une utilisation fréquente de ceux qui s'en servent moins fréquemment, mais dont les coûts sont plus élevés. La présence de valeurs à coût zéro peut être élevée dans les cas de non-usage. Une approche de modélisation statistique potentielle pour ces données se sert des modèles de dispersion exponentielle de Tweedie ; ces derniers ont été employés dans le cadre d'études actuarielles et d'analyses de survie, mais ont rarement été adoptés pour les données relatives au coût des soins de santé. Les distributions Tweedie sont une somme de n variables aléatoires indépendantes gamma, où n suit une loi

In addition, they potentially allow for the decomposition of the variance into one part associated with frequency and another associated with severity. Using simulated data and an existing dataset of healthcare costs associated with prostate cancer treatment in Ontario, we aimed to assess estimation by Tweedie models.

de Poisson. Il est d'ailleurs possible d'analyser ces distributions au moyen du modèle linéaire général (MLG). De plus, elles peuvent potentiellement rendre possible la décomposition de la variance en une partie associée à la fréquence et une autre partie associée à la sévérité. Au moyen de données simulées et d'un vrai jeu de données portant sur les coûts associés au traitement du cancer de la prostate en Ontario, nous visons à évaluer l'estimation à partir des modèles de Tweedie.

[Tuesday May 28/mardi 28 mai, 14:00-14:15]

Madeline Ward (University of Guelph) , **Anu Stanley** (University of Guelph) , **Lorna Deeth** (University of Guelph) , **Rob Deardon** (University of Calgary) , **Zeny Feng** (University of Guelph) , **Lise Trotz-Williams** (Wellington-Dufferin-Guelph Public Health)

Evaluation of School Absenteeism Surveillance Systems for Influenza Outbreaks in Wellington-Dufferin-Guelph, Ontario

Évaluation des systèmes de surveillance de l'absentéisme scolaire lors d'éclosions de grippes à Wellington-Dufferin-Guelph en Ontario

Wellington-Dufferin-Guelph Public Health uses daily reported school absenteeism data to aid in detecting the start of influenza season. When absence in excess of 10% is reported, an alarm is raised and investigated. However, generally this results in many false alarms. To attempt to reduce the required alarm follow-up time while maintaining advanced notice prior to an outbreak, several model-based approaches for predicting the onset of the influenza season were explored. These included an exponentially weighted moving average model, logistic regression with and without seasonality terms and random intercepts for school year, and a generalized estimating equation model. Different aggregations and lags (in the regression models) of absenteeism as a predictor variable were considered, and performance of each surveillance model was evaluated using a false alarm rate and a measure of true alarm timeliness.

La santé publique de Wellington-Dufferin-Guelph utilise les données quotidiennes d'absentéisme scolaire pour aider dans la détection du début de la saison de la grippe. Lorsqu'un taux d'absence de plus de 10% est signalé, une alerte est déclenchée et enquêtée. Par contre, ceci donne souvent lieu à de fausses alarmes. Pour essayer de réduire la durée du suivi requis tout en maintenant le préavis avant une épidémie, plusieurs approches fondées sur un modèle ont été explorées pour prédire le début de la saison de la grippe. Parmi celles-ci il y a eu le modèle de la moyenne mobile pondérée exponentiellement, la régression logistique avec et sans terme de saisonnalité et interceptions aléatoires pour l'année scolaire ainsi qu'un modèle d'estimation d'équation généralisé. Différents agrégations et décalages (dans les modèles de régression) de l'absentéisme comme variable prédictive ont été considérés et la performance de chaque modèle de surveillance a été évaluée en utilisant un faux taux d'alerte et une mesure de la rapidité d'une vraie alerte.

[Tuesday May 28/mardi 28 mai, 14:15-14:30]

Justin Wayne Dyck (University of Manitoba) , **Mahmoud Torabi** (University of Manitoba)

Statistical Models for Spatially Misaligned Data: An Application to Ischemic Heart Disease in Manitoba

Modèles statistiques pour données spatiales désalignées : une application aux maladies cardiaques ischémiques au Manitoba

Disease mapping of spatially referenced data is an important area of epidemiology, as it provides valuable information for policy makers and service providers. Modeling of geographically aggregated data provides methods for detection of spatial disease patterns. Often data is collected at different aggregation levels and can therefore be misaligned in space. A motivating example is the usage of administrative data, which is aggregated by postal code, to identify ischemic heart disease

La cartographie des maladies de données spatiales référencées est un domaine important de l'épidémiologie car elle fournit des renseignements importants aux décideurs politiques et aux fournisseurs de services. La modélisation de données agrégées géographiquement fournit des méthodes pour la détection de profils épidémiologiques spatiaux. Les données sont souvent recueillies à différents niveaux d'agrégation et peuvent donc être désalignées dans l'espace. Un exemple est l'utilisation de données administratives, qui sont agrégées par codes postaux, pour l'iden-

(IHD) rates. Census data can provide valuable information on population demographics and economic status. However, the census dissemination areas are constructed differently than postal code areas, and hence no one-to-one conversion exists. In this talk, we develop an approach for predicting values of a variable from a misaligned and geographically aggregated dataset, such as the Census, so it can be used as an accurate predictor for a response variable (i.e., IHD) from a differently aggregated dataset.

tification des taux de maladie cardiaque ischémique (MCI). Les données du recensement peuvent fournir de l'information importante sur la démographie et le statut économique de la population. Par contre, les aires de diffusion du recensement sont construites différemment des régions des codes postaux, par conséquent aucune conversion directe n'existe. Dans cet exposé, nous développons une approche pour prédire les valeurs d'une variable provenant d'un ensemble de données désalignées agrégées géographiquement, tel que le recensement, pour qu'elle puisse être utilisée comme prédicteur précis d'une variable de réponse (c'est-à-dire MCI) à partir d'un ensemble différent de données agrégées.

[Tuesday May 28/mardi 28 mai, 14:30-14:45]

Jinhui Ma (McMaster University) , **Hon Yiu So** (University of Waterloo) , **Lauren Griffith** (McMaster University) , **Cynthia Balion** (McMaster University) , **Mylinh Doung** (McMaster University) , **Carol Bassim** (McMaster University) , **Chris Verschoor** (McMaster University) , **Edwin van den Heuvel** (Eindhoven University of Technology) , **Parminder Raina** (McMaster University)

Imputation Strategies for Handling Missing Spirometry Data in Population-Based Studies

Stratégies d'imputation pour le traitement des données spirométriques manquantes dans les études basées sur une population

Spirometry has increasingly been utilized in population-based studies to assess the presence of lung disease at the population level. Though spirometry testing is not an invasive procedure, participants with conditions which place them at high risk for adverse effects are often excluded from spirometry tests. Moreover, the assurance of high-quality spirometry testing remains challenging in population-based studies. It is common that 10 to 30% of participants in population-based studies do not have valid spirometry data due to exclusion criteria or poor quality of the spirometry test. Missing spirometry data can lead to serious bias and undermine the validity of research results. The objectives of the present study are to identify factors associated with missing spirometry data, assess the potential impact of missing spirometry data on the conclusions drawn from the data, and propose an appropriate strategy to handle them, using the Canadian Longitudinal Study on Aging as an example.

La spirométrie est de plus en plus utilisée dans les études basées sur une population pour évaluer l'incidence de maladies pulmonaires dans la population. Même si le test spirométrique n'est pas une procédure invasive, les participants atteints d'un trouble de santé les mettant à risque de subir des effets indésirables en sont souvent exclus. De plus, rien n'assure entièrement que le test spirométrique dans ce genre d'études soit de grande qualité. Pour 10 % à 30 % des participants aux études basées sur une population, il est fréquent que les données spirométriques obtenues ne soient pas valides en raison de critères d'exclusion ou de la qualité médiocre du test spirométrique. Les données spirométriques manquantes peuvent être une source de biais sérieux et miner la validité des résultats de recherche. La présente étude vise à identifier les facteurs associés aux données spirométriques manquantes, évaluer l'effet potentiel de ces données manquantes sur les conclusions que l'on peut en tirer et proposer une stratégie efficace pour les traiter, avec comme exemple l'Étude longitudinale canadienne sur le vieillissement.

[Tuesday May 28/mardi 28 mai, 14:45-15:00]

Roya Gavanji (University of Saskatchewan) , **Cindy Xin Feng** (University of Saskatchewan) , **Catherine Trask** (University of Saskatchewan)

Identifying Risk Factors Associated with High Risk of Occupational Injury in Saskatchewan Using Machine Learning Methods

Identifier les facteurs de risque associés au risque élevé d'accidents de travail en Saskatchewan en utilisant des méthodes d'apprentissage machine

Occupational fatalities are a serious public health con-

En Saskatchewan (SK), les décès reliés au travail constituent

cern in Saskatchewan (SK), requiring a better understanding of the contributing factors. The objective of this study was to identify risk factors associated with fatal claims using logistic regression and machine learning methods in a large population-based sample of occupational workers. Models were developed using SK workers' compensation board claims data from 2007-2016, with fatality status as the outcome and potential risk factors. Machine learning methods such as lasso, ridge, and elastic net logistic regressions were utilized to analyze the data. Models' predictive accuracy was then compared (sensitivity, specificity, and area under the curve). Ultimately, it is hoped that the result of the study will provide useful information for policymakers to design targeted interventions to reduce the burden of fatal injury claims on the workers, the employers, the healthcare system, and the compensation system.

une préoccupation importante de santé publique nécessitant une meilleure compréhension des facteurs contributifs. L'objectif de cette étude visait l'identification des facteurs de risque reliés aux décès en utilisant la régression logistique et des méthodes d'apprentissage machine pour analyser un échantillon de travailleurs provenant d'une grande population. Des modèles ont été élaborés en utilisant les données des demandes d'indemnisation des travailleurs en Saskatchewan de 2007 à 2016, avec la mortalité comme variable réponse et comme facteurs de risque potentiels. Des méthodes d'apprentissage machine telles que le lasso, le ridge et le filet élastique en régression logistique ont été utilisées pour analyser les données. La précision des prédictions des modèles a alors été comparée (sensibilité, spécificité et aire sous la courbe). Il est souhaité que les résultats de cette étude fourniront de l'information utile aux décideurs pour l'élaboration d'interventions ciblées pour réduire le fardeau des accidents mortels des travailleurs, des employeurs, du système de santé et du système d'indemnisation.

Advances in Estimation Methods
Progrès en matière de méthodes d'estimation

Chair/Président: Mélina Mailhot

Room/Salle: 143 (ST)

Abstract/Résumé

[Tuesday May 28/mardi 28 mai, 13:30-13:45]

Lahiru R. Wickramasinghe (University of Manitoba) , **Alexandre Leblanc** (University of Manitoba) , **Saman Muthukumarana** (University of Manitoba)

Model-Based Estimation of Baseball Batting Metrics

Estimation fondée sur le modèle des mesures de la performance au bâton au baseball

We introduce an approach to model the batting outcomes of baseball batters based on the weighted likelihood approach and make use of our methodology to estimate commonly used baseball batting metrics. The weighted likelihood allows the other batters to contribute to the inference so that the relevant information they contain is not lost and the weights are determined based on their dissimilarities with the target batter. MAMSE weights are used as the likelihood weights. For comparison, we implemented a semiparametric Bayesian approach based on Dirichlet process which enables borrowing information across batters while providing a natural clustering mechanism. We demonstrate and compare these approaches using 2018 Major League Baseball (MLB) batters' data.

Nous présentons une approche pour modéliser les résultats au bâton des frappeurs au baseball fondée sur une approche par vraisemblance pondérée. Puis, nous utilisons notre méthode pour estimer des mesures fréquemment utilisées de la performance au bâton. La vraisemblance pondérée permet aux autres frappeurs de contribuer à l'inférence pour que l'information pertinente qu'ils possèdent ne soit pas perdue et les poids sont déterminés selon leurs différences avec le frappeur cible. Les poids MAMSE sont utilisés pour pondérer la vraisemblance. À des fins de comparaison, nous avons implémenté une approche bayésienne semiparamétrique fondée sur un processus de Dirichlet ce qui permet l'emprunt d'information à travers les frappeurs tout en offrant un mécanisme naturel de regroupement. Nous démontrons et comparons ces approches à l'aides des données sur les frappeurs de la Ligue majeure de baseball (LMB) de 2018.

[Tuesday May 28/mardi 28 mai, 13:45-14:00]

Shakhawat Hossain (University of Winnipeg) , **Le An Lac** (University of Winnipeg)

Efficient Estimation in Partially Linear Single-Index Models for Binary Longitudinal Data

Estimation efficace dans des modèles à un seul indice partiellement linéaires pour des données longitudinales binaires

We consider the estimation strategies of partially linear single-index models (SIM) for binary longitudinal data. A semiparametric method based on a combination of a non-parametric approach and generalized estimation equations (GEE) is introduced to estimate the two parameter vectors as well as the unknown link function. We develop the pretest and shrinkage estimation methods for SIM when some regression parameters are subject to restrictions. We fit two models: one includes all the covariates and the other restricts the regression parameters based on the auxiliary information. The unrestricted and restricted estimators then are combined optimally to get the pretest and shrinkage estimators. Asymptotic properties of these estimators including biases and risks will be discussed. A simulation study is conducted to assess the performance of the estimators

Nous considérons les stratégies d'estimation des modèles à un seul indice partiellement linéaires pour les données longitudinales binaires. On présente une méthode semi-paramétrique basée sur la combinaison d'une approche non paramétrique et d'équations d'estimation généralisées pour estimer les deux vecteurs de paramètres, ainsi que la fonction de liaison inconnue. Nous mettons au point les méthodes d'estimation des prétests et du rétrécissement pour les modèles à un seul indice lorsque certains paramètres de régression font l'objet de restrictions. Nous ajustons deux modèles : l'un inclut toutes les covariables et l'autre restreint les paramètres de régression en fonction de l'information auxiliaire. Les estimateurs non restreints et restreints sont ensuite combinés de façon optimale pour obtenir les estimateurs de prétest et de rétrécissement. Nous examinerons les propriétés asymptotiques de ces estimateurs, ainsi que les biais et les risques. Nous réalisons une étude de simulation pour évaluer l'efficacité

with respect to the unrestricted estimator and an empirical application will be used to illustrate the usefulness of our methodology.

des estimateurs par rapport à celle de l'estimateur non restreint, et nous utilisons une application empirique pour illustrer l'utilité de notre méthodologie.

[Tuesday May 28/mardi 28 mai, 14:00-14:15]

Victoire Michal (Université de Montréal) , **David Haziza** (Université de Montréal) , **Sixia Chen** (University of Oklahoma)

Efficient Multi-Robust Estimation in the Presence of Influential Units

Estimation efficace multi-robuste en présence d'unités influentes

Item nonresponse is a common issue in surveys. To reduce the bias of unadjusted estimators, it is common practice to impute the missing values, leading to the creation of a completed data file. In practice, one must also face the problem of influential units in the sample, which make the commonly used estimators of population totals/means very unstable. To reduce the impact of influential units, we develop a robust version of multiply robust estimators using the conditional bias of a unit. The latter is a measure of influence of a unit that accounts for both sampling and nonresponse. We will present the results of a simulation study to show the benefits of the proposed method in terms of bias and efficiency.

La non-réponse d'un item est un problème courant dans les sondages. Pour réduire le biais des estimateurs non-ajustés, la pratique courante est d'imputer les valeurs manquantes donnant ainsi lieu à la création d'un fichier de données complet. En pratique, on doit aussi faire face au problème d'unités influentes dans l'échantillon, ce qui rend les estimateurs couramment utilisés de totaux et de moyennes de la population très instables. Pour diminuer l'impact des unités influentes, nous élaborons une version robuste des estimateurs multiplement robustes en utilisant le biais conditionnel d'une unité. Ce dernier est une mesure de l'influence d'une unité qui représente à la fois l'échantillonnage et la non-réponse. Nous présenterons les résultats d'une étude de simulation pour démontrer les avantages de la méthode proposée en termes de biais et d'efficacité.

[Tuesday May 28/mardi 28 mai, 14:15-14:30]

Julien Miron (Université de Genève) , **Benjamin Poilane** (Université de Genève) , **Eva Cantoni** (Université de Genève)

Robust Estimation of Polytomous Logistic Regression Models

Estimation robuste de modèles de régression logistique polytomique

Polytomous logistic regression is a useful model when the response variable can take one of C possible values. The relationship between the expectation of the response variable and the covariates is then setup with a (logistic) link function. The model is generally fitted by maximum likelihood, which has well-known properties and is easily implemented. However, one drawback of this method is its non robustness to contamination in the response variable and in the design matrix. A robust estimator built on the generalized linear models proposal of Cantoni and Ronchetti (2001) as well as an optimal B-robust estimator are proposed. The first approach applies a bounding function to the residuals, whereas the second one acts on the score function to downweight outlying observations. We compare our proposals with alternatives, namely the density power divergence estimator (Castilla et al., 2018) and a generalized method of weighted moments (Wang, 2014).

La régression logistique polytomique est un modèle utile quand la variable de réponse peut prendre une parmi valeurs possibles. La relation entre ce qui est attendu de la variable de réponse et les covariables est alors établie avec une fonction lien (logistique). Le modèle est généralement ajusté par la méthode du maximum de vraisemblance, dont les propriétés et la facilité de mise en œuvre sont bien connues. Par contre, un inconvénient de cette méthode est son absence de robustesse contre la contamination de la variable de réponse et aussi de la matrice de design. Un estimateur robuste établi d'après la proposition de Cantoni et Ronchetti (2001) de modèles linéaires généralisés de même qu'un estimateur robuste B optimal sont proposés. La première approche applique une fonction limitante aux résidus, tandis que la seconde agit sur la fonction de caractérisation pour sous-pondérer les données aberrantes. Nous comparons nos propositions à d'autres, notamment à l'estimateur de divergence de puissance de densité (Castilla et coll. (2018)) et à la méthode généralisée des moments pondérés (Wang (2014)).

[Tuesday May 28/mardi 28 mai, 14:30-14:45]

Luc Villandré (HEC Montreal) , **Patrick Brown** (University of Toronto) , **Thierry Duchesne** (Université Laval) , **Nancy Reid** (University of Toronto) , **Jean-François Plante** (HEC Montréal)

Advances in Estimation Methods
Progrès en matière de méthodes d'estimation

Integrated Nested Laplace Approximation (INLA) Estimation for a Spatio-Temporal Regression Model Applicable to Large Datasets

Estimation de l'approximation de Laplace imbriquée (ALI) pour un modèle de régression spatiotemporel applicable à de grands jeux de données

Massive spatio-temporal datasets are a computational challenge for conventional regression models. The Multi-Resolution Approximation (MRA) is a Bayesian model designed to address that issue. The MRA has not yet been extended to spatio-temporal data, and in practice, cannot easily estimate the hyperparameters' posteriors. We therefore propose the INLA-MRA method, which uses the Integrated Nested Laplace Approximation (INLA) to estimate posteriors. We test the method on simulated data generated under the assumed model, and then test it in the presence of misspecification. We fit the model to a real dataset comprising temperatures recorded on five consecutive days in northern Quebec. Simulations revealed that INLA-MRA produces posterior estimates centered close to the true parameter values. Model misspecification lead to larger variances in posteriors. In the real data analyses, INLA-MRA was especially helpful for producing predictions in regions where blocks of data were missing.

Les ensembles de mégadonnées spatiotemporelles posent un défi informatique de taille pour les modèles de régression conventionnels. L'approximation multirésolution (AMR) est un modèle bayésien conçu pour aborder ce problème. L'AMR n'a pas encore été adaptée pour les données spatiotemporelles. D'ailleurs, en pratique, elle ne peut pas aisément estimer les données a posteriori des hyperparamètres. Par conséquent, nous proposons la méthode ALI-AMR, qui se sert de l'approximation de Laplace imbriquée (ALI) pour estimer les données a posteriori. Nous testons la méthode à partir de données simulées et générées en fonction du modèle, puis la testons en présence d'erreurs de spécification. Nous adoptons le modèle à un jeu de données réelles composé de températures enregistrées pendant cinq jours consécutifs dans le nord du Québec. Les simulations ont révélé que la ALI-AMR produit des estimations a posteriori centrées près des valeurs réelles du paramètre. Les erreurs de spécification du modèle mènent à de plus grandes variances dans les données a posteriori. Dans le cadre d'analyses de données réelles, la ALI-AMR a été tout particulièrement utile pour produire des prédictions dans les régions où il manquait des blocs de données.

[Tuesday May 28/mardi 28 mai, 14:45-15:00]

Xiufang Liu (University of Regina) , **Dianliang Deng** (University of Regina) , **Dehui Wang** (Jilin University)

Estimating the Quantile Function for History Process with Time-Dependent Covariates and Censoring Mechanism

Estimation de la fonction quantile pour processus historique avec covariables à dépendance chronologique et mécanisme de censure

Most of the existing literatures on longitudinal data analysis have focused on modeling the conditional mean. Recently, analyses estimating the cumulative mean function (CMF) for history process has generated significant interest as an important component in health treatment evaluation. However, there is little literature studying the estimation of the cumulative quantile function (CQF) for history process. In this paper, a novel approach to estimating the CQF for history process with time-dependent covariates and right censored time-to-event variable is developed using the inverse probability weighting method. The consistency of the proposed estimator is derived. The performance of the CQF is investigated via extensive simulations and we illustrate the application by analyzing a real data set from a multicenter automatic defibrillator implantation trial.

La littérature actuelle relative à l'analyse de données longitudinales porte essentiellement sur la modélisation de la moyenne conditionnelle. Récemment, l'analyse de l'estimation de la fonction moyenne cumulative (FMC) des processus historiques a suscité un intérêt considérable, ces derniers étant une composante importante de l'évaluation des traitements de santé. Cependant, il n'existe que peu d'études sur l'estimation de la fonction quantile cumulative (FQC) pour les processus historiques. Dans cette présentation, nous développons une nouvelle approche pour estimer la fonction quantile cumulative (FQC) pour les processus historiques avec covariables à dépendance chronologique et variable de temps d'événement censurée à droite, en nous fondant sur la méthode de pondération par probabilité inverse. Nous dérivons la convergence de l'estimateur proposé. Nous étudions la performance de la fonction quantile cumulative (FQC) via des simulations approfondies et nous illustrons l'application en analysant un jeu de données réelles tirées d'un essai multicentrique d'implantation de défibrillateurs automatiques.

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

Chair/Président: Susie Fortier

Organizer/Responsable: Susie Fortier

Room/Salle: 201 (ENA)

Abstract/Résumé

[Tuesday May 28/mardi 28 mai, 15:30-16:35]

Jack Gambino (Statistics Canada)

The Evolving Role of Non-Survey Data in Official Statistics

L'évolution du rôle des données non issues d'enquêtes dans les statistiques officielles

For decades, administrative data such as birth records and tax data have played a major role in official statistics. In recent years, efforts to exploit other administrative and non-survey data have increased significantly. Some statistical agencies have adopted an admin-first policy. To meet an information need, the agency first looks at existing data sources, turning to the survey option as a last resort. Perhaps this change would have occurred naturally over time, in part due to cost pressures. But it became essential due to the gradual decline in response rates. For some surveys, the decline has been substantial, leading to questions about the validity of their estimates. In this presentation, we look at the role of survey and non-survey data and their integration in a statistical system. The focus is on a suitable infrastructure and how that infrastructure can be used. Finally we consider some risks involved in combining survey and non-survey data for official statistics.

Pendant des décennies, les données administratives telles les registres des naissances et données fiscales ont joué un rôle prépondérant dans les statistiques officielles. Ces dernières années pourtant, on a redoublé d'efforts pour exploiter d'autres données administratives et des données non issues d'enquête. Certains organismes de statistique ont adopté une politique de « priorité à l'administratif » : pour répondre à un besoin d'information, ils exploitent d'abord les sources de données existantes avant de se tourner vers l'option enquête en dernier ressort. Peut-être cette évolution se serait-elle faite naturellement avec le temps, en raison notamment de pressions économiques. Mais elle est aujourd'hui devenue essentielle, en raison de la diminution progressive des taux de réponse. Dans le cas de certains sondages, ce déclin a été considérable, donnant lieu à des questions quant à la validité des estimations. Dans cette présentation, nous examinons le rôle des données obtenues par sondage ou d'autres moyens et leur intégration dans un système statistique. L'accent est mis sur l'infrastructure nécessaire et comment exploiter cette dernière. Enfin, nous examinons les risques que présente l'inclusion de telles données dans les statistiques officielles.

Analytics in Sports
Analyse sportive

Chair/Président: Shirley Mills

Organizer/Responsable: Shirley Mills

Room/Salle: 116 (ICT)

Abstract/Résumé

[Tuesday May 28/mardi 28 mai, 15:30-16:00]

Tim B. Swartz (Simon Fraser University) , **Rajitha Silva** (University of Sri Jayewardenepura) , **Lucas Wu** (Simon Fraser University) , **Joan Hu** (Simon Fraser University)

Soccer Insights

Regards sur le soccer

Soccer (the beautiful game) has caught my fancy. In this talk, I discuss several aspects of strategy that are informed by data. In particular I will discuss the substitution rule, the dependence of goal scoring on situation and the effectiveness of crossing where the last topic is addressed via player tracking data. Some of the statistical buzzwords that are relevant to the presentation are temporal smoothing, prior elicitation, WinBUGS, autoregressive covariance matrices and nonparametrics.

Le soccer, ce magnifique sport, a attiré mon attention. Dans cet exposé, j'aborde plusieurs aspects de la stratégie qui s'appuient sur des données. Je discuterai notamment de la règle de substitution, de la dépendance des tirs au but par rapport au contexte et de l'efficacité de la traversée, où le dernier sujet est abordé au moyen des données de suivi des joueurs. Certains des mots du jargon statistique qui sont pertinents pour la présentation sont : lissage temporel, élicitation a priori, WinBUGS, matrices de covariance autorégressives et méthodes non-paramétriques.

[Tuesday May 28/mardi 28 mai, 16:00-16:30]

Michael E. Schuckers (St. Lawrence University)

Statistical Analysis of the National Hockey League Entry Draft

Analyse statistique du repêchage dans la Ligue nationale de hockey

The National Hockey League (NHL) Entry Draft, which takes play in June of each year, is the mechanism by which the NHL allocates new players to its teams. In this talk we will discuss some statistical analyses of historical NHL Entry Drafts. These analyses include methods for valuing each individual draft selection as well as some methods for predicting future performance of players. One important aspect is the quality of rankings that the NHL provides via their Central Scouting Service (CSS) ranking of draft eligible players, which is published each April. Of particular note will be comparisons between CSS rankings, a supervised statistical model and the actual order of the NHL Entry Draft. Finally, we will propose some improvements to the models in the public sphere.

Tenu en juin tous les ans, le repêchage dans la Ligue nationale de hockey (LNH) est le mécanisme permettant à la ligue d'adjoindre de nouveaux joueurs à ses équipes. Notre propos porte sur certaines analyses statistiques de repêchages historiques dans la LNH. Ces analyses comportent des méthodes d'évaluation de chaque choix au repêchage et quelques méthodes pour prévoir le rendement futur des joueurs. Un facteur important du processus : la qualité des classements que fournit la LNH à l'aide de la sélection des joueurs admissibles au repêchage que publie annuellement en avril son Bureau central de dépistage (BCD). À noter en particulier, les comparaisons entre les sélections du BCD, un modèle statistique supervisé et le classement réel au repêchage de la LNH. Nous proposons enfin quelques améliorations aux modèles utilisés dans la sphère publique.

[Tuesday May 28/mardi 28 mai, 16:30-17:00]

Nathan Sandholtz (Simon Fraser University) , **Jacob Mortensen** (Simon Fraser University) , **Luke Bornn** (Sacramento Kings; Simon Fraser University)

Measuring Spatial Allocative Efficiency in Professional Basketball

Mesure de l'allocation spatiale optimale des ressources dans le domaine du basketball professionnel

In professional basketball, allocative efficiency is fundamentally a spatial problem—the distribution of shot attempts within a lineup is highly dependent on court location. Despite the importance of spatial context, there are very few analyses that have explicitly accounted for this critical factor. The main idea behind our approach is to compare a player's field goal percentage (FG%) to his field goal attempt (FGA) rate in context of his four teammates on the court for a given lineup. To this end, we build Bayesian hierarchical models to estimate player FG% and FGA rates at every location in the offensive half-court using publicly available data from the National Basketball Association (NBA). Then, by pairing a player's lineup-specific FGA rankings with his corresponding FG% rankings, we can detect areas where the lineup exhibits inefficient allocation of shots, quantify and visualize the points that are consequently lost, and identify which players are responsible.

Dans le domaine du basketball professionnel, l'allocation optimale des ressources est fondamentalement un problème spatial — la distribution des tentatives de tir dans un alignement dépend fortement de l'emplacement sur le terrain. Malgré l'importance du contexte spatial, très peu d'analyses ont explicitement pris en compte ce facteur déterminant. L'idée principale derrière notre approche vise à comparer le pourcentage de tirs d'un joueur avec son taux de tentatives de tirs dans le contexte de la position de ses quatre coéquipiers sur le terrain pour un alignement donné. À cette fin, nous créons des modèles hiérarchiques bayésiens pour estimer les taux de pourcentage de paniers et de tentatives de paniers des joueurs à chaque endroit de la moitié du terrain de l'attaquant en utilisant des données publiques existantes de la National Basketball Association (NBA). Ensuite, en appariant les classements tentatives de tirs spécifiques à un alignement de joueurs avec ses classements de pourcentage de paniers sur le terrain correspondants, nous pouvons détecter les zones où l'alignement présente une allocation inefficace des tirs, quantifier et visualiser les points qui sont par conséquent perdus, et déterminer les joueurs responsables de ces points perdus.

Tenure and Promotion: Insightful Tips from the Applicants and Reviewers
Permanence et promotion : conseils de candidats et d'évaluateurs

Chair/Président: Hua Shen

Organizer/Responsable: Hua Shen

Room/Salle: 142 (AD)

Abstract/Résumé

[Tuesday May 28/mardi 28 mai, 15:30-17:00]

Hua Shen (University of Calgary) , **Richard Lockhart** (Simon Fraser University) , , **Xikui Wang** (University of Manitoba) , , **Wendy Lou** (University of Toronto) , , **Louis-Paul Rivest** (Université Laval) , , **Hugh Chipman** (Acadia University) , , **Linglong Kong** (University of Alberta)

Tenure and Promotion: Insightful Tips from the Applicants and Reviewers

Titularisation et promotion : des conseils judicieux de la part de candidats et d'évaluateurs

The New Investigator Committee exists to provide junior academics with the opportunity to gain advices from insightful mentors and build for continuous success. One of the milestones and challenges faced by new investigators is to be awarded tenure and granted promotion in their first few years in academia. The committee has prepared a panel discussion on tenure and promotion consisting of invited speakers at different stages of their careers and having different experiences and perspectives. The intent of this session is to provide new investigators with the opportunity to learn about the successes of individuals who have tenure and receive powerful tips to prepare for thorough assessment processes. The panelists who either have extensive experiences or served in tenure review positions either internally or externally or recently successfully went through the process have been invited to share their great insights.

Le comité des nouveaux chercheurs existe pour offrir l'occasion aux jeunes universitaires d'obtenir des conseils de la part de mentors et de se préparer en permanence pour atteindre la réussite. Un des jalons et défis que les nouveaux chercheurs rencontrent est de devenir titulaire et de recevoir une promotion durant leurs premières années dans le monde universitaire. Le comité a préparé un atelier de discussion portant sur la titularisation et les promotions comprenant des invités participants situés à différentes étapes dans leur carrière et ayant différentes expériences et perspectives. L'objectif de cette séance est d'offrir aux nouveaux chercheurs l'occasion d'apprendre à partir de la réussite de personnes titulaires et de bénéficier des conseils judicieux pour se préparer à tous les procédés d'évaluation. Des invités qui ont beaucoup d'expérience, qui ont été titulaires à l'interne ou à l'externe ou qui ont récemment réussi à terminer la procédure seront présents pour partager leurs opinions.

New perspectives and challenges in analysis of linked genomic and phenomic data
Nouvelles perspectives et nouveaux défis en analyse de données génomiques et phénomiques couplées

Chair/Président: Jinko Graham

Organizer/Responsable: Jinko Graham

Room/Salle: 102 (ICT)

Abstract/Résumé

[Tuesday May 28/mardi 28 mai, 15:30-16:00]

Quan Long (University of Calgary) , **Shengjie Lu** (University of Calgary) , **Chen Cao** (University of Calgary) , **Zhi Xiong** (Shantou University) , **Xuewen Lu** (University of Calgary)

Less-Is-More and More-Is-Less in Integrating Multi-Scale Omics with Polygenic Phenotype Predictors

« Moins donne plus » et « plus donne moins » dans l'intégration de l'omique multiéchelle dans les prédicteurs de phénotype polygéniques

A goal of genomics is to develop a genome-based phenotype predictor. Empowered by the advancement of high-throughput technologies, in addition to genomes, researchers can assess transcriptomes and methylomes. How to statistically integrate them in a phenotype predictor is tricky. This is because, due to overfitting, more data may cause an increased model complexity that offsets the benefit of the extra information. We develop a novel approach that leverages omics to reduce, instead of increase, the model complexity. In our method, the role of omics is to provide priori knowledge that helps group genomic variants effectively, leading to fewer coefficients in the predictor. Testing it on benchmark datasets in plants and humans demonstrates that it substantially increases the power of prediction. This research leads to a new path towards integration of multi-scale data with polygenic predictors along the line of the "less-is-more" principle in the circumstance of "more-is-less".

L'un des objectifs de la génomique est de développer un prédicteur de phénotype basé sur le génome. Grâce à l'avancée des technologies à haut débit, outre le génome, les chercheurs peuvent également évaluer les transcriptomes et les méthylomes. Mais il est délicat de les intégrer statistiquement dans un prédicteur de phénotype. En effet, à cause du surajustement, un surplus de données risque de rendre le modèle plus complexe, annulant l'avantage des informations supplémentaires. Nous développons une nouvelle approche qui exploite l'omique pour réduire, plutôt qu'augmenter, la complexité du modèle. Dans notre méthode, le rôle de l'omique est de fournir des connaissances a priori qui aident à regrouper les variantes génomiques efficacement, réduisant ainsi le nombre de coefficients du prédicteur. Nos tests sur des jeux de données de référence sur des plantes et des humains montrent qu'elle augmente considérablement la puissance de prédiction. Ces recherches nous conduisent vers une nouvelle façon d'intégrer des données multiéchelle avec des prédicteurs polygéniques selon le principe du « moins donne plus » dans une situation de « plus donne moins ».

[Tuesday May 28/mardi 28 mai, 16:00-16:30]

Lloyd T Elliott (Simon Fraser University)

Towards Modern Machine Learning for Genome-Wide Association

Vers un apprentissage machine moderne pour l'association pangénomique

In collaboration with the University of Oxford Department of Statistics and the Wellcome Centre for Integrative Neuroimaging, a genome-wide association study was conducted on 3,144 phenotypes derived from brain magnetic resonance imaging (MRI) in 8,428 participants of the UK Biobank consortium. Classical analyses of genetic effects can provide heritability, genetic correlation and enrichment analyses. We discuss results from classical analyses, and provide directions for

En collaboration avec le département de statistiques et le Wellcome Centre for Integrative Neuroimaging de l'University of Oxford, une étude sur l'association pangénomique a été menée sur 3 144 phénotypes issus de l'imagerie par résonance magnétique (IRM) du cerveau de 8 428 participants du UK Biobank consortium. Les études classiques sur les effets génétiques peuvent fournir des analyses d'héritabilité ainsi que de corrélation et d'enrichissement génétiques. En plus de traiter des résultats d'analyses classiques et de proposer un mode d'emploi pour intégrer l'étude

bringing genome-wide association studies into the predictive frameworks used by Bayesian statistics and Deep Learning, and outline new frontiers in digital medicine.

des associations pangénomiques aux cadres prédictifs utilisés en statistique et apprentissage profond bayésiens, nous définissons les nouvelles frontières de la médecine numérique.

[Tuesday May 28/mardi 28 mai, 16:30-17:00]

Kun Liang (University of Waterloo) , **Yu Gao** (University of Waterloo)

Controlling the False Discovery Rate of GWAS

Contrôle du taux de fausses découvertes dans les études d'association pangénomiques (GWAS)

For complex traits or diseases, the reported genomic loci only explain small proportions of the trait variations in genome-wide association studies (GWAS). The use of the traditional familywise error rate can be overly stringent; we aim to control the false discovery rate. Due to the extensive linkage disequilibrium among SNPs, defining the true and false positives at the SNP level is difficult. We propose to define them on the level of blocks, which are formed by grouping correlated SNPs. We further develop a novel method to control the false discovery rate by conditioning on a set of tentative causal SNPs. Simulation studies and a real application will be presented to illustrate the use of our method.

En ce qui concerne les maladies ou traits complexes, les locus génomiques rapportés n'expliquent qu'en petites parties les variations des traits dans les études d'association pangénomiques. L'emploi du FWER (family-wise error rate) peut être trop rigoureux ; le but est de contrôler le taux de fausses découvertes. En raison du considérable déséquilibre de liaison parmi les SNP, il est difficile de distinguer les vrais positifs des faux positifs au niveau du SNP. Nous proposons de les définir au niveau des blocs, qui sont formés en regroupant les SNP corrélés. Nous concevons en plus une nouvelle méthode pour contrôler le taux de fausses découvertes au moyen d'un conditionnement d'un ensemble de SNP causal provisoire. Des études de simulation et une application réelle seront présentées pour illustrer le fonctionnement de notre méthode.

Advances in model-based clustering of complex data
Progrès des méthodes de groupage par modèle pour données complexes

Chair/Président: Linglong Kong

Organizer/Responsable: Bei Jiang

Room/Salle: 146 (SB)

Abstract/Résumé

[Tuesday May 28/mardi 28 mai, 15:30-16:00]

Thomas Brendan Murphy (University College Dublin) , **Michael Fop** (University College Dublin) , **Luca Scrucca** (Università degli Studi di Perugia)

Model-Based Clustering with Sparse Covariance Matrices

Regroupement selon un modèle avec des matrices de covariance éparses

Finite Gaussian mixture models are widely used for model-based clustering of continuous data. Nevertheless, since the number of model parameters scales quadratically with the number of variables, these models can be easily over-parameterized. For this reason, parsimonious models have been developed via covariance matrix decompositions or assuming local independence. We introduce mixtures of Gaussian covariance graph models for model-based clustering with sparse covariance matrices. A penalized likelihood approach is employed for estimation and a general penalty term on the graph configurations can be used to induce different levels of sparsity and incorporate prior knowledge. With this approach, sparse component covariance matrices are directly obtained. The framework results in a parsimonious model-based clustering of the data via a flexible model for the within-group joint distribution of the variables.

Les modèles de mélange gaussien fini sont largement utilisés pour le regroupement de données continues selon un modèle. Cependant, comme le nombre de paramètres du modèle est quadratique par rapport au nombre de variables, ces modèles peuvent facilement être surparamétrés. C'est pourquoi des modèles parcimonieux ont été créés par décomposition de matrices de covariance ou avec l'hypothèse d'une indépendance locale. Nous présentons des mélanges de modèles graphiques gaussiens de covariance pour le regroupement selon un modèle avec des matrices de covariance éparses. On utilise une approche de vraisemblance pénalisée pour l'estimation ; un terme de pénalité général sur les configurations graphiques peut être utilisé pour induire différents niveaux de faible densité et intégrer les connaissances préalables. Cette approche permet d'obtenir directement des matrices de covariance éparses de composantes. Le cadre donne lieu à un regroupement des données parcimonieux au moyen d'un modèle souple pour la distribution conjointe des variables à l'intérieur d'un groupe.

[Tuesday May 28/mardi 28 mai, 16:00-16:30]

Bei Jiang (University of Alberta) , **Adrian E. Raftery** (University of Washington) , **Russell J. Steele** (McGill University) , **Naisyin Wang** (University of Michigan)

Balancing Inferential Integrity and Disclosure Risk: A Mixture Modeling Approach

Équilibrer l'intégrité de l'inférence et le risque de divulgation : une approche de modélisation par mélange

In the context of survey sampling, Rubin (1993) proposed to release multiply imputed synthetic datasets with the target sensitive values replaced by values drawn from posterior predictive distributions under proper imputation models. However, information loss due to incorrect model specification can weaken or invalidate the inference obtained from synthetic data. We discuss a new masking framework through data augmentation as a potential remedy. The new framework can always guarantee valid inferences using synthetic datasets, and al-

Dans le contexte d'échantillonnage d'enquêtes, Rubin (1993) a proposé de publier des jeux de données imputés dont les valeurs sensibles à la cible sont remplacées par des valeurs tirées de lois prédictives a posteriori selon des modèles d'imputation adéquats. Toutefois, la perte d'information causée par une erreur de spécification d'un modèle peut affaiblir ou invalider l'inférence obtenue à partir des données synthétiques. En guise de solution possible, nous examinons un nouveau cadre de masquage par l'entremise de l'augmentation des données. Ce nouveau cadre garantit toujours la validité des inférences au moyen de

lows data users to obtain their desired data utility while satisfying disclosure requirements. This framework can be extended through mixture modelling and combined with other existing methods to accommodate different levels of disclosure protection. We demonstrate through simulations and an illustrative example that the new framework outperforms the classical multiple imputation approach to preserving data utility while providing good disclosure protection.

données synthétiques et permet aux utilisateurs d'obtenir l'utilité des données souhaitée tout en respectant les exigences en matière de divulgation. Il peut aussi être adapté grâce à la modélisation par mélange et combiné à d'autres méthodes actuelles pour fonctionner selon différents niveaux de protection de la divulgation. Nous démontrons au moyen de simulations et d'un exemple explicatif que le nouveau cadre surpasse l'approche par imputations classique dans la préservation de l'utilité des données, tout en offrant une bonne protection de la divulgation.

[Tuesday May 28/mardi 28 mai, 16:30-17:00]

Gongjun Xu (University of Michigan) , **Yuqi Gu** (University of Michigan)

Learning Attribute Patterns in High-Dimensional Structured Latent Attribute Models

Apprentissage des structures d'attributs dans les modèles à attributs latents structurés en haute dimension

Structured latent attribute models (SLAMs) are a special family of discrete latent variable models widely used in the social and biological sciences. This work considers the problem of learning significant latent attribute patterns from a SLAM with potentially high-dimensional patterns. We address the theoretical identifiability issue, propose a penalized likelihood method for the selection of the attribute patterns, and further establish the selection consistency in the overfitted SLAM with diverging number of latent mixture components. The good performance of the proposed methodology is illustrated by simulation studies and two real datasets in educational assessment.

Les modèles à attributs latents structurés (MALS) sont une famille spéciale de modèles à variables latents discrètes communément utilisés en sciences sociales et biologiques. Cette présentation explore le problème de l'apprentissage de structures d'attributs latents significatifs à partir d'un MALS à structures potentiellement de haute dimension. Nous abordons le problème de l'identifiabilité théorique, proposons une méthode de vraisemblance pénalisée pour la sélection des structures d'attributs, puis établissons la convergence de la sélection dans le MALS surajusté avec un nombre divergent de composantes de mélange latentes. Nous illustrons la bonne performance de la méthode proposée par des études de simulation et deux jeux de données réelles tirées de l'évaluation pédagogique.

CANSSI Postdoctoral Showcase
Présentations des stagiaires postdoctoraux de l'INCASS

Chair/Président: John Braun

Organizer/Responsable: John Braun

Room/Salle: 109 (SS)

Abstract/Résumé

[Tuesday May 28/mardi 28 mai, 15:30-16:00]

Whitney K. Huang (University of Victoria) , **Francis Zwiers** (University of Victoria) , **Adam Monahan** (University of Victoria)

Modeling Compound Wind and Precipitation Extremes Using a Large Climate Model Ensemble

Modélisation de vents et de précipitations extrêmes en utilisant un ensemble de grands modèles climatiques

The concurrence of precipitation and wind extremes can have severe impacts on buildings and infrastructure. Thus, there is a pressing need to reliably estimate the magnitude of such compound extreme events and their potential changes in a changing climate. We therefore develop statistical methods for modeling such compound extreme events (i.e., simultaneous wind and precipitation extremes). Specifically, semi-parametric methods that jointly estimate the bulk and tails of bivariate probability densities are developed to assess the risk of the compound precipitation and wind extremes. A large (35 members, 26 billion observations for each ensemble) moderate (50km) resolution North American regional climate model ensemble of the period from 1951 to 2100 under a climate change scenario is used to both assess the performance of the proposed methods and to facilitate the non-stationary modeling of compound extremes.

La conjonction de précipitations et de vents extrêmes peut avoir de graves conséquences sur les bâtiments et les infrastructures. Il y a donc un besoin pressant de pouvoir estimer efficacement la magnitude de tels évènements extrêmes composés ainsi que leurs changements potentiels dans un climat changeant. Nous avons donc développé des méthodes statistiques pour modéliser de tels événements (c'est-à-dire des précipitations et des vents extrêmes simultanés). Plus particulièrement, des méthodes semi-paramétriques qui estiment conjointement l'ensemble et les ailes des densités de probabilités sont développées pour évaluer le risque de précipitations et de vents extrêmes composés. Un large (35 membres, 26 milliards d'observations pour chaque ensemble) ensemble de modèles à résolution modérée (50 km) sur le climat régional nord américain de la période allant de 1951 à 2100 dans un scénario de changement climatique est utilisé pour évaluer la performance des méthodes proposées et pour faciliter la modélisation non-stationnaire des extrêmes composés.

[Tuesday May 28/mardi 28 mai, 16:00-16:30]

David Soave (Ontario Institute for Cancer Research) , **Jerry Lawless** (University of Waterloo)

Regularized Regression Methods for Two Phase Studies

Méthodes de régression régularisée pour les études en deux phases

Large prospective cohorts like the Canadian Partnership for Tomorrow Project (CPTP) follow participants longitudinally and capture incident cases of disease. In two-phase studies, researchers select a subset of the complete cohort based on observed outcomes and covariates and measure additional, possibly expensive, variables. In the CPTP, blood samples are collected and stored when participants enroll. During follow-up a small fraction of the cohort will present with rare diseases and researchers often select a matched case-control sample and obtain biological information from their blood sam-

Toute grande cohorte prospective comme celle du Projet de partenariat canadien Espoir pour demain (PPCED) fait un suivi longitudinal des participants et procède à la capture des cas nouveaux de maladie. Dans les études en deux phases, les chercheurs choisissent un sous-ensemble de toute la cohorte en fonction des résultats et covariables observés et mesurent d'autres variables possiblement coûteuses. Au moment de l'inscription des participants au PPCED, des échantillons de leur sang sont recueillis et stockés. Pendant le suivi, une petite partie de la cohorte souffrira de maladies rares et les chercheurs choisissent souvent un échantillonnage correspondant de cas témoins. Ils obtiennent des

ples, seeking to discover biomarkers that predict disease. Such biological information is commonly represented by a large dataset with many variables including gene expression values or DNA genotypes. This presentation will describe variable selection methods based on regularized regression for two-phase study data. Illustrations will be based on predictive modeling of acute myeloid leukemia risk.

données biologiques à partir des échantillons sanguins avec pour but de découvrir des biomarqueurs susceptibles de prévoir des maladies. Ces données biologiques sont couramment représentées par un grand ensemble de données et de multiples variables, dont des valeurs d'expression génétique ou des génotypes d'ADN. Cette présentation offre une description des méthodes de sélection de variables fondées sur la régression régularisée pour les données d'études en deux phases. Une modélisation prédictive du risque de leucémie myéloïde aiguë illustre nos propos.

[Tuesday May 28/mardi 28 mai, 16:30-17:00]

Luc Villandré (McGill University), **Aurélie Labbe** (HEC Montréal), **Ilinca-Ruxandra Ibanescu** (Jewish General Hospital) **Isabelle Hardy** (Centre Hospitalier de l'Université de Montréal), **Bluma Brenner** (Jewish General Hospital), **Michel Roger** (Centre Hospitalier de l'Université de Montréal), **David Stephens** (McGill University)

HIV transmission cluster inference using Bayesian phylogenetics

Inférence des grappes de transmission du VIH par l'intermédiaire de la phylogénétique bayésienne

Phylogenetics is the field concerned with the inference of the ancestral links between organisms. Phylogenetic models represent the ancestry of genomic sequences with a tree structure known as a phylogeny. Studies have used phylogenetics to investigate HIV transmission among men who have sex with men (MSMs) in Quebec, revealing many transmission clusters. Conventional phylogenetic clustering approaches rely on arbitrary numerical cutpoints, and are therefore hard to tune properly. The current study proposes *DM-PhyClus*, a Bayesian phylogenetic clustering algorithm. We apply it to a sample of 3,704 real HIV-1 sequences collected among MSMs in Quebec, and measure the expansion of transmission chains attributable to known clusters. Simulations reveal that DM-PhyClus can outperform conventional methods in terms of mean cluster recovery. Cluster estimates obtained from the real data overlap moderately with those from conventional methods. DM-PhyClus facilitates transmission cluster detection by eliminating the need for arbitrary cutpoints.

La *phylogénétique* est un domaine d'expertise portant sur l'inférence et les liens ancestraux entre organismes. Les modèles phylogénétiques représentent l'ascendance des séquences génomiques au moyen d'une structure arborescente que l'on appelle phylogénie. Des études ont employé la phylogénétique pour examiner la transmission du VIH parmi des hommes ayant des relations homosexuelles (HRH) au Québec, révélant plusieurs groupements de transmission. Les approches conventionnelles de regroupements phylogénétiques reposent sur des critères numériques arbitraires, ce qui rend leur réglage d'autant plus difficile. L'étude actuelle présente *DM-PhyClus* : un algorithme bayésien de regroupement phylogénétique. Nous l'appliquons à un échantillon de 3704 séquences véritables de VIH-1 recolté auprès de HRHs au Québec, et mesurons l'étendue des chaînes de transmission attribuables aux groupes connus. Les simulations démontrent que la performance de DM-PhysClus est supérieure aux méthodes conventionnelles en termes de la moyenne du taux de récupération des groupes. Les estimations de regroupement recueillies à partir des données réelles chevauchent modérément celles obtenues à partir de méthodes conventionnelles. DM-PhyClus facilite la détection de groupes de transmission en éliminant le besoin de dépendre des critères arbitraires.

New Approaches for Functional and Longitudinal Data
Nouvelles approches des données fonctionnelles et longitudinales

Chair/Président: Cindy Xin Feng

Room/Salle: 201 (ENG)

Abstract/Résumé

[Tuesday May 28/mardi 28 mai, 15:30-15:45]

Janie Coulombe (McGill University) , **Erica Moodie** (McGill University) , **Robert Platt** (McGill University)

Two Weighted Estimators for the Treatment Effect that Account for Covariate-Dependent Monitoring Times and Confounding in Longitudinal Studies

Deux estimateurs pondérés pour l'effet de traitement qui considèrent les temps de mesure informatifs et les effets confondants dans les études longitudinales

Electronic health records (EHRs) are increasingly used to assess the effect of treatments on longitudinal outcomes. However, the analysis of EHRs data is prone to confounding, mediation and missing outcomes which, when not considered, may lead to biased estimates of the treatment effect. Imbalances due to covariate-dependent monitoring patterns may also modify the estimate. For instance, sicker patients tend to be over-represented in EHRs, and their outcomes tend to be worse than healthier patients. Several methods have been proposed to handle covariate-dependent monitoring times when looking at longitudinal and continuous outcomes but most focus only on one or two of the four biasing factors above, yet these commonly occur simultaneously. Building on the work of Buzkova Lumley (2009, Stat Med), we propose and demonstrate the use of two novel weighted estimators for treatment effect that address confounding, mediation, missing outcomes and covariate-dependent monitoring times.

Les dossiers de santé électroniques (DSEs) sont utiles pour évaluer l'effet de traitements sur des mesures longitudinales. Cependant, leur analyse est sujette aux effets confondants, à la médiation et aux mesures manquantes qui peuvent mener à une estimation biaisée de l'effet du traitement. Les déséquilibres causés par les temps de mesure informatifs peuvent aussi modifier l'estimation. Par exemple, les patients très malades ont tendance à être surreprésentés dans les DSEs, et leurs mesures à être plus graves comparées aux patients plus sains. Plusieurs méthodes ont été proposées pour traiter les temps de mesure informatifs, mais elles se concentrent sur l'un ou deux des quatre facteurs mentionnés ci-haut, alors que ces facteurs se produisent souvent simultanément. Basés sur le travail de Buzkova et Lumley (2009, Stat Med), nous proposons et démontrons deux nouveaux estimateurs pondérés pour l'effet de traitement qui tiennent compte des effets confondants, de la médiation, des mesures manquantes et des temps de mesure informatifs.

[Tuesday May 28/mardi 28 mai, 15:45-16:00]

Md Erfanul Hoque (University of Manitoba) , **Elif Acar** (University of Manitoba) , **Mahmoud Torabi** (University of Manitoba)

A D-Vine Copula Model for Unbalanced and Unequally Spaced Longitudinal Data

Un modèle de copule en vigne D pour les données longitudinales déséquilibrées et inégalement espacées

In most longitudinal studies, the number and timing of measurements differ across study subjects. Statistical analysis of such data requires accounting for both the unbalanced study design and the spacing of repeated measurements. In this work, we propose a time-heterogeneous D-vine model that allows for time adjustment in the dependence structure of unbalanced and unequally spaced longitudinal data. The performance of the longitudinal D-vine model is evaluated through simulation studies as well as by a real data application and

Dans la plupart des études longitudinales, le nombre de mesures et leur moment précis varient selon les sujets d'étude. Pour réaliser l'analyse statistique de telles données, il faut tenir compte à la fois du concept d'étude déséquilibré et de l'espacement des mesures répétées. Dans le cadre de ce travail, nous proposons un modèle en vigne D à temps hétérogène qui permet d'ajuster le temps dans la structure de dépendance des données longitudinales déséquilibrées et inégalement espacées. La performance du modèle en vigne D longitudinal est évaluée au moyen d'études en simulation et d'une application sur des données réelles, puis elle

is compared to those of the time-homogeneous models to investigate the impact of ignoring the time interval effects.

est comparée à celles des modèles à temps homogène dans le but d'étudier les conséquences qui surviennent lorsque l'on ignore les effets des intervalles.

[Tuesday May 28/mardi 28 mai, 16:00-16:15]

Adam Kashlak (University of Alberta)

Symmetrization for Exact Nonparametric Functional ANOVA

Symétrisation pour analyse de variance fonctionnelle non paramétrique exacte

Testing for equality of means and covariances among functional data groups has received a lot of attention from both parametric approaches via Gaussian processes and nonparametric ones reliant on permutation tests. In this work, we advance nonparametric testing by devising an exact test via a type of Khintchine inequality, a symmetrisation result for random variables in Banach spaces. This approach combines the computational speed of parametric methods with the distribution free benefits of permutation tests. The methodology is very general and can be extended to other data comparisons across categories.

Les tests d'égalité des moyennes et des covariances entre groupes de données fonctionnelles suscitent beaucoup d'attention, avec des approches paramétriques via processus gaussiens et d'autres non paramétriques fondées sur des tests de permutation. Dans cette présentation, nous préconisons le test non paramétrique et créons un test exact via un type d'inégalité de Khintchine, un résultat de symétrisation pour les variables aléatoires dans les espaces de Banach. Cette approche combine la vitesse de calcul des méthodes paramétriques et les avantages relatifs à la distribution des tests de permutation. La méthodologie est très générale et peut être étendue à la comparaison de données d'autres catégories.

[Tuesday May 28/mardi 28 mai, 16:15-16:30]

Marie-Hélène Descary (Université du Québec à Montréal)

Recovering Covariance from Functional Fragments

Reconstituer la fonction de covariance à partir de fragments de données fonctionnelles

The problem of nonparametric estimation of a covariance function on the unit square is considered given a sample of discretely observed fragments of functional data. When each sample path is only observed on a subinterval of length delta¡1, one has no statistical information on the unknown covariance outside a delta-band around the diagonal. The problem seems unidentifiable without parametric assumptions, but we show that nonparametric estimation is feasible under suitable smoothness and rank conditions on the unknown covariance. This remains true even when observation is discrete, and we give precise deterministic conditions on how fine the observation grid needs to be relative to the rank and fragment length for identifiability to hold. We show that our conditions translate the estimation problem to a low-rank matrix completion problem, and construct a nonparametric estimator in this vein. We illustrate our method by simulation and provide theory to show the validity of the model.

L'estimation non-paramétrique d'une fonction de covariance sur le carré d'unité à partir d'un échantillon de fragments de données fonctionnelles observées de façon discrète est considérée. Chaque courbe de l'échantillon est observée sur un sous-intervalle de longueur delta¡1, il n'y a donc aucune information statistique sur la fonction de covariance à l'extérieur d'une bande delta autour de la diagonale. Le problème semble non-identifiable sans présupposés paramétriques. Néanmoins, nous montrons que l'estimation nonparamétrique est possible en utilisant des conditions sur le rang et la régularité de la fonction de covariance. Ceci reste vrai même lorsque les fragments sont observés en K points ; nous donnons la relation entre K, le rang et la longueur des fragments afin que le problème soit identifiable. Nous montrons que le problème d'estimation se traduit par un problème de complétion de matrice à faible rang, et nous construisons un estimateur non-paramétrique dans la même veine. Nous illustrons notre méthode à l'aide de simulations et développons la théorie afin de justifier la validité de notre modèle.

[Tuesday May 28/mardi 28 mai, 16:30-16:45]

Yijun Xie (University of Waterloo) , **Adam Kolkiewicz** (University of Waterloo) , **Greg Rice** (University of Waterloo)

Functional Normality Test

Test de normalité fonctionnel

Normality tests form important methods in statistics, as a lot of data analysis tools rely on the assumption of a Gaussian distribution. Researchers have developed numerous normality tests for scalar or multivariate data. However, normality tests for functional data have not been thoroughly studied. One of the existing approaches proposes a test based on functional principal components analysis (fPCA), which may fail when non-Gaussian signals are not in the leading principal components. In this work, we propose a normality test based on the projection pursuit method, which overcomes this insufficiency. In particular, we develop efficient algorithms for projection pursuit and optimization in a functional space. In our simulation study, we demonstrate that the new method achieves non-trivial power when compared with the previous method. We apply the new method to fertility rate, stock price, and weather data, and obtain interpretable results.

Les tests de normalité sont des méthodes très importantes en statistique, puisque bon nombre d'outils d'analyse de données reposent sur l'hypothèse de loi gaussienne. Les chercheurs ont mis au point de nombreux tests de normalité pour les données scalaires ou multivariées. Cependant, les tests de normalité pour les données fonctionnelles n'ont pas encore été étudiées de manière approfondie. L'une des approches existantes propose un test fondé sur l'analyse en composantes principales fonctionnelles (ACPf), qui risque d'échouer lorsque des signaux non gaussiens ne figurent pas dans les composantes principales importantes. Dans cette présentation, nous proposons un test de normalité fondé sur la méthode de poursuite de projection, qui permet de surmonter cette insuffisance. En particulier, nous développons des algorithmes efficaces pour la poursuite de projection et l'optimisation dans un espace fonctionnel. Dans notre étude de simulation, nous montrons que cette nouvelle méthode présente une puissance non négligeable par rapport à la méthode précédente. Nous appliquons la méthode proposée à des données de taux de fertilité, des prix d'actions et des données météorologiques et obtenons des résultats interprétables.

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

Chair/Président: Nicholas Mitsakakis

Room/Salle: 105 (SB)

Abstract/Résumé

[Tuesday May 28/mardi 28 mai, 15:30-15:45]

Qihuang Zhang (University of Waterloo) , **Grace Yi** (University of Waterloo)

Analysis of Bivariate Responses in Genetic Association Studies with Measurement Error and Misclassification

Analyse de réponses bivariées dans des études d'association génétique avec erreurs de mesure et classification erronée

In genetics, pleiotropy refers to the phenomenon that one single gene influences multiple traits. Statistical methods based on jointly modeling the association of a gene with a continuous trait and a binary trait have been widely used in genetic association studies. However, the effects of response mismeasurement are often neglected. In many settings, ignorance of mismeasurement in variables usually results in biased estimation. We consider the setting with a bivariate outcome vector that contains a continuous component and a binary component, both subject to mismeasurement. We propose an estimating function approach to handle measurement error in continuous response and misclassification in binary response simultaneously. The proposed estimators are consistent and robust to certain model misspecification. Simulation studies demonstrate that the proposed method successfully corrects the biasedness resulting from the error in variables.

Dans le domaine de la génétique, la pléiotropie désigne le phénomène quand un gène unique influence plusieurs traits. Les méthodes statistiques fondées sur la modélisation conjointe de l'association d'un gène avec un trait continu et un trait binaire ont été largement utilisées dans les études d'association génétiques. Par contre, les effets des erreurs de mesure de la réponse sont souvent négligés. Dans plusieurs contextes, ignorer les erreurs de mesure dans les variables résulte en des estimations biaisées. Nous examinons le contexte d'un vecteur de résultats bivarié qui contient une composante continue et une composante binaire, toutes deux soumises à des erreurs de mesure. Nous proposons une approche de fonction d'estimation pour traiter simultanément les erreurs de mesure dans la réponse continue et la classification erronée dans la réponse binaire. Les estimateurs proposés sont convergents et robustes face à certaines erreurs de spécification du modèle. Des études de simulation démontrent que la méthode proposée corrige avec succès le biais résultant des erreurs dans les variables.

[Tuesday May 28/mardi 28 mai, 15:45-16:00]

Joycelyne E Ewusie (University of Ottawa) , **Kelly Burkett** (University of Ottawa) , **Marie-Hélène Roy-Gagnon** (University of Ottawa)

Using External Controls to Account for Mating Asymmetry in Maternal Genetic Association

Prendre en compte l'asymétrie d'accouplement dans les études d'effets génétiques maternels en utilisant des donnée externes de parents témoins

In studying the effects of maternal and child genes on the risk of a disease with early onset, current analytic approaches depend on the assumption of mating symmetry when using case-parent trio data. Mating symmetry refers to the assumption that for any possible parental genotype pair, the frequency of a given mother and father genotype assignment in the population is equal to the reverse genotype assignment. The violation of this assumption (mating asymmetry) may lead to spurious maternal associations in studies of case-parent triads. Our study modifies a popular log-linear modeling approach to test and accommodate this assumption by us-

Les études de l'effet des gènes de la mère et de son enfant sur le risque de maladies se développant à un jeune âge utilisent souvent des échantillons de trios cas-parents. Ces études doivent poser l'hypothèse de symétrie d'accouplement qui implique que pour chaque paire possible des génotypes des deux parents, la fréquence d'une combinaison spécifique de génotypes du père et de la mère est la même que la combinaison inverse. Lorsque cette hypothèse n'est pas vérifiée (asymétrie d'accouplement), de fausses associations génétiques maternelles peuvent être détectées dans les études de trios cas-parents. Notre étude modifie un modèle log-linéaire populaire afin de tester et tenir compte de cette hypothèse en utilisant des statistiques sommaires obtenues de parents témoins pro-

ing summary statistics obtained from control parents for an external dataset. Simulations based on real data will be performed using different scenarios of mating asymmetry to assess the effect on the type 1 error and power of the test.

venant d'ensembles de données externes. Nous effectuerons des simulations basées sur des données réelles en utilisant différents scénarios d'asymétrie d'accouplement afin d'évaluer l'effet sur l'erreur de type I et la puissance des tests.

[Tuesday May 28/mardi 28 mai, 16:00-16:15]

Mei Dong (University of Saskatchewan) , **Longhai Li** (University of Saskatchewan) , **Lloyd Balbuena** (University of Saskatchewan)

Using External Cross Validation for Measuring Predictivity of Selected Features with Application to Genome Wide Predictive Analysis for Alzheimer's Disease

Validation croisée externe pour mesurer la prédictivité de certaines caractéristiques et application à l'analyse prédictive pangénomique de la maladie d'Alzheimer

Prediction is an important goal for current genome wide association studies and feature selection is an essential step to reduce noninformative single nucleotide polymorphisms (SNPs) due to the large size of genome data. However, selection bias is easily caused by using the whole dataset to select SNPs, which has not aroused enough attention. We use a semi-real and real dataset of Alzheimer's disease to assess the selection bias by comparing the predictive accuracy between models built by SNPs selected from the whole dataset and training data in each fold. Lasso, elastic net, and hyper-lasso are implemented to find the best classifier. Bias ranges from 0 to 30% accuracy for different feature subsets and datasets. The identified SNPs (except APOE) do not help in improving predictive accuracy (30% ER). Hyper-lasso has better predictive performance than the other two. Hence, cross validation external to the feature selection process must be implemented to obtain an honest prediction.

La prévision est un objectif important des études d'association pangénomique actuelles et la sélection de caractéristiques est une étape essentielle qui permet de réduire les polymorphismes singuliers de nucléotides (PSNs) non informatifs au vu du gros volume de données sur le génome. Cependant, un biais de sélection risque de se présenter si l'on utilise l'ensemble du jeu de données pour sélectionner les PSNs, un problème qui n'a pas fait l'objet d'assez de recherches. Nous utilisons un jeu de données semi-réelles et réelles sur la maladie d'Alzheimer pour évaluer le biais de sélection en comparant l'exactitude prédictive des modèles construits avec des PSNs sélectionnés sur l'ensemble du jeu de données et sur les données de formation pour chaque groupe. Nous mettons en œuvre le lasso, l'elastic net et l'hyper-lasso pour trouver le meilleur classificateur. Le biais varie de 0 à 30 % de précision selon les sous-ensembles de caractéristiques et les jeux de données. Les PSNs identifiés (sauf APOE) n'aident pas à améliorer l'exactitude prédictive (30 %). Nous concluons aussi que l'hyper-lasso présente une meilleure performance prédictive que les deux autres méthodes. Par conséquent, il faut employer une validation croisée externe au processus de sélection des caractéristiques pour obtenir une prédiction honnête.

[Tuesday May 28/mardi 28 mai, 16:15-16:30]

Changjiang Xu (University of Toronto) , **Gary Bader** (University of Toronto) , **Veronique Voisin** (University of Toronto) , **Ruth Isserlin** (University of Toronto) , **Jeff Liu** (University of Toronto)

Gene Set Analysis Using GSEA and GSVA: Performance Comparison and Improvement

Analyse d'ensembles de gènes à l'aide des méthodes d'enrichissement des ensembles de gènes (GSEA) et de la variation des ensembles de gènes (GSVA) : comparaison et amélioration de la performance

We compare GSEA and GSVA methods for gene set analysis by simulations in various scenarios. We examine the performance of the gene set analysis methods by calculating statistical power, type I error rate, and area under the ROC curve (AUC) for each method. Besides the original GSEA method, we further propose modified GSEA methods, GSEA.max and GSEA.min, to test

Nous comparons les méthodes GSEA et GSVA d'analyse des ensembles de gènes à l'aide de simulations selon divers scénarios. Nous examinons la performance des méthodes d'analyse d'ensembles de gènes en calculant pour chacune la puissance d'un test statistique, le taux d'erreurs de type-I et l'aire sous la courbe ROC (ASC). Outre la méthode GSEA originale, nous proposons aussi des méthodes GSEA modifiées, GSEA.max et GSEA.min

up- and down-regulated gene sets separately. We also compare the performance of the GSEA methods with sample permutation and gene permutation. The simulation results show: 1) the original GSEA with sample permutation has similar statistical power and AUC as the GSVA; 2) the GSEA.max and GSEA.min have better performance than the original GSEA and GSVA; 3) for each method the performance decreases as the correlation between genes increases; 4) the GSEA methods with sample permutation and gene permutation have similar statistical power when the correlation between genes is small.

afin d'évaluer séparément des ensembles de gènes régulés à la hausse ou à la baisse. Nous comparons aussi la performance des méthodes GSEA avec la permutation des échantillons et celle des gènes. Les résultats de la simulation indiquent que : 1) la puissance statistique de la méthode GSEA originale avec permutation des échantillons et AUC est semblable à celle de la méthode GSVA ; 2) la performance de GSEA.max et GSEA.min est meilleure que celle des méthodes GSEA et GSVA originales ; 3) la performance de chaque méthode diminue à mesure que s'accroît la corrélation entre les gènes ; 4) la puissance statistique des méthodes GSEA avec permutation des échantillons et avec permutation des gènes est semblable lorsqu'il y a une faible corrélation entre les gènes.

[Tuesday May 28/mardi 28 mai, 16:30-16:45]

Osvaldo Espin-Garcia (Dalla Lana School of Public Health, University of Toronto / Lunenfeld-Tanenbaum Research Institute) , **Radu Craiu** (University of Toronto) , **Shelley Bull** (University of Toronto Lunenfeld-Tanenbaum Research Institute)

Optimal Two-Phase Designs in Post-Genome-Wide Association Studies (GWAS)

Plan optimal en deux phases pour les études d'association post-pangénomiques (GWAS)

Two-phase sampling designs (TPDs) use data from phase 1, for example, outcome and/or inexpensive GWAS SNPs, to inform choice of a subsample in phase 2 for expensive next-generation sequencing. Inference is made on the missing-by-design sequence variants (SV) by combining information from phases 1 2. Given a postulated SV effect size and a SNP-SV haplotype distribution, we develop two approaches for optimal TPDs under a semi-parametric maximum likelihood framework: (1) a Laplace multiplier (LM) method based on an analytical expression for the variance-covariance matrix (VCM) subject to a budget constraint, and (2) a genetic algorithm (GA) that performs a direct search of the phase 2 sampling space. We evaluate VCM optimality criteria useful in experimental design. Comprehensive simulation studies to assess the empirical properties of LM and GA suggest that the optimal TPDs can render higher power compared to intuitive designs while preserving type 1 error.

Les plans d'échantillonnage en deux phases (PDP) utilisent des données de phase 1, p. ex. des SNP GWAS de résultat et/ou peu coûteux, pour conditionner le choix d'un sous-échantillon en phase 2 pour le séquençage de génération suivante plus coûteux. Une inférence est faite sur les variantes de séquence (VS) délibérément omises en combinant des informations des phases 1 2. Pour une ampleur d'effet sur les VS postulée et une distribution SNP-VS haplotype, nous développons deux approches pour un PDP optimal dans un cadre de maximum de vraisemblance semiparamétrique : (1) une méthode de multiplicateur de Laplace (ML) fondée sur une expression analytique de la matrice de variance-covariance (MVC) sous contrainte budgétaire, et (2) un algorithme génétique (AG) qui effectue une recherche directe sur l'espace d'échantillonnage de la phase 2. Nous évaluons les critères d'optimalité de la MVC utiles pour le plan d'expérience. La réalisation d'études complètes de simulation des propriétés empiriques du ML et de l'AG suggère que le PDP optimal est plus puissant qu'un plan intuitif tout en préservant l'erreur de type 1.

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

Chair/Président: Anita Brobbey

Room/Salle: 143 (ST)

Abstract/Résumé

[Tuesday May 28/mardi 28 mai, 15:30-15:45]

Shabnam Fani (University of Calgary) , **Hua Shen** (University of Calgary) , **Xuewen Lu** (University of Calgary) , **Jingjing Wu** (University of Calgary)

Semiparametric Regression with the U-Shaped Baseline Hazard Function in the Additive Hazards Model

Régression semiparamétrique avec fonction de risque de base en forme de U dans le modèle de risque additif

When employing the natural shape constraint knowledge obtained from prior studies in survival data analysis, both the selection difficulties including bandwidth and tuning parameter of nonparametric approaches and restrictions of parametric models can be avoided. We use this technique to derive a nonparametric estimator of the U-shaped baseline hazard function in the semiparametric additive hazards regression model for exact data, interval-censored data and a combination of both. A new algorithm, which overcomes the issue of finding the anti-mode of the U-shaped baseline hazard function, is developed for computing estimators of the baseline hazard function and regression coefficients simultaneously. In simulation studies, we compare the performance of the proposed approach incorporating shape information with that of the B-spline method, and show that the proposed method increases the efficiency of estimators for both the baseline hazard function and regression coefficients.

Lors de l'utilisation de l'information sur la contrainte de forme naturelle provenant d'études antérieures dans l'analyse de données de survie, les difficultés de sélection telles que le choix des paramètres de lissage et d'ajustement des approches non paramétriques et les restrictions des modèles paramétriques peuvent toutes deux être évitées. Nous utilisons cette technique pour obtenir un estimateur non paramétrique de la fonction de risque de base en forme de U dans le modèle de régression de risque additif semi-paramétrique pour données exactes, données avec censure par intervalle et une combinaison des deux. Le nouvel algorithme, qui surmonte l'obstacle de trouver l'anti-mode dans la fonction de risque de base en forme de U, est élaboré pour calculer simultanément les estimateurs de la fonction de risque de base et les coefficients de régression. En utilisant des données de simulations, nous comparons la performance de l'approche proposée en intégrant l'information sur la forme avec celle de la méthode B-spline et nous démontrons que la méthode proposée augmente l'efficacité des estimateurs pour la fonction de risque de base et pour les coefficients de régression.

[Tuesday May 28/mardi 28 mai, 15:45-16:00]

Longlong Huang (University of the Fraser Valley) , **Karen Kopciuk** (Cancer Control Alberta; University of Calgary) , **Xuewen Lu** (University of Calgary)

Adaptive Group Bridge Selection in the Semiparametric Accelerated Failure Time Model

Sélection bridge groupé adaptive dans le modèle semiparamétrique du temps de défaillance accéléré

Huang, Kopciuk and Lu (2016) considered group selection in a semiparametric accelerated failure time (AFT) model using Stute's weighted least squares and a group bridge penalty when covariates can be naturally grouped in survival analysis. Although the group bridge penalized approach can eectively remove unimportant groups, it cannot eectively remove unimportant variables within the important groups. To overcome this limitation, the adaptive group bridge method is proposed. We show that the adaptive group bridge method enjoys the powerful oracle property. Simulation studies indicate that the

Huang, Kopciuk et Lu (2016) ont examiné la sélection groupée dans un modèle semiparamétrique du temps de défaillance accéléré (TDA) en utilisant les moindres carrés pondérés de Stute et une pénalité bridge groupé quand les covariables peuvent être groupées naturellement dans une analyse de survie. Même si l'approche bridge groupé pénalisée peut efficacement retirer les groupes sans importance, elle ne peut pas retirer efficacement les variables sans importance des groupes importants. Pour surmonter cette limitation, la méthode bridge groupé adaptive est proposée. Nous démontrons que la méthode bridge groupé adaptive jouit de la propriété oracle. Des études de simulation indiquent que

adaptive group bridge approach for the AFT model can correctly identify important groups and within-group individual variables even with high censoring rates in high-dimensional data. The PBC data is analyzed to illustrate the application of the proposed method.

l'approche bridge groupé adaptive pour le modèle TDA peut correctement identifier les groupes importants ainsi que les variables individuelles importantes à l'intérieur des groupes même dans le cas de taux élevés de censure dans des données de grandes dimensions. Les données PBC sont analysées pour illustrer l'application de la méthode proposée.

[Tuesday May 28/mardi 28 mai, 16:00-16:15]

Jingyu Cui (University of New Brunswick) , **Renjun Ma** (University of New Brunswick) , **M. Tariq Hasan** (University of New Brunswick)

Generalized Linear Mixed Model with Crossed Random Effects

Modèles linéaires mixtes généralisés avec effets croisés aléatoires

A lot of studies are conducted to analyze data with nested random effects, where the random effects are subject to a hierarchical structure. However, in reality, it is often the case that the random effects are crossed. For example, in education studies, there is no hierarchical structure between middle and high schools. Instead, they are crossed. Also, in medical and health studies, patients living in the same ward may go to different practices, while the patients affiliated with the same practice may be from different areas. Thus, the random effects "ward" and "practice" are crossed. In the presentation, a Tweedie generalized linear model with crossed distribution-free random effects is used to accommodate a wide class of data. In addition, the orthodox best linear unbiased predictors of random effects are adopted to obtain the optimal estimating function for the regression parameters. The proposed model will be demonstrated using Gamma-distributed data.

Plusieurs études sont menées pour analyser des données avec effets aléatoires emboîtés où les effets aléatoires sont soumis à une structure hiérarchique. En réalité, il arrive souvent que les effets aléatoires soient croisés. Par exemple, dans des études en éducation, il n'y a pas de structure hiérarchique entre les écoles intermédiaires et secondaires ; elles sont plutôt croisées. De plus, dans des études médicales, des patients qui résident dans la même salle peuvent se rendre dans des pratiques différentes tandis que des patients de la même pratique peuvent provenir de zones différentes. Les effets aléatoires « salle » et « pratique » sont donc croisés. Dans cet exposé, un modèle Tweedie linéaire généralisé avec effets aléatoires croisés sans distribution spécifiée est utilisé pour intégrer une large classe de données. De plus, les meilleurs prédicteurs linéaires non-biaisés d'effets aléatoires suivant l'approche orthodoxe sont utilisés pour obtenir la fonction d'estimation optimale pour les paramètres de régression. Le modèle proposé sera démontré à l'aide de données à distribution gamma.

[Tuesday May 28/mardi 28 mai, 16:15-16:30]

Kaida Cai (University of Calgary) , **Xuewen Lu** (University of Calgary) , **Hua Shen** (University of Calgary)

Bi-Level Variable Selection for Multivariate Failure Time Data with Observed Heterogeneity

Sélection de variables à deux niveaux pour les données multivariées de temps de défaillance avec hétérogénéité observée

In this work, we propose a bi-level and an adaptive bi-level variable selection method for multivariate failure time data. In the case of observed heterogeneity, the average effects of covariates on different failure modes and the individual effects of the same covariate varies between different failure modes are both our concerns. Comparing with some classical L1 norm penalty methods, the proposed penalty methods can select the covariates with significant average effects and the covariates with significant individual effects between different failure modes simultaneously. Based on the simulation results, our methods perform better than the classical penalty methods, especially in terms of removing

Nous proposons ici une méthode de sélection de variables à deux niveaux et à deux niveaux adaptatifs pour les données multivariées de temps de défaillance. Dans le cas de l'hétérogénéité observée, nous nous intéressons à la fois aux effets moyens des covariables sur différents modes de défaillance et à ceux d'une même covariable variant entre différents modes de défaillance. En comparant avec certaines méthodes classiques de pénalité en norme L1, il est possible avec les méthodes de pénalité proposées de sélectionner simultanément les covariables avec des effets moyens importants et celles avec des effets individuels marqués entre divers modes de défaillance. Basées sur des résultats de simulation, nos méthodes offrent une meilleure performance que les méthodes classiques de pénalité, en particulier pour le retrait de covariables

insignificant covariates. An algorithm based on cycle coordinate descent is proposed to carry out the proposed methods. We investigate the asymptotic oracle properties of the proposed methods and construct generalized cross validation method for the tuning parameter selection.

négligeables. Nous mettons de l'avant un algorithme basé sur une descente cyclique des coordonnées pour l'exécution des méthodes proposées. Nous examinons les propriétés oracle asymptotiques des méthodes proposées et élaborons une méthode de validation croisée généralisée pour la sélection du paramètre d'ajustement.

SSC 2018 Impact Award Address
2018 Prix pour impact de la SSC

Chair/Président: Peter X Song

Organizer/Responsable: Carl James Schwarz

Room/Salle: 102 (ICT)

Abstract/Résumé

[Wednesday May 29/mercredi 29 mai, 08:40-09:50]

Geneviève Gauthier (HEC Montreal)

The Use of Filters and Large Databases in Financial Engineering

L'utilisation des filtres et des grandes bases de données en ingénierie financière

Financial models have unobservable variables that are sometimes difficult to estimate. Whether it is market volatility that varies continuously, risk of market correction or risk premiums, all of these variables are necessary for good measurement and good risk management, but are not directly observable. We use filtering techniques coupled with the incredible wealth of financial databases (we have access to intraday data, over a long period, in addition to large sets of derivatives such as options) to estimate these unobservable variables. The results of this extensive research program are achieved in collaboration with several colleagues and graduate students. I would like to thank Diego Amaya, Jean-François Bégin, Mathieu Boudreault, Christian Dorion, Delia Doljanu, Pascal Francois, Frédéric Godin, Tommy Thomassin for their precious collaboration.

Les modèles financiers comportent des variables non observables qui sont parfois difficiles à estimer. La volatilité des marchés qui varie de façon continue, le risque de correction boursière et les primes de risques, sont toutes des variables nécessaires pour une bonne mesure et une bonne gestion des risques, mais ne sont pas directement observables. Nous utilisons des techniques de filtrage couplées à l'incroyable richesse des bases de données financières (nous avons accès à des données intrajournalières, sur une longue période, en plus de vastes ensembles de produits dérivés tels les options) pour estimer ces variables non observables. Les résultats de ce vaste programme de recherche sont obtenus en collaboration avec plusieurs collègues et des étudiants de deuxième et troisième cycle. J'aimerais remercier Diego Amaya, Jean-François Bégin, Mathieu Boudreault, Christian Dorion, Delia Doljanu, Pascal François, Frédéric Godin, Tommy Thomassin pour leur précieuse collaboration.

CRM-SSC Prize in Statistics invited Address
Allocution de la récipiendaire du Prix CRM-SSC en statistique

Chair/Président: Alejandro Murua

Room/Salle: 148 (ST)

Abstract/Résumé

[Wednesday May 29/mercredi 29 mai, 08:40-09:50]

Johanna G. Neslehova (McGill University)

Tales of tails, tiles and ties in dependence modeling

La queue, la tuile, le bris d'égalité et leur rôle dans les modèles de dépendance

Modeling dependence between random variables is omnipresent in statistics. When rare events with high impact are involved, such as severe storms, floods or heat waves, the issue is both of great importance for risk management and theoretically challenging. Combining extreme-value theory with copula modeling and rank-based inference yields a particularly flexible and promising approach to this problem. I will present three recent advances in this area. One will tackle the question of how to account for dependence between rare events in the medium regime, in which asymptotic extreme-value models are not suitable. The other will explore what can be done when a large number of variables is involved and how a hierarchical model structure can be learned from large-scale rank correlation matrices. Finally, I won't resist giving you a glimpse of the notoriously intricate world of rank-based inference for discrete or mixed data.

La queue, la tuile, le bris d'égalité et leur rôle dans les modèles de dépendance La modélisation de la dépendance entre variables aléatoires est omniprésente en statistique. S'agissant d'événements rares à fort impact, tels que des orages violents, des inondations ou des vagues de chaleur, la question revêt une grande importance pour la gestion des risques et pose des défis théoriques. Une approche hautement flexible et prometteuse s'appuie sur la théorie des valeurs extrêmes, la modélisation par copules et l'inférence fondée sur les rangs. Je présenterai trois avancées récentes dans ce domaine. Nous nous intéresserons d'abord à la prise en compte de la dépendance en régime moyen, lorsque les modèles asymptotiques de valeurs extrêmes ne conviennent pas. Nous verrons ensuite quoi faire lorsque le nombre de variables est grand et comment une structure de modèle hiérarchique peut être apprise à partir de matrices de corrélation de rangs de grande taille. Enfin, je ne résisterai pas à l'envie de vous initier à l'univers complexe de l'inférence basée sur les rangs pour les données discrètes ou mixtes.

Recent Advances in Risk Theory
Dernières avancées en théorie des risques

Chair/Président: Bin Li

Organizer/Responsable: Bin Li

Room/Salle: 201 (ENG)

Abstract/Résumé

[Wednesday May 29/mercredi 29 mai, 10:20-10:50]

Alexey Kuznetsov (York University) , **Runhuan Feng** (University of Illinois at Urbana-Champaign) , **Fenghao Yang** (Royal Bank of Canada)

Exponential Functionals of Levy Processes and Variable Annuity Guaranteed Benefits

Fonctionnelles exponentielles de processus Levy et annuités variables avec prestations garanties

I will discuss the problem of calculating various risk measures for certain embedded options, known as variable annuity guaranteed benefits. Mathematically the problem boils down to finding the distribution of one curious random variable – the exponential functional of a Levy process. Such random variables have rich analytical structure and they are connected to many areas of pure and applied probability. Thus, in addition to the specific application of exponential functionals to variable annuity guaranteed benefits, I will also give an overview of the theory and applications of exponential functionals.

Je discuterai du problème du calcul de différentes mesures de risque pour certaines options intégrées connues sous le nom d'annuités variables avec prestations garanties. Mathématiquement, le problème se résume à trouver la distribution d'une curieuse variable aléatoire – la fonctionnelle exponentielle d'un processus Levy. De telles variables aléatoires possèdent une riche structure analytique et elles sont liées à plusieurs domaines des probabilités pures et appliquées. En plus de l'application spécifique des fonctionnelles exponentielles aux annuités variables avec prestations garanties, je donnerai un aperçu de la théorie et des applications des fonctionnelles exponentielles.

[Wednesday May 29/mercredi 29 mai, 10:50-11:20]

Jiandong Ren (Western University) , **Wenjun Jiang** (Western University) , **Hanping Hong** (Western University)

Reinsurance Policies with the Maximal Synergy Potential

Polices de réassurance ayant le potentiel maximal de synergie

Two types of optimality criteria are commonly used when studying optimal reinsurance in the literature: maximizing the expected utility (EU) and minimizing risks. These criteria give different optimal policies. In practice, insurance companies are likely to consider both criteria when negotiating reinsurance policies. One approach is to maximize EU under some risk constraints, as was done in Bernard and Tian (2010). An alternative approach was in fact proposed in Borch (1960), which assumed that the admissible reinsurance policies in maximizing EU should be such that the reduction in variance of losses through the reinsurance transaction is maximized. In this paper, following Borch (1960), we first identify the set of reinsurance policies that minimize the total risks, measured by distortion risk measures, shared by the two parties, then we take this set of policies as admissible and determine the Pareto

Deux types de critères d'optimalité sont souvent utilisés dans la littérature pour étudier la réassurance optimale : maximisation de l'espérance de l'utilité et minimisation des risques. Ces critères donnent des politiques optimales différentes. Dans la pratique, les compagnies d'assurance sont susceptibles de prendre en compte les deux critères lorsqu'elles négocient des polices de réassurance. Une approche consiste à maximiser l'espérance de l'utilité sous certaines contraintes de risque, comme l'ont fait Bernard et Tian (2010). En fait, Borch (1960) a proposé une autre approche, selon laquelle les polices de réassurance admissibles pour maximiser l'espérance de l'utilité devraient être telles que la réduction de la variance des pertes par la transaction de réassurance soit maximisée. Dans cet article, en suivant l'approche de Borch (1960), nous déterminons d'abord l'ensemble des polices de réassurance qui minimisent les risques totaux, mesurés par des mesures de risque de distorsion, partagés par les deux parties, puis nous considérons cet ensemble de polices comme admis-

optimal policies that maximize the EU of the two parties

sible et déterminons les polices optimales Pareto qui maximisent l'espérance de l'utilité des deux parties.

[Wednesday May 29/mercredi 29 mai, 11:20-11:50]

Yi Lu (Simon Fraser University) , **Shuanming Li** (University of Melbourne) , **Kristina Sendova** (University of Western Ontario)

The Expected Discounted Penalty Functions: From Infinite Time to Finite Time

Les fonctions de pénalité actualisées attendues : d'un temps infini à fini

In this talk, I will present some recent results on the finite-time ruin problems for the classical risk model and that perturbed by diffusion. Specifically, we study the finite-time expected discounted penalty function (EDPF). Using some techniques developed recently, we show that the finite-time EDPFs can be expressed in terms of their corresponding ones under the infinite-time horizon. In the perturbed risk model case, we find that the finite-time ruin probability due to oscillations and the finite-time ruin probability caused by a claim can also be expressed in terms of the corresponding quantities under the infinite-time horizon. Numerical examples are given when claims follow an exponential distribution.

Lors de cet exposé, je présenterai de récents résultats concernant les problèmes de ruine en temps fini du modèle de risque classique et de celui perturbé par la diffusion. Nous étudions spécifiquement la fonction de pénalité actualisée attendue (FPAA) en temps fini. Grâce à certaines techniques conçues dernièrement, nous démontrons que les FPAA en temps fini peuvent être exprimées par des fonctions correspondantes selon l'horizon en temps infini. Dans le cas du modèle de risque perturbé, nous avons découvert que la probabilité de ruine à temps fini causée par les oscillations et celle causée par une déclaration de sinistre peuvent être exprimées par des quantités correspondantes en fonction de l'horizon en temps infini. Nous procurons de nombreux exemples présentant des déclarations de sinistre suivant une loi exponentielle.

Extreme values
Valeurs extrêmes

Chair/Président: Gail B. Ivanoff

Organizer/Responsable: Gail B. Ivanoff

Room/Salle: 146 (SB)

Abstract/Résumé

[Wednesday May 29/mercredi 29 mai, 10:20-10:40]

Stilian Stoev (University of Michigan, Ann Arbor) , **Zheng Gao** (University of Michigan, Ann Arbor)

Concentration of Maxima and the Fundamental Limits of Exact Support Recovery in High Dimensions

Concentration des maxima et les limites fondamentales du support de redressement exact en haute dimension

We study the estimation of the support (set of non-zero components) of a high-dimensional signal observed with additive and dependent noise. With the usual parameterization of the size of the support set and the signal magnitude, we characterize a phase-transition phenomenon akin to the Ingster's signal detection boundary. Namely, when the signal is above the so-called strong classification boundary, thresholding estimators achieve asymptotically perfect support recovery. This is so under arbitrary error dependence assumptions, provided that the marginal error distribution has rapidly varying tails. Conversely, under mild dependence conditions on the noise, we show that no thresholding estimators can achieve perfect support recovery if the signal is below the boundary. For log-concave error densities, the thresholding estimators are shown to be optimal and hence the strong classification boundary is universal. The results are based on a concentration of maxima phenomenon.

Nous étudions l'estimation du support (ensemble de composants non nuls) d'un signal de haute dimension observé avec un bruit additif et dépendant. Au moyen du paramétrage habituel de la taille de l'ensemble de support et de l'amplitude du signal, nous caractérisons un phénomène de transition de phase semblable à la limite de détection du signal d'Ingster. En d'autres termes, lorsque le signal est au-dessus de la limite dite forte de classification, les estimateurs de seuillage permettent d'obtenir un redressement asymptotique parfait du support. Il en est ainsi dans le cas d'hypothèses arbitraires sur la dépendance de l'erreur, à condition que la distribution marginale de l'erreur ait des queues qui varient rapidement. Inversement, dans des conditions de légère dépendance au bruit, nous montrons qu'aucun estimateur de seuillage ne peut atteindre un redressement parfait du support si le signal est au-dessous de la limite. Pour les densités d'erreurs logarithmiques concaves, les estimateurs de seuillage se révèlent optimaux, de sorte que la forte limite de classification est universelle. Les résultats sont basés sur une concentration du phénomène des maxima.

[Wednesday May 29/mercredi 29 mai, 10:40-11:00]

Clemonell Lord Baronat Bilayi-Biakana (University of Ottawa) , **Gail Ivanoff** (University of Ottawa) , **Rafal Kulik** (University of Ottawa)

Heavy-Tailed Long Memory Stochastic Volatility Model with Leverage

Modèle de volatilité stochastique à mémoire longue et à queue lourde avec effet de levier

The tail empirical process (TEP) is an important tool used in nonparametric estimation of extremal quantities, like the Hill estimator of the index of regular variation, or various risk measures. In this talk, we consider a heavy-tailed long memory stochastic volatility model with leverage. We establish central and non-central limit theorems for the TEP via a martingale long-memory decomposition. Thanks to this probabilistic framework, we subsequently prove weak convergence of integral functionals of the tail empirical process. In turn, this

Le processus empirique de queue est un outil important utilisé dans l'estimation non paramétrique des quantités extrêmes, comme l'estimateur de Hill de l'indice de variation régulière, ou diverses mesures du risque. Dans le présent exposé, nous examinons un modèle de volatilité stochastique à mémoire longue et à queue lourde avec effet de levier. Nous établissons des théorèmes centraux et non-centraux limites pour le processus empirique de queue au moyen d'une martingale de décomposition à mémoire longue. Grâce à ce cadre probabiliste, nous démontrons par la suite la faible convergence des fonctions intégrales du processus

provides a unified approach to central limit theorems for estimators of the tail index and central/non central limit theorems of risk measures. This work is done under co-supervision of Professors Gail Ivanoff and Rafal Kulik.

empirique de la queue. Ainsi, on obtient une approche unifiée des théorèmes centraux limites pour les estimateurs de l'indice de queue, et des théorèmes centraux et non-centraux limites des mesures du risque. Ces travaux sont effectués sous la codirection des professeurs Gail Ivanoff et Rafal Kulik.

[Wednesday May 29/mercredi 29 mai, 11:00-11:20]

Natalia Nolde (University of British Columbia) , **Jinyuan Zhang** (INSEAD)

Conditional Extremes in Asymmetric Financial Markets

Les extrêmes conditionnels dans les marchés financiers asymétriques

The global financial crisis revealed the great extent to which systemic risk can jeopardize stability of the financial system. An effective methodology to quantify systemic risk is at the heart of the process of identifying the so-called systemically important financial institutions for regulatory purposes as well as to investigate key drivers of systemic contagion. The paper proposes a method for dynamic forecasting of CoVaR, a popular measure of systemic risk. As a first step, we develop a semi-parametric framework using extreme value theory (EVT) to model the conditional distribution of a bivariate random vector given that one of the components takes on a large value, taking into account important features of financial data such as asymmetry and heavy tails. In the second step, we embed the proposed EVT method into a dynamic framework via a bivariate GARCH process. An empirical analysis is conducted to demonstrate the performance of the proposed methodology.

La crise financière mondiale a révélé à quel point le risque systémique peut compromettre la stabilité du système financier. Une méthodologie efficace de quantification du risque systémique est au cœur du processus d'identification des institutions financières dites d'importance systémique à des fins réglementaires et d'enquête sur les principaux facteurs de contagion systémique. Dans cet article, nous proposons une méthode de prévision dynamique de la valeur à risque conditionnelle, une mesure populaire du risque systémique. Dans un premier temps, nous élaborons un cadre semi-paramétrique au moyen de la théorie des valeurs extrêmes pour modéliser la distribution conditionnelle d'un vecteur aléatoire bivarié étant donné que l'une des composantes prend une grande valeur, en tenant compte des caractéristiques importantes des données financières, comme l'asymétrie et les queues lourdes. Dans un deuxième temps, nous intégrons la méthode de la théorie des valeurs extrêmes proposée dans un cadre dynamique au moyen d'un processus GARCH bivarié. Nous effectuons une analyse empirique pour démontrer l'efficacité de la méthodologie proposée.

[Wednesday May 29/mercredi 29 mai, 11:20-11:40]

Rafal Kulik (University of Ottawa) , **Philippe Soulier** (Paris-Nanterre)

Limit Theorems for Empirical Cluster Functionals

Théorèmes limites pour les fonctionnelles de grappe empiriques

Limit theorems for empirical cluster functionals are discussed. Conditions for weak convergence are provided in terms of tail and spectral tail processes and can be verified for a large class of multivariate time series, including geometrically ergodic Markov chains. Applications include asymptotic normality of blocks and runs estimators for the extremal index and other cluster indices. Results for multiplier bootstrap processes are also provided.

Nous discutons des théorèmes limites pour les fonctionnelles de grappe empiriques. Les conditions pour une convergence faible sont fournies par des processus de queue et de queue spectrale et peuvent être vérifiées pour une vaste catégorie de séries temporelles multivariées, y compris des chaînes de Markov géométriquement ergodiques. Les applications comprennent la normalité asymptotique des blocs et d'estimateurs de parcours pour l'indice des extrêmes et autres indices de grappes. Sont aussi présentés des résultats de processus de bootstrap multiplicateur.

Building the Pipeline: the International Data Science in Schools Project
Construire l'avenir : le projet "International Data Science in Schools Project"

Chair/Président: Alison L. Gibbs

Organizer/Responsable: Alison L. Gibbs

Room/Salle: 102 (ICT)

Abstract/Résumé

[Wednesday May 29/mercredi 29 mai, 10:20-10:50]

Alison L. Gibbs (University of Toronto)

Introduction to the International Data Science in Schools Project

Introduction au Projet International de la Science des Données dans les Écoles

What should every school student know about learning from data? How can we encourage talented secondary school students to pursue advanced studies in Data Science? And how could we design a course for students and their teachers with these goals in mind? Motivated by these questions, the International Data Science in Schools Project (IDSSP) is working on a curriculum for a data science course for senior secondary school students, assuming only a limited background in mathematics and no computer programming prerequisites. IDSSP is an international collaboration between statisticians, computer scientists, and educators, with representatives from Australia, Canada, England, Germany, the Netherlands, New Zealand and the United States. The Canadian participation in the IDSSP is supported by the SSC. We'll describe the project's goals, approaches, and outcomes to date, and seek your input on the project and your ideas for how the project can have a greater impact.

Qu'est-ce que chaque élève devrait connaître sur l'apprentissage à partir de données ? Comment pouvons-nous encourager des étudiants talentueux du secondaire à poursuivre des études dans la science des données ? Et comment pourrions-nous concevoir un cours pour des étudiants et leurs enseignants avec ces objectifs en tête ? Motivé par ces questions, le Projet International de la Science des Données dans les Écoles (PISDE) travaille sur un curriculum pour un cours de science des données pour étudiants au deuxième cycle du secondaire, présumant seulement une formation limitée en mathématiques et aucun prérequis en programmation. PISDE est une collaboration internationale entre statisticiens, informaticiens et éducateurs provenant de l'Australie, Canada, Angleterre, Allemagne, les Pays-Bas, la Nouvelle-Zélande et les États-Unis. La participation canadienne est appuyée par la SSC. Nous décrirons les buts, approches et résultats actuels du projet et nous solliciterons votre avis sur le projet et sur comment il pourrait avoir un plus grand impact.

[Wednesday May 29/mercredi 29 mai, 10:50-11:20]

Wesley Burr (Trent University)

Case Studies in Data Science Education: Limits and Scope

Études de cas dans l'enseignement de la science des données : limites et champ d'application

The International Data Science in Schools Project consists of 17 modular topics which have had curriculum developed: two proposed years of material (years 11 and 12 of secondary school). The first year has seven serial topics covering the Data Science Learning Cycle, while the second has ten parallel modules. The case studies were developed for the second year, to aid instructors in learning and understanding the importance and relevance of given topics. In this talk we will examine two case studies (Time Series and Maps) as instructive examples of our philosophy in this project: where should

Le Projet International de la Science des Données dans les Écoles propose 17 sujets modulaires pour lesquels un programme d'études a été élaboré : du matériel pour deux années proposées (les 11e et 12e années du secondaire). Dès la première année, une série de sept sujets couvre tout le cycle d'apprentissage en science des données, tandis que dix modules parallèles sont étudiés en deuxième année. Les études de cas ont été conçues pour la deuxième année, afin d'aider les enseignants à apprendre et comprendre l'importance et la pertinence de certains sujets. Nous allons examiner ici deux études de cas (séries et cartes temporelles) qui servent d'exemples instructifs de notre conception

boundaries be set? how closed-form should projects be? what topics are most suitable for case studies? when should maturity and sophistication in analyses be introduced? The requirements for teaching data science to secondary students can be quite different than university or college, due to age, maturity and experience, and case studies can be an effective way to motivate and inspire students.

de ce projet : où devons-nous établir des limites ? À quel point les projets doivent être de forme fermée ? Quels sujets se prêtent mieux à une étude de cas ? Quand doit-on présenter les aspects de l'échéance et de la sophistication des analyses ? En raison de leur âge, maturité et expérience, les exigences pour l'enseignement de la science des données aux élèves du secondaire peuvent largement différer de celles en vigueur à l'université ou au collège et les études de cas peuvent être efficaces pour les motiver et les inspirer.

[Wednesday May 29/mercredi 29 mai, 11:20-11:50]

Robert Gould (UCLA)

Implementing a Data Science Course in Secondary Schools

Instaurer un cours sur la science des données dans les écoles secondaires

Implementing a data science curriculum in secondary schools, such as the one proposed by the International Data Science in Schools Project, may sound daunting. Using our experiences implementing the Mobilize Introduction to Data Science (IDS) course in the Los Angeles Unified School District as a case study, we will discuss the challenges faced and offer approaches that may provide viable pathways to establishing data science curricula in local schools. IDS is now taught in 14 school districts and 42 schools, and the success is due in part to a strong professional development program and the ability to situate the course within the mathematics curriculum. We'll discuss the need for professional development, attitudes towards data science by teachers, counselors and administrators, and offer suggestions for establishing similar programs elsewhere.

Instaurer un curriculum sur la science des données dans les écoles secondaires, tel que proposé par le Project International de la Science des Données dans les Écoles, peut sembler intimidant. En utilisant notre expérience dans l'implémentation du cours Mobiliser l'Introduction à la Science des Données (ISD) dans le Los Angeles Unified School District comme étude de cas, nous discuterons des défis rencontrés et nous présenterons des approches qui peuvent fournir des méthodes viables pour la création d'un curriculum en science des données dans les écoles locales. ISD est maintenant enseigné dans 14 districts scolaires et dans 42 écoles et son succès est dû en partie à un solide programme de développement professionnel ainsi qu'à la capacité de situer le cours à l'intérieur du programme de mathématiques. Nous discuterons de la nécessité du développement professionnel, de l'attitude des enseignants, conseillers et gestionnaires face à la science des données et nous offrirons des suggestions pour l'instauration d'un programme similaire.

CJS Award Address
Allocution du récipiendaire du Prix de la RCS

Chair/Président: Louis-Paul Rivest

Organizer/Responsable: Louis-Paul Rivest

Room/Salle: 122 (ICT)

Abstract/Résumé

[Wednesday May 29/mercredi 29 mai, 10:20-11:25]

Radu Craiu (University of Toronto)

LISA for BART

Algorithme d'échantillonnage de vraisemblance gonflée pour les arbres additifs de régression bayésienne

Markov chain Monte Carlo (MCMC) sampling from a posterior distributions corresponding to a massive data set can be computationally prohibitive since producing one sample requires a number of operations that is linear in the data size. A new communication-free parallel method, the Likelihood Inflating Sampling Algorithm (LISA) is introduced. LISA significantly reduces computational costs by randomly splitting the data set into smaller subsets and running MCMC methods independently in parallel on each subset using different processors. Each processor is used to run an MCMC chain that samples sub-posterior distributions which are defined using an inflated likelihood function. We develop a strategy to combining the draws from different sub-posteriors to study the Bayesian Additive Regression Trees (BART). The performance of the method is tested using simulated data and a large socio-economic analysis.

L'échantillonnage par la méthode de Monte Carlo par chaîne de Markov effectué à partir de probabilités a posteriori, dans le cadre d'un ensemble de mégadonnées, peut être prohibitif sur le plan calculatoire, car la production d'un échantillon nécessite un certain nombre d'opérations linéaires avec la taille des données. Nous présentons une nouvelle méthode parallèle sans communication : l'algorithme d'échantillonnage de vraisemblance gonflée (AEVG). Cet algorithme réduit significativement les coûts de calculs en divisant aléatoirement l'ensemble des données en sous-ensembles plus petits et en appliquant les méthodes Monte Carlo par chaîne de Markov de manière indépendante et parallèle sur chaque sous-ensemble au moyen de différents processeurs. Chaque processeur est utilisé pour mettre en œuvre une méthode Monte Carlo par chaîne de Markov qui échantillonne les probabilités a posteriori définies à l'aide d'une fonction de vraisemblance gonflée. Nous élaborons une stratégie pour combiner les tirages de différentes distributions a posteriori afin d'étudier les arbres additifs de régression bayésienne (AARB). L'efficacité de la méthode est testée à l'aide de données simulées et d'une vaste étude socio-économique.

Statistical challenges and methods for clinical trial design
Défis et méthodes statistiques pour la conception d'essais cliniques

Chair/Président: Depeng Jiang

Organizer/Responsable: Depeng Jiang

Room/Salle: 142 (AD)

Abstract/Résumé

[Wednesday May 29/mercredi 29 mai, 10:20-10:50]

Ying Yuan (University of Texas MD Anderson Cancer Center) , **Ruitao Lin** (University of Texas MD Anderson Cancer Center) , **Daniel Li** (Juno Therapeutics) , **Lei Nie** (Food and Drug Administration) , **Katherine Warren** (National Cancer Institute)

Time-To-Event Bayesian Optimal Interval Design to Accelerate Phase I Trials

Plan d'intervalle optimal bayésien de temps avant l'évènement pour accélérer les essais de phase I

Late-onset toxicity is common for novel molecularly targeted agents and immunotherapy. It causes major logistic difficulty for existing adaptive phase I trial designs, which require the observance of toxicity early enough to apply dose escalation rules for new patients. We propose the time-to-event Bayesian optimal interval (TITE-BOIN) design to accelerate phase I trials by allowing for real-time dose assignment decisions for new patients while some enrolled patients' toxicity data are still pending. Similar to the rolling six design, the TITE-BOIN dose escalation/de-escalation rule can be tabulated before the trial begins, making it transparent and simple to implement, but is more flexible in choosing the target DLT rate and has higher accuracy to identify the MTD. Compared to the more complicated model-based time-to-event continuous reassessment method (TITE-CRM), the TITE-BOIN has comparable accuracy to identify the MTD, but is simple to implement with better overdose control.

La toxicité tardive est un problème commun dans les nouveaux agents moléculaires ciblés et l'immunothérapie. Elle cause de graves difficultés logistiques pour les plans d'essais de phase I adaptatifs existants, qui exigent que la toxicité soit observée assez tôt pour permettre l'application de règles d'augmentation de dose aux nouveaux patients. Nous proposons un plan d'intervalle optimal bayésien de temps avant l'évènement dit TITE-BOIN, qui permettra d'accélérer les essais de phase I puisque la détermination des doses pourra se faire en temps réel pour les nouveaux patients avant même de connaitre les données de toxicité de tous les patients inscrits. De même que dans le plan « rolling six » (six mobiles), la règle d'augmentation/réduction de la dose TITE-BOIN peut être tabulée avant le début de l'essai, ce qui facilite sa mise en application et la rend transparente ; par ailleurs, le plan permet plus de souplesse dans le choix de la toxicité qui limite la dose et plus de précision dans la détermination de la dose maximale tolérée. Par comparaison à la méthode plus compliquée basée sur un modèle de réévaluation continue du temps avant l'évènement (TITE-CRM), la méthode TITE-BOIN présente une précision comparable quant à la dose maximale tolérée, mais elle est plus simple à mettre en œuvre et permet un meilleur contrôle de la surdose.

[Wednesday May 29/mercredi 29 mai, 10:50-11:20]

Bo Huang (Pfizer) , **Xiaodong Luo** (Sanofi) , **Hui Quan** (Sanofi)

Design and Monitoring of Survival Trials in the Presence of Non-Proportional Hazard

Conception et suivi d'essais de survie en présence de risque non-proportionnel

With the emergence of novel therapies such as immunotherapies, the proportional hazard assumption, commonly assumed in oncology clinical trials, may no longer be valid. The use of the log-rank test can result in significant power loss and the hazard ratio is difficult to interpret. We propose a flexible method

Avec l'émergence de nouvelles thérapies telles que les immunothérapies, l'hypothèse de risque proportionnel couramment présumée lors d'essais cliniques en oncologie peut ne plus être valide. L'utilisation du test du log-rank peut entraîner une perte significative de puissance et le rapport de risque est difficile à interpréter. Nous proposons une méthode flexible pour la

to conduct study design and monitoring based on the restricted mean survival time (RMST). We illustrate that, with event time and censoring time following a piecewise exponential distribution, the RMSTs and their variance-covariance structure can be conveniently computed, which greatly facilitates study design and monitoring. As the number of pieces of the exponential distributions can be arbitrary, this approach can handle a wide range of scenarios. One hypothetical example is presented to demonstrate its potential use.

conception et le suivi d'une étude fondée sur le temps de survie moyen restreint (TSMR). Nous démontrons que lorsque les temps d'événement et de censure suivent une distribution exponentielle par morceaux, les TSMRs et leurs structures de variance-covariance peuvent être aisément calculées, ce qui facilite grandement la conception et le suivi de l'étude. Comme le nombre de morceaux de la distribution exponentielle peut être arbitraire, cette approche peut s'appliquer à une grande variété de scénarios. Un exemple hypothétique est présenté pour démontrer son utilité potentielle.

[Wednesday May 29/mercredi 29 mai, 11:20-11:50]

Suyu Liu (MD Anderson Cancer Center) , **Beibei Guo** (Louisiana State University) , **Ying Yuan** (MD Anderson Cancer Center)

A Bayesian Phase I/II Trial Design for Immunotherapy

Un plan d'essai bayésien de phases I et II pour l'immunothérapie

Immunotherapy is an innovative treatment approach that stimulates a patient's immune system to fight cancer. It demonstrates characteristics distinct from conventional chemotherapy and stands to revolutionize cancer treatment. We propose a Bayesian phase I/II dose-finding design that incorporates the unique features of immunotherapy by simultaneously considering three outcomes: immune response, toxicity and efficacy. The objective is to identify the biologically optimal dose, defined as the dose with the highest desirability in the risk-benefit tradeoff. An Emax model is utilized to describe the marginal distribution of the immune response. Conditional on the immune response, we jointly model toxicity and efficacy using a latent variable approach. Using the accumulating data, we adaptively randomize patients to experimental doses based on the continuously updated model estimates. A simulation study shows that our proposed design has good operating characteristics.

L'immunothérapie est une approche thérapeutique novatrice qui stimule le système immunitaire du patient pour combattre le cancer. Elle présente des caractéristiques distinctes de la chimiothérapie conventionnelle et est sur le point de révolutionner le traitement du cancer. Nous proposons un modèle bayésien d'établissement de la posologie de phases I et II qui intègre les caractéristiques uniques de l'immunothérapie en tenant simultanément compte de trois résultats : réponse immunitaire, toxicité et efficacité. L'objectif est d'établir la posologie biologiquement optimale, définie comme la posologie la plus souhaitable dans le compromis risque-avantage. Nous utilisons un modèle Emax pour décrire la distribution marginale de la réponse immunitaire. Selon la réponse immunitaire, nous modélisons conjointement la toxicité et l'efficacité au moyen d'une approche à variables latentes. À l'aide des données accumulées, nous randomisons de façon adaptative les patients aux posologies expérimentales en nous basant sur les estimations du modèle continuellement mises à jour. Une étude de simulation montre que la conception que nous proposons présente de bonnes caractéristiques de fonctionnement.

Advances in Distribution Theory
Progrès en matière de théorie de la distribution

Chair/Président: Alberto Nettel-Aguirre

Room/Salle: 109 (SS)

Abstract/Résumé

[Wednesday May 29/mercredi 29 mai, 10:20-10:35]

Liu Yi (University of Alberta) , **Peng Liu** (University of Alberta) , **Rui Zhu** (University of Alberta) , **Linglong Kong** (University of Alberta) , **Bei Jiang** (University of Alberta) , **Di Niu** (University of Alberta)

Optimal Smooth Approximation for Quantile Matrix Factorization

Approximation optimale lisse pour la factorisation quantile de matrice

Matrix estimation and factorization have wide applications in signal processing and recommender systems. Most existing matrix factorization methods adopt a squared loss function and aim to recover a low-rank matrix to interpret conditional means of matrix entries given noisy observations. Quantile matrix factorization (QMF) adopts a check loss in matrix factorization and can better explain the central tendency of data under realistic noise distributions such as skewed and heavy-tailed noise. In this paper, we propose NsQMF, a nearly optimal smooth approximation procedure for QMF by extending Nesterov's optimal smooth approximation technique to nonconvex matrix factorization problems. We theoretically show that solving the nonsmooth QMF problem is equivalent to solving the proposed smooth approximation. We then present an efficient algorithm to solve QMF by adapting alternating minimization and a singular value projection algorithm to the proposed NsQMF.

L'estimation et la factorisation de matrices ont plusieurs applications dans le traitement des signaux et dans les systèmes recommandeurs. La plupart des méthodes actuelles de factorisation de matrices utilisent la fonction de perte quadratique et ont pour but de recouvrer une matrice de faible rang pour interpréter les moyennes conditionnelles des entrées des matrices lorsqu'il y a des observations bruitées. La factorisation quantile de matrice (FQM) utilise la fonction de perte check dans la factorisation de matrice et contribue à mieux expliquer la tendance centrale des données sous des hypothèses distributionnelles réalistes sur le bruit telles que le bruit asymétrique et à aile lourde. Dans cet exposé, nous présentons NsFQM, une procédure d'approximation lisse presque optimale pour la FQM en étendant la technique d'approximation lisse optimale de Nesterov aux problèmes de factorisation de matrice non-convexe. Nous démontrons théoriquement que résoudre le problème de la FQM non-lisse est équivalent à résoudre l'approximation lisse proposée. Nous présentons ensuite un algorithme efficace pour résoudre la FQM en adaptant un algorithme de minimisation en alternance et un algorithme de prévision à valeur singulière au NsFQM proposé.

[Wednesday May 29/mercredi 29 mai, 10:35-10:50]

Yuan Sun (University of Michigan, Ann Arbor) , **Xuming He** (University of Michigan, Ann Arbor)

A Model-Based Bootstrap Method for Regional Quantiles Treatment Effects Detection with a Quantile Regression Rank Test

Méthode bootstrap fondée sur un modèle afin de détecter des effets de traitement sur les quantiles régionaux avec un test de rang par régression quantile

Quantile treatment effects are often considered in a quantile regression framework to adjust for the effects of covariates. In this study, we focus on the problem of testing whether the treatment effects are significant for a set of quantile levels (e.g., lower quantiles). We propose a quantile regression rank test, which is a generalization of the rank score test at an individual quantile level. This test statistic allows us to detect the treatment effect for a prespecified quantile interval by integrating the regression rank score over a region of interest. A model-

Les effets de traitement sur les quantiles sont souvent pris en compte dans un cadre de régression quantile pour tenir compte des effets des covariables. Dans cette étude, nous nous concentrons sur la difficulté à vérifier si les effets du traitement sont significatifs sur un ensemble de quantiles (p. ex., les quantiles inférieurs). Nous proposons un test de rang par régression quantile, qui est une généralisation du test de rang par score à un niveau de quantile individuel. Ce test statistique nous permet de détecter un effet de traitement pour un intervalle quantile prédéfini en intégrant le score de rang de régression sur une région d'intérêt. Nous met-

based bootstrap method is constructed to estimate the null distribution of the test statistic. A simulation study is conducted to demonstrate the validity and usefulness of the proposed test. We further apply our method to analyze the 2016 US birth weight data and the SP 500 index data.

tons sur pied une méthode bootstrap fondée sur un modèle pour estimer la distribution nulle de la statistique du test. Nous menons une étude de simulation pour démontrer la validité et l'utilité du test proposé. Nous appliquons également notre méthode pour analyser les données de 2016 relatives au poids à la naissance des Américains et les données de l'indice SP 500.

[Wednesday May 29/mercredi 29 mai, 10:50-11:05]

Matthew Pietrosanu (University of Alberta) , **Dengdeng Yu** (University of Toronto) , **Linglong Kong** (University of Alberta)

Extending Partial Quantile Regression to Multidimensional Functional Linear Models via Tensor Decomposition

Extension de la régression quantile partielle aux modèles linéaires fonctionnels multidimensionnels par la décomposition tensorielle

We have previously shown partial quantile regression (PQR) to be a robust procedure for functional linear quantile regression. PQR employs partial quantile covariance in a supervised approach to extract a basis and estimate functional coefficients. We have demonstrated that PQR is applicable to functional covariates of a single variable, although the extension to functions of multiple variables was not immediate. In this presentation, we propose a generalization of PQR and implement a procedure to estimate multidimensional functional linear models. This approach makes particular use of tensor decomposition to reduce the dimension of functional data and block relaxation techniques for model estimation. We discuss the asymptotic properties of our estimators using techniques in empirical process theory. Lastly, we illustrate the application of our data using standard simulations and real-world neuroimaging tensor data and explore its performance under various conditions and parameters.

Nous avons précédemment démontré que la régression quantile partielle (RQP) est une procédure robuste pour la régression quantile linéaire fonctionnelle. RQP utilise la covariance quantile partielle dans une approche supervisée pour extraire une base et pour estimer les coefficients fonctionnels. Nous avons démontré que RQP peut s'appliquer à des covariables fonctionnelles d'une seule variable même si l'extension aux fonctions à variables multiples n'est pas immédiate. Dans cet exposé, nous proposons une généralisation de RQP et nous établissons une procédure pour estimer des modèles linéaires fonctionnels multidimensionnels. Cette approche utilise la décomposition tensorielle pour réduire la dimension des données fonctionnelles et les techniques de relaxation par blocs pour l'estimation du modèle. Nous discutons les propriétés asymptotiques de nos estimateurs en utilisant des techniques de la théorie du processus empirique. Finalement, nous illustrons l'application de nos données à l'aide de simulations et de données tensorielles réelles en neuroimagerie et nous examinons sa performance sous différentes conditions et paramètres.

[Wednesday May 29/mercredi 29 mai, 11:05-11:20]

Yi Lian (McGill University) , **Yi Yang** (McGill University) , **Robert Platt** (McGill University)

Tweedie Compound Poisson Model in the Reproducing Kernel Hilbert Space

Modèle de Poisson composé Tweedie appliqué à l'espace de Hilbert à noyau reproduisant

The Tweedie Compound Poisson model is a class of generalized linear model widely used in insurance cost prediction. Being a linear model in nature, the model's capability to capture complex non-linear patterns is limited. Motived by the increased need in accurate prediction of insurance costs in healthcare and business, we proposed a more flexible non-parametric Tweedie model in a reproducing kernel Hilbert space. We developed an efficient multi-layer algorithm for estimating the insurance cost. Classic profile likelihood approach was used to estimate the (nuisance) index and disper-

Le modèle de Poisson composé Tweedie est un type de modèle linéaire généralisé fréquemment utilisé dans la prédiction du coût d'une assurance. Étant donné la nature linéaire du modèle, son potentiel à saisir des tendances non linéaires complexes est limité. Encouragé par le besoin grandissant de prédiction juste du coût des assurances dans les entreprises et dans le domaine des soins de santé, nous avons proposé un modèle Tweedie non paramétrique à polyvalence supérieure dans un espace de Hillbert à noyau reproduisant. Nous avons conçu un algorithme multicouche efficace pour estimer le coût d'une assurance. Nous avons adopté une approche de vraisemblance profilée classique pour es-

sion parameters. We compared the efficiency and accuracy of our model to those of existing methods. Results showed the superiority of our model and algorithm. We also implemented our algorithm in an R package.

timer l'indice (de nuisance) et les paramètres de dispersion. Nous avons comparé l'efficacité et la précision de notre modèle par rapport aux méthodes actuelles, et les résultats obtenus démontrent sa supériorité (et aussi celle de notre algorithme). Nous avons aussi intégré notre algorithme dans une bibliothèque R.

[Wednesday May 29/mercredi 29 mai, 11:20-11:35]

René Ferland (Université du Québec à Montréal) , **François Watier** (Université du Québec à Montréal)

Goal Achieving Probabilities of Regime-Switching Mean-Variance Portfolios

Probabilités de réussite des objectifs de portefeuilles moyenne-variance à changement de régime

In a continuous-time market model where stocks prices are driven by a Brownian motion and the volatility follows a regime-switching process, we will study first passage time properties related to an optimal mean-variance strategy. In particular, we will compute the probability to reach a random time, within the investment horizon, when the individual's generated wealth is large enough so he can safely reinvest all of his money in a bank account with fixed interest and eventually achieve his financial goal.

Dans le cadre d'un modèle de marché en temps continu, où les prix des actions sont guidés par un mouvement brownien et la volatilité suit un processus de changement de régime, nous étudierons les propriétés temporelles d'un premier passage relatif à une stratégie optimale basée sur la moyenne et la variance. Tout particulièrement, nous calculerons la probabilité d'atteindre un moment aléatoire, à l'intérieur d'un horizon de placement, lorsque le capital d'un individu sera assez élevé pour réinvestir tout son argent dans un compte de banque à intérêt fixe, puis un jour atteindre son objectif financier.

[Wednesday May 29/mercredi 29 mai, 11:35-11:50]

Salma Saad (University of Regina) , **Andrei Volodin** (University of Regina)

Asymptotic Analysis of Method of Moments

Analyse asymptotique de la méthode des moments

An estimation of parameters of the binomial distribution by a sample of fixed size n, when both parameters m and p are unknown, has remained an important statistical problem for more than three quarters of a century. Known estimates of m usually underestimate the true value. We consider only the Method of Moments and its modifications for estimation of parameters m and p of the binomial distribution. We also apply the delta method for the proof of asymptotic normality of the joint distribution of the estimators of m and p by the Method of Moments. We are mostly interested in the bias and variance of the Method of Moments and its modifications estimators. To achieve these goals it is necessary to solve the following problems: (1) derivation of estimates of parameters of binomial distribution by the Method of Moments; and (2) Derivation of the parameters of asymptotic normality by the delta-method.

L'estimation des paramètres de la distribution binomiale par un échantillon de taille fixe n, lorsque les deux paramètres m et p sont inconnus, demeure un problème statistique important depuis plus de trois quarts de siècle. Les estimations connues de m sous-estiment en général la valeur vraie. Nous ne considérons que la méthode des moments et ses modifications pour l'estimation des paramètres m et p de la distribution binomiale. Nous appliquons également la méthode delta pour la preuve de la normalité asymptotique de la distribution conjointe des estimateurs de m et p par la méthode des moments. Nous nous intéressons surtout au biais et à la variance de la méthode des moments et de ses estimateurs de modifications. Pour atteindre ces objectifs, il est nécessaire de résoudre les problèmes suivants : 1) dérivation des estimations des paramètres de distribution binomiale par la méthode des moments, et 2) dérivation des paramètres de normalité asymptotique par la méthode delta.

Novel Biostatistical Methods
Nouvelles méthodes biostatistiques

Chair/Président: Angelo J. Canty

Room/Salle: 201 (ENA)

Abstract/Résumé

[Wednesday May 29/mercredi 29 mai, 10:20-10:35]

Thierry Chekouo (University of Calgary) , **Himadri Mukherjee** (University of Minnesota Duluth)

Model-Based Clustering and Gene Selection via Bayesian Hierarchical Hidden Markov Models

Regroupement fondé sur un modèle et sélection des gènes via des modèles hiérarchiques bayésiens de Markov cachés

I will present a Bayesian hierarchical model that simultaneously performs clustering and feature selection. The model is combined with hidden Markov processes with three states for modeling functional dependence between features. The three states of the hidden Markov process allow us to obtain biologically meaningful clusters and to better discriminate them. Both simulation studies and gene expression analysis in a kidney cancer study illustrate the reliability and success of this method. We used Gene Ontology to define functional similarities between genes.

Dans cet exposé, je présenterai un modèle hiérarchique bayésien qui effectue simultanément un regroupement et une sélection de variables. Le modèle est combiné avec des processus de Markov cachés à trois états pour modéliser la dépendance fonctionnelle entre les gènes. Les trois états du processus de Markov caché nous permettent d'obtenir des groupes biologiquement pertinents et de mieux les discriminer. Les études de simulation et l'analyse de l'expression génétique dans une étude sur le cancer du rein illustrent la fiabilité et le succès de cette méthode. Nous avons utilisé l'ontologie des gènes pour définir les similitudes fonctionnelles entre les gènes.

[Wednesday May 29/mercredi 29 mai, 10:35-10:50]

Maryam Yetunde Onifade (University of Ottawa) , **Kelly Burkett** (University of Ottawa)

Comparison of Mixed Model-Based Approaches for Correcting for Population Substructure with Application to Extreme Phenotype Sampling

Comparaison d'approches à modèles mixtes pour corriger l'effet de la sous-structure de la population et application à l'échantillonnage de phénotypes extrêmes

Mixed models have been useful in correcting for confounding due to population stratification and hidden relatedness in genome wide association studies. This class of models includes linear mixed models (LMM) and generalised linear mixed models (GLMM). Existing mixed model approaches to correct for population substructure have been investigated with both continuous and case/control response variables. However, they have not been investigated in the context of extreme phenotype sampling (EPS), where genetic covariates are only collected on samples having extreme response variable values. In this work, we compare the performance of existing mixed model approaches (LTMLM, GMMAT) with EPS data analysed as a binary trait. We use simulation to estimate the type 1 error of all approaches when there is confounding. Since LMMs are commonly used even with binary traits, we also analysed the data using a LMM. This work was done under the supervision of Kelly Burkett.

Les modèles mixtes sont utiles pour corriger l'effet de confusion dû à la stratification de la population et à la parenté cachée dans les études d'association pangénomique. Cette classe de modèles inclut les modèles mixtes linéaires (MML) et les modèles mixtes linéaires généralisés (MMLG). Les modèles mixtes utilisés pour corriger l'effet de la sous-structure de la population ont été étudiés dans les cas de variables de réponses continues et cas/témoins. Cependant, ils n'ont fait l'objet d'aucune étude dans le contexte de « l'échantillonnage de phénotypes extrêmes » (EPE), où l'on ne collecte de covariables génétiques que sur des échantillons présentant des valeurs de variables réponses extrêmes. Dans cette présentation, nous comparons la performance des approches à modèles mixtes (LTMLM, GMMAT) avec analyse de données EPE comme trait binaire. Nous utilisons une simulation pour estimer l'erreur de type 1 de toutes les approches dans les cas de confusion. Puisque les MML sont communément utilisées même pour les traits binaires, nous analysons également les données avec un MML. Ces travaux ont été effectués sous la supervision de Kelly Burkett.

Novel Biostatistical Methods
Nouvelles méthodes biostatistiques

[Wednesday May 29/mercredi 29 mai, 10:50-11:05]

Wendimagegn Alemayehu (University of Alberta) , **Cynthia Westerhout** (University of Alberta)

On Statistical Modeling of the Relationship of Temporal Change of Biomarkers with Clinical Outcomes

La modélisation statistique de la relation du changement temporel des biomarqueurs au moyen de résultats cliniques.

The role of biomarkers both as diagnostic and prognostic tools for risk assessment of a disease is emerging. Most experience has focused on relating biomarkers measured at one time point with subsequent outcomes. However, there is an evolving interest in the effect of temporal change in biomarkers over time. In such cases, the statistical challenge is to model the longitudinally measured predictor variable against an outcome variable. The most common practice in the literature can be generalized as a two-stage model in which a measure of the change is first estimated and then used as a predictor variable to be modeled in the second step. Here we propose an alternative modeling approach that integrates the two stages and primarily estimates the association between underlying temporal change and the outcome. Using simulations and application to real data, we demonstrate that our approach results in more consistent and efficient estimates of the measure of variable associations.

Le rôle des biomarqueurs à la fois en tant qu'outils de diagnostic et de pronostic pour évaluer les risques de contracter une maladie est en voie d'émergence. La plupart des expériences se sont concentrées à faire le lien entre les biomarqueurs mesurés à un point dans le temps et les résultats subséquents. Cependant, l'effet du changement temporel chez les biomarqueurs au fil du temps attire de plus en plus l'attention. Dans de tels cas, le défi statistique consiste à modéliser la variable prédictive mesurée longitudinalement par rapport à une variable de résultat. La pratique la plus couramment rencontrée dans la documentation est généralement représentée par un modèle en deux étapes : on estime tout d'abord la mesure d'un changement, puis on s'en sert ensuite en guise de variable prédictive qui devra être modélisée à la deuxième étape. Nous proposons ici une approche de modélisation alternative qui intègre les deux étapes et estime essentiellement l'association entre le changement temporel sous-jacent et le résultat. Au moyen de simulations et de son application sur des données réelles, nous démontrons que notre approche procure des estimations plus convergentes et efficaces de la mesure de l'association des variables.

[Wednesday May 29/mercredi 29 mai, 11:05-11:20]

Changchang Xu (University of Toronto) , **Shelley Bull** (University of Toronto) , **Shelley Bull** (University of Toronto)

Improving Mixture Cure Modelling of Molecular Genetic Biomarkers in Cancer Prognosis by Penalized Maximum Likelihood

Amélioration de la modélisation de traitement par mélange des biomarqueurs génétiques moléculaires dans le prognostic du cancer par le maximum de vraisemblance pénalisé

When a study sample includes a large proportion of long-term survivors, mixture cure (MC) models that separately assess biomarker associations with long-term recurrence-free survival and time to disease recurrence are preferred to proportional hazards models. Standard maximum likelihood estimation (MLE) may be biased in small or sparse samples (i.e. with few recurrences). We extend Firth-type penalized likelihood estimation (PLE) developed for bias reduction in the exponential family to the Weibull-logistic MC, using the Jeffreys invariant prior. Via simulation studies, we evaluate PLE, as well as type 1 error and power obtained using Wald-type and likelihood ratio statistics, in comparison to MLE. In samples with few events, the MC-PLEs had mean bias closer to zero and smaller mean squared error than MC-MLEs, and could be obtained in samples when the MLEs are infinite. We illustrate the practical

Lorsque l'échantillon pour une étude comprend une grande proportion de survivants à long terme, l'utilisation de modèles de traitement par mélange (TM) qui évaluent séparément les associations de biomarqueurs avec la survie sans récurrence à long terme et avec le temps écoulé avant la récurrence est préférée à celle de modèles à risques proportionnels. L'estimateur du maximum de vraisemblance (EMV) normatif peut être biaisé dans le cas d'échantillons petits ou épars (c.-à-d. avec peu de récurrences). Nous élargissons l'estimation de vraisemblance pénalisée (EVP) de type Firth développée pour réduire le biais dans la famille exponentielle du TM logistique Weibull, à l'aide de la distribution a priori invariante de Jeffreys. Par des études en simulation, nous évaluons l'EVP, l'erreur de type 1 et la puissance obtenue en utilisant la statistique de type Wald et celle du ratio de vraisemblance, en comparaison avec l'EMV. Dans les échantillons avec relativement peu d'événements, en plus d'avoir un biais moyen plus près de zéro et une plus petite erreur quadratique moyenne que les TM-

utility of the methodology in a breast cancer cohort with long-term follow-up.

EMV, les TM-EVP pouvaient être obtenu avec des échantillons dans lesquels les EMV étaient infinis. Nous illustrons l'utilité pratique de la méthodologie dans une cohorte de sujets atteints de cancer du sein avec un suivi à long terme.

[Wednesday May 29/mercredi 29 mai, 11:20-11:35]

Rajib Dey (McGill University) , **Paramita Saha Chaudhuri** (McGill University)

Estimation of Time-Dependent Predictive Accuracy in the Presence of Competing Risks

Estimation de l'exactitude prédictive dépendante du temps en présence de risques concurrents

Evaluating a candidate biomarker or developing a predictive model score for event-time outcomes is frequently an important clinical goal. However, model development and assessment may be complicated in the presence of competing risks. The time-dependent incident/dynamic cause-specific (CS) area under the ROC curve (AUC) proposed by Saha and Heagerty (2010, Biometrics 66, 999-1011) is an appealing semiparametric measure to capture the predictive performance of a biomarker and incorporate competing risks. We propose a local and a global non-parametric estimator for the time-dependent CS AUC from censored survival data in the presence of competing risks. The first proposed estimator is a local average of time-dependent CS AUCs. The second estimator is based on modelling the CS AUC as a function of t through fractional polynomials. We investigated the performance of the proposed estimators through both simulation and real-life data analysis.

Évaluer un biomarqueur potentiel ou élaborer un score de modèle prédictif pour des données de survie sont des objectifs cliniques fréquents. Par contre, l'élaboration et l'évaluation du modèle peuvent être compliquées en présence de risques concurrents. L'aire sous la courbe ROC (AUC) par cause (CS) incident/dynamique dépendante du temps proposée par Saha et Heagerty (2010, Biometrics 66, 999-1011) est une mesure semiparamétrique intéressante pour saisir la performance prédictive d'un biomarqueur et pour incorporer les risques concurrents. Nous proposons un estimateur non paramétrique local et un global pour la CS AUC provenant de données de survie censurées en présence de risques concurrents. Le premier estimateur est une moyenne locale des CS AUCs dépendantes du temps. Le deuxième estimateur se base sur la CS AUC en fonction de t par des polynômes fractionnaires. Nous examinons la performance des estimateurs proposés par l'analyse de données réelles et de simulation.

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

Chair/Président: Daniel Zi Yang

Room/Salle: 119 (SA)

Abstract/Résumé

[Wednesday May 29/mercredi 29 mai, 10:20-10:35]

Mili Roy (University of Calgary)

Joint Analysis of Correlated Non-Gaussian Continuous Outcomes: Impact of Conditional Dependence

Analyse conjointe de résultats continus non gaussiens corrélés : impact de la dépendance conditionnelle

Generalized linear mixed models (GLMM) are well suited for joint analysis of multiple correlated non-Gaussian continuous outcomes. However, conventional GLMMs rely on the assumption of conditionally independent outcomes, given subject-specific random effects. In this talk, we adopt the class of Gaussian copula mixed models (GCMMs) to investigate the impact of incorrectly assuming the outcomes are conditionally independent when they are not. GCMMs generalize conventional GLMMs to disparate non-Gaussian outcomes and provide a flexible way of incorporating conditional dependence among them. We evaluate the finite-sample performance of maximum likelihood estimates for GCMMs empirically via simulations vis-à-vis those obtained from 'naive' GLMMs based on conditional independence. Our results show that the 'naive' analysis tends to yield estimates with severe bias and incorrect SEs. Our analysis of comet assay data highlights this inadequacy and illustrates the flexibility of GCMMs.

Les modèles mixtes linéaires généralisés (MMLG) sont bien adaptés à l'analyse conjointe de multiples résultats continus non gaussiens corrélés. Cependant, les MMLGs conventionnels reposent sur l'hypothèse de résultats conditionnellement indépendants, sachant les effets aléatoires spécifiques au sujet. Dans cette présentation, nous adoptons la classe des modèles mixtes de copules gaussiennes (MMCG) pour étudier l'impact de présumer à tort que les résultats sont conditionnellement indépendants quand ils ne le sont pas. Les MMCGs généralisent les MMLGs conventionnels à des résultats non gaussiens disparates et permettent d'y intégrer une dépendance conditionnelle en toute souplesse. Nous évaluons empiriquement la performance des estimations du maximum de vraisemblance des MMCGs en échantillon fini via des simulations, par rapport à celle des MMLGs « naïfs » basés sur une indépendance conditionnelle. Nos résultats montrent que l'analyse « naïve » a tendance à produire des estimations très biaisées et des erreurs-types incorrectes. Notre analyse de données de test des comètes souligne cette insuffisance et illustre la souplesse des MMCGs.

[Wednesday May 29/mercredi 29 mai, 10:35-10:50]

James G. MacKinnon (Queen's University) , **Morten O. Nielsen** (Queen's University) , **Matthew D. Webb** (Carleton University)

Wild Bootstrap Inference with Multiway Clustering

Inférence bootstrap sauvage avec regroupements multivoies

We study cluster-robust inference for regression models with clustering in two dimensions. Two different cluster-robust variance estimators (CRVEs) are shown to be consistent, but one of them requires stronger conditions than the other. We then propose several wild bootstrap procedures and prove that they are asymptotically valid under conditions that differ for the two CRVEs. For one of the CRVEs, the bootstrap can fail when there is actually no clustering in one or both dimensions. Simulations suggest that bootstrap inference based on the other CRVE is often much more accu-

Nous étudions l'inférence à regroupement robuste pour les modèles de régression avec regroupement en deux dimensions. Deux différents estimateurs de variance à regroupement robuste (EVRRs) ont été démontrés comme étant convergent mais l'un d'entre eux nécessite des conditions plus fortes que l'autre. Nous proposons plusieurs procédures bootstrap sauvages et nous démontrons qu'elles sont asymptotiquement valides sous des conditions différentes que celles pour les deux EVRRs. Pour une des EVRRs, le bootstrap peut échouer quand il n'y a aucun regroupement dans une ou dans les deux dimensions. Des simulations indiquent que l'inférence bootstrap fondée sur l'autre EVRR est souvent beau-

rate than inference based on the t distribution, especially when there are few clusters in at least one dimension. An empirical example confirms that bootstrap inferences can differ substantially from conventional ones.

coup plus précise que l'inférence fondée sur la distribution t, particulièrement quand il y a peu de regroupements dans au moins une dimension. Un exemple empirique confirme que les inférences bootsrap peuvent être considérablement différentes des inférences conventionnelles.

[Wednesday May 29/mercredi 29 mai, 10:50-11:05]

Ismaila Ba (Université du Québec à Montréal) , **Jean François Coeurjolly** (Université du Québec à Montréal)

Regularization Techniques for Inhomogeneous Gibbs Point Process Models with a Diverging Number of Covariates

Méthodes de régularisation pour des processus ponctuels de Gibbs inhomogènes avec un nombre divergent de covariables

The Gibbs point processes (GPP) constitute a large class of point processes with interaction between points. Depending on the relative distance between the points, this interaction can be attractive or repulsive. Feature selection procedures are an important topic in high-dimensional statistical modeling. In this talk, pseudo-likelihood approach regularized with convex and non-convex penalty functions is proposed to handle this kind of problem for inhomogeneous GPP. We particularly investigate the setting where the number of covariates diverges as the domain of observation increases. Furthermore, under some conditions provided on the spatial GPP and on the penalty functions, we show that the oracle property, the consistency and the asymptotic normality hold. Through simulation experiments, we validate our theoretical results and finally, an application to tropical forestry datasets illustrates the use of the proposed approach.

Les processus ponctuels de Gibbs (GPP) constituent une large classe de processus ponctuels avec une interaction entre les points. Dépendamment de la position relative entre les points, cette interaction peut générer de l'attraction ou de la répulsion. Les procédures de sélection de variables sont une thématique importante en modélisation statistique en grande dimension. Dans cet exposé, une approche de pseudo-vraisemblance avec des fonctions de pénalité convexes et non convexes est proposée pour traiter ce type de problèmes pour les GPP inhomogènes. Nous étudions particulièrement le cas où le nombre de covariables diverge quand la fenêtre d'observation s'aggrandit. De plus, sous certaines conditions fournies sur le GPP spatial et les fonctions de pénalité, nous démontrons une propriété d'oracle, la consistance et la normalité asymptotique pour nos estimateurs. À travers des expériences de simulation, nous validons nos résultats théoriques et, enfin, une application aux jeux de données de forêt tropicale illustre l'utilisation de l'approche proposée.

[Wednesday May 29/mercredi 29 mai, 11:05-11:20]

Saumen Mandal (University of Manitoba)

Optimal Designs Subject to Achieving Equality of Variances of the Estimators of Linear Functions of Parameters

Conceptions optimales soumises à l'atteinte de l'égalité des variances des estimateurs de fonctions linéaires des paramètres

In some regression models, it may be of interest to construct designs subject to achieving the equality of variances of the estimators of some parameters or parametric functions of interest. Motivated by this fact, we focus on the problem of finding constrained approximate designs by optimizing a criterion subject to constraints. The constraints are based on the equality of variances of the estimators of the linear parametric functions of interest. We approach this problem by initially formulating the Lagrangian function with the constraints. Invoking the general equivalence theorem and making the directional derivatives of the Lagrangian function to zero, we create a compound optimization criterion and solve the problem by means of a simultaneous optimization tech-

Dans certains modèles de régression, il pourrait être intéressant de construire des conceptions soumises à l'atteinte de l'égalité des variances des estimateurs de certains paramètres ou fonctions paramétriques d'intérêt. Motivés par cette occurrence, nous nous concentrons à trouver des conceptions approximatives restreintes en optimisant un critère soumis à des contraintes. Les contraintes sont basées sur l'égalité des variances des estimateurs des fonctions paramétriques d'intérêt. Nous abordons ce problème en formulant d'abord la fonction lagrangienne avec les contraintes. Puis en invoquant le théorème d'équivalence générale et en fixant les dérivées directionnelles de la fonction lagrangienne à zéro, nous créons un critère d'optimisation composé et résolvons le problème au moyen d'une technique d'optimisation simultanée. Les approches sont formulées pour un modèle de régression général et

nique. The approaches are formulated for a general regression model and are explored through some examples.

sont examinées dans certains exemples.

[Wednesday May 29/mercredi 29 mai, 11:20-11:35]

Harlan Campbell (University of British Columbia) , **Daniel Lakens** (Eindhoven University of Technology)

Can We Disregard the Whole Model? Non-Inferiority Testing for Omnibus Effects in Linear Models

Peut-on ne pas prendre en compte le modèle entier ? Tests de non infériorité des effets omnibus dans les modèles linéaires

In a multivariate setting, showing evidence "in favour of the null" can be challenging. Indeed, frequentist tests for this task are rarely suggested or cited in the literature. In this talk, we introduce two novel non-inferiority tests to be used in ANOVA and linear regression analyses. These tests correspond to standard F-tests for eta-squared and R-squared, and, with an appropriate non-inferiority margin, can be used to determine if the variance explained by the explanatory variables is at most negligible. We compare the operating characteristics of these tests with popular Bayesian testing schemes and conduct a small simulation study to study type 1 error and power.

Dans une configuration multivariée, il peut être difficile d'établir une preuve « en faveur de l'hypothèse nulle ». En fait, les tests fréquentistes à cette fin sont rarement suggérés ou mentionnés dans la littérature. Nous présentons ici deux nouveaux tests de non infériorité à être utilisés dans des analyses de la variance (ANOVA) et de régression linéaire. Ces tests correspondent à des tests F normatifs pour eta carré et R carré, et, avec une marge de non infériorité appropriée, ils peuvent être utilisés pour déterminer si la variance expliquée par les variables explicatives est tout au plus négligeable. Nous comparons les caractéristiques de fonctionnement de ces tests avec les tests bayésiens populaires et menons une étude de simulation à petite échelle pour en examiner l'erreur de première espèce et la puissance.

Best Practices in Experiential Learning
Pratiques exemplaires en apprentissage expérientiel

Chair/Président: Sohee Kang

Organizer/Responsable: Sohee Kang

Room/Salle: 119 (SA)

Abstract/Résumé

[Wednesday May 29/mercredi 29 mai, 13:30-14:00]

Shirley Mills (Carleton University)

Experiential Learning via Co-ops and Internships

L'apprentissage expérientiel par les coopératives et les stages

Since 1979, the Statistics Honours and Graduate programs at Carleton University have been offering students the opportunity to gain workplace experience via co-ops and internships. In this session we examine recruitment and retention implications, outline requirements of these options, discuss the placement process and the breadth of work undertaken by students, and outline implications for classroom teaching and student and faculty research.

Depuis 1979, les programmes de spécialisation et les études de cycle supérieur en statistique de l'Université de Carleton offrent aux étudiants la possibilité d'acquérir une expérience professionnelle par le biais de coopératives et de stages. Au cours de cette session, nous examinons les implications pour le recrutement et la rétention, définissons les exigences de ces options, discutons du processus de placement et de l'ampleur des travaux entrepris par les étudiants, ainsi que des implications pour l'enseignement en classe et pour la recherche effectuée par les étudiants et les professeurs.

[Wednesday May 29/mercredi 29 mai, 14:00-14:30]

Albert Y. Kim (Smith College)

Moderndive: Statistical Inference Using the Tidyverse

Moderndive : l'inférence statistique avec le tidyverse

We propose an approach to teaching statistical inference using tidyverse data science tools in the R statistical language. We argue that this tidyverse-centric approach is both 1) feasible since tidyverse is more intuitive for new R coders to learn than base R and 2) valuable since on our proposed path to statistical inference, students will also learn to use data science tools applicable beyond the classroom. Our approach involves teaching "just enough" data visualization wrangling with the ggplot2 dplyr packages to equip students with the necessary computational tools for the journey. Using these tools, students can then learn both explanatory and predictive regression modeling and the infer package, which is a new R package that makes statistical inference tidy and transparent. A combination of pen/paper exercises and tactile sampling/resampling simulations also help keep students engaged.

On propose une approche pour l'enseignement de l'inférence statistique en utilisant des outils de la science des données ≪tidyverse≫ avec le langage statistique R. On propose que cette approche soit aussi bien 1) faisable car le tidyverse est plus facile à apprendre que le ≪base R≫ pour les programmeurs débutants en R, que 2) profitable car sur notre parcours proposé vers l'inférence statistique, les étudiants vont aussi apprendre à utiliser des outils de la science des données valables au-delà de la classe. Notre approche consiste à enseigner juste assez de visualisation et de transformation de données avec les bibliothèques R ggplot2 et dplyr pour équiper les étudiants avec les outils informatiques nécessaires pour leur apprentissage. Avec ces outils, les étudiants seront en mesure d'apprendre des modèles de régression d'explication et de prédiction, ainsi que la bibliothèque R de l'inférence statistique. Cette dernière est une nouvelle bibliothèque R qui rend l'inférence statistique rangée et transparente. Des exercices sur papier et des simulations tactiles d'échantillonnage ou de ré-échantillonnage vont aussi nous aider à garder les étudiants intéressés.

[Wednesday May 29/mercredi 29 mai, 14:30-15:00]

Nathan A. Taback (University of Toronto)

Best Practices in Experiential Learning
Pratiques exemplaires en apprentissage expérientiel

ASA DataFest@UofT and Beyond
ASA DataFest@UofT, et plus encore

ASA DataFest is like a hackathon for undergraduate students, except the problem is a data science problem, rather than a programming problem. Teams of students get a dataset on day 1 and work on the problem until day 2 where they present their results. After two days of intense data wrangling, analysis, and presentation design, each team is allowed a few minutes and no more than two slides to impress a panel of judges. Prizes are given for various categories. ASA DataFest brings together the data science community. Undergraduate students do the work, but they are assisted by roving consultants who are graduate students, faculty, and industry professionals. The event provides an experiential learning opportunity for a large number of students. I will discuss my experiences as a faculty who has organized ASA DataFest@UofT, and offer practical advice for faculty who are contemplating organizing a data science competition.

ASA DataFest est comme un marathon de programmation pour les étudiants de premier cycle, sauf que le problème à résoudre est relié à la science des données plutôt qu'à la programmation. Des équipes d'étudiants reçoivent un jeu de données au jour 1 et doivent tenter de résoudre le problème avant le jour 2, durant lequel ils devront présenter leurs résultats. Après deux jours intensifs de préparation préalable des données, d'analyses et de design de présentation, chaque équipe a droit à quelques minutes et à pas plus de deux diapositives pour impressionner le jury. Des prix sont décernés pour plusieurs catégories. ASA DataFest permet à la communauté en sciences des données de se rassembler. Bien que des étudiants de premier cycle soient responsables d'accomplir le travail, ils sont tout de même soutenus par des conseillers itinérants composés de diplômés, du corps enseignant et de professionnels de l'industrie. L'événement offre à un grand nombre d'étudiants une occasion de vivre une expérience d'apprentissage. Je témoignerai de mon expérience en tant qu'enseignant ayant contribué à organiser ASA DataFest@UofT et offrirai des conseils pratiques aux enseignants qui songeraient à organiser une compétition en science des données.

Isobel Loutit Lecture
Allocution Isobel Loutit

Chair/Président: Chunfang Lin

Organizer/Responsable: Chunfang Lin

Room/Salle: 102 (ICT)

Abstract/Résumé

[Wednesday May 29/mercredi 29 mai, 13:30-14:35]

Max D. Morris (Iowa State University)

A Brief History of Statistical Computer Experiments

Bref historique des expériences informatiques en statistique

For several decades, statisticians and researchers from other professions have discussed how output generated from computer models can be usefully analyzed, sometimes along with data from other sources. In this talk, I'll review some of the efforts in the area statisticians broadly call "computer experiments". I'll offer some examples of research and applications, but the focus will be on the big picture of how this work fits into modern statistics than with detailed technical descriptions. While it predates my review, it is interesting to note that Isobel Loutit's World War II-era work with Northern Electric involved the validation – a common point of concern with modern computer models – of analogue calculations of ballistic trajectories. As we rely more and more on formal computational models to express what we know, analysis aimed at understanding and predicting what we do not know will require continued development of methodology that takes advantage of these models.

Depuis plusieurs décennies, statisticiens et chercheurs d'autres disciplines discutent de comment analyser les résultats de modèles informatiques, combinés ou non à des données provenant d'autres sources. Dans cette présentation, je passerai en revue certains efforts dans ce domaine que les statisticiens appellent au sens large « expériences informatiques ». Je donnerai des exemples de recherches et de leurs applications, mais j'insisterai surtout sur la façon dont ces travaux s'intègrent dans la statistique moderne, sans m'attarder sur le détail technique. Bien qu'antérieurs à cette étude, il est intéressant de noter que les travaux qu'a effectué Isobel Loutit avec Northern Electric pendant la Seconde Guerre Mondiale portaient sur la validation – sujet de préoccupation commun dans les modèles informatiques modernes – de calculs analogiques de trajectoires balistiques. Aujourd'hui, alors que nous utilisons de plus en plus des modèles computationnels formels pour exprimer ce que nous savons, il nous faut, pour analyser, comprendre et prédire ce que nous ne savons pas, continuer à élaborer des méthodes qui exploitent ces modèles.

Data Integration and Distributed Inference
Intégration de données et inférence distribuée

Chair/Président: Xikui Wang

Organizer/Responsable: Xikui Wang

Room/Salle: 122 (ICT)

Abstract/Résumé

[Wednesday May 29/mercredi 29 mai, 13:30-14:00]

Peter X Song (University of Michigan)

Integrative Data Analytics via Distributed Inference Functions

Analyse intégrative des données au moyen des fonctions d'inférence distribuée

This talk concerns integrative data analytics and distributed inference in data integration. As data sharing from related studies become of interest, statistical methods for a joint analysis of all available datasets are needed in practice to achieve better statistical power and detect signals that are otherwise impossible based on a single dataset alone. A major challenge arising from integrative data analytics pertains to principles of information aggregation, learning data heterogeneity, inference and algorithms for model fusion. Information aggregation has been studied extensively by many statistics pioneers, which lay down the foundation of data integration. In this process, it is of critical importance to accommodate data heterogeneity, otherwise the analysis will result in biased estimation and misleading inference. Distributed inference function and divide-and-conquer algorithms will be discussed with theoretical arguments and numerical examples.

Cet exposé porte sur l'analyse intégrative des données et l'inférence répartie dans l'intégration des données. Lorsque l'échange de données issues d'études connexes devient intéressant, des méthodes statistiques pour une analyse conjointe de tous les ensembles de données disponibles sont nécessaires dans la pratique pour obtenir une meilleure puissance statistique et détecter des signaux qui seraient autrement impossibles sur la seule base d'un seul ensemble de données. L'un des principaux défis que pose l'analyse intégrative des données concerne les principes d'agrégation de l'information, l'hétérogénéité des données d'apprentissage, l'inférence et les algorithmes de fusion des modèles. De nombreux pionniers de la statistique, qui ont jeté les bases de l'intégration des données, ont étudié en profondeur l'agrégation de l'information. Dans ce processus, il est crucial de tenir compte de l'hétérogénéité des données, sinon l'analyse donnerait lieu à des estimations biaisées et à une inférence trompeuse. Nous discuterons de la fonction d'inférence répartie et des algorithmes diviser pour régner avec des arguments théoriques et des exemples numériques.

[Wednesday May 29/mercredi 29 mai, 14:00-14:30]

Su Chen (University of Memphis) , **Wilfried Karmaus** (University of Memphis)

A Nonparametric Test of Variance Heterogeneity in DNA Methylation Influenced by Genetic Variants

Un test non paramétrique de l'hétérogénéité de la variance dans une méthylation de l'ADN influencée par des variants génétiques

A number of studies had shown genetic variants influence the mean value of various quantitative traits in human. Recently, a few studies reveal that the genetic variants are also associated with the variance of different complex human traits, including intermediate phenotypes such as DNA methylation (DNAm). The variance heterogeneity in DNAm across different genotypes may help explain individual susceptibility to environmental exposure. To accommodate the non- normal DNA methylation data, we propose a nonparametric test of equality invariability using the kernel density func-

Plusieurs études ont démontré que les variants génétiques influencent la valeur moyenne de plusieurs traits quantitatifs chez l'humain. Récemment, quelques études révèlent que les variants génétiques sont aussi associés à la variance de différents traits humains complexes, y compris des phénotypes intermédiaires comme la méthylation de l'ADN (mADN). L'hétérogénéité de la variance de la mADN dans différents génotypes pourrait contribuer à expliquer la prédisposition d'un individu à l'exposition environnementale. Pour gérer les données anormales de la méthylation de l'ADN, nous proposons un test non paramétrique de l'invariabilité de l'égalité au moyen de l'estimation fonction-

tional estimate of scale parameters. Our proposed test outperforms existing nonparametric test of equality of variances such as the Fligner-Killeen test, and bootstrapped Levene's test in intensive simulation studies. We applied our proposed method to test the variance heterogeneity in DNAm associated with genetic variants with and without adjusting for relevant covariates.

nelle par noyau des paramètres d'échelle. Le test que nous proposons est plus performant dans des études en simulation intensive que les tests de l'égalité des variances actuels comme le test de Fligner-Killeen et le test de Levene par bootstrap. Nous avons appliqué notre méthode proposée au test d'hétérogénéité de la variance de la mADN associée à des variants génétiques avec et sans ajustement des covariables pertinentes.

[Wednesday May 29/mercredi 29 mai, 14:30-15:00]

You Liang (Ryerson University) , **Xikui Wang** (University of Manitoba) , **Lysa Porth** (University of Manitoba)

Risk Management for Heavy Tailed and Tail Dependent Claims

Gestion du risque pour les réclamations à queue lourde et à dépendance de queue

Risk management is an important issue in the insurance industry and financial institutions. The primary goal is to model, measure and manage risk in a variety of settings. One particular challenge is the aggregation of dependent claims in the insurance industry, especially for the property and casualty sector. We will focus on exploring the asymptotic behaviour of the tail probability and risk measure for aggregate and tail dependent claims from heavy tailed distributions. The Archimedean survival copula is used to model tail dependency. Simulation of the Archimedean survival copula and estimation of value at risk and conditional value at risk will be emphasized.

La gestion du risque est un problème important dans le secteur de l'assurance et des institutions financières. L'objectif premier consiste à modéliser, mesurer et gérer le risque selon des paramètres très diversifiés. Un problème particulier est celui de l'agrégation de réclamations de personnes à charge dans l'industrie de l'assurance, plus précisément dans le secteur des biens et accidents. Nous nous appliquons à explorer le comportement asymptotique de la queue de la loi de probabilité et la mesure du risque pour les réclamations agrégées et à dépendance de queue selon des distributions à queue lourde. Utilisée pour modéliser la dépendance de queue, la copule archimédienne de survie fera l'objet d'une simulation que nous mettrons en lumière de même que l'estimation de la valeur et de la valeur conditionnelle à risque.

Recent progress for quantile regression analysis
Récents progrès en analyse par régression quantile

Chair/Président: Dianliang Deng

Organizer/Responsable: Dianliang Deng

Room/Salle: 109 (SS)

Abstract/Résumé

[Wednesday May 29/mercredi 29 mai, 13:30-14:00]

Dianliang Deng (University of Regina) , **Mashfiqul Chowdhury** (University of Regina) , **Mashfiqul Chowdhury** (University of Regina)

Quantile Regression Analysis for Gene Expression Data

Analyse de régression quantile pour données d'expression génique

Temporal gene expression data contains ample information to characterize gene function and is now widely used in biomedical research. A dense temporal gene expression usually shows various patterns in expression levels under different biological conditions. Existing literature models the gene trajectory using the mean function. Moreover, temporal gene expression curves generally show a strong degree of heterogeneity between multiple conditions. As a result, the rate of change of gene expressions may be different in non-central locations and a mean model can not explore the non-central location of the distribution. We will discuss the linear quantile mixed model to analyze gene expression data. Based on the estimated quantile regression parameters, the statistical performance of proposed method is investigated using a simulation study. Further, the proposed method is used to analyze a dataset of 18 genes in P. aeruginosa, expressed in 24 biological conditions.

Les données d'expression génique temporelle, qui contiennent une abondance d'informations permettant de caractériser la fonction génique, sont désormais largement utilisées en recherche biomédicale. Une expression génique temporelle dense présente généralement des profils différents selon les conditions biologiques. Les publications existantes modélisent la trajectoire d'un gène en utilisant la fonction moyenne. De plus, les courbes d'expression génique temporelle montrent généralement une forte hétérogénéité selon les conditions. Par conséquent, le rythme de variation de l'expression génique peut être différent dans les emplacements loin du centre, or un modèle de moyenne ne peut pas explorer ces parties non centrales de la distribution. Dans cette présentation, nous discuterons d'un modèle mixte à quantiles linéaires pour l'analyse des données d'expression génique. Modèle fondé sur des paramètres de régression quantile estimée, nous en étudions la performance statistique via une étude de simulation. Puis nous utilisons la méthode proposée pour analyser un jeu de données de 18 gènes de P. aeruginosa, exprimés dans 24 conditions biologiques.

[Wednesday May 29/mercredi 29 mai, 14:00-14:30]

Mei Ling Huang (Brock University) , **Jenny Tieu** (Brock University)

A Nonparametric Quantile Regression Method

Méthode de régression quantile non paramétrique

Quantile regression (QR) estimates conditional quantiles and has wide applications in the real world. Estimating high conditional quantiles is an important problem. The regular QR method often sets a linear or nonlinear model, then estimates the coefficients to obtain the estimated conditional quantile. This approach may be restricted by the model setting. To overcome this problem, this paper proposes a direct nonparametric QR method. The asymptotic properties of this direct estimator are given. Monte Carlo simulations show good

La régression quantile estime les quantiles conditionnels et a de vastes applications dans le monde réel. L'estimation de quantiles conditionnels élevés est un problème important. La méthode de régression quantile usuelle établit souvent un modèle linéaire ou non linéaire, puis estime les coefficients pour obtenir le quantile conditionnel estimé. Cette approche peut être restreinte par le paramétrage du modèle. Pour surmonter ce problème, cet article propose une méthode directe non paramétrique de régression quantile. On donne les propriétés asymptotiques de cet estimateur direct. Les simulations de Monte Carlo montrent une bonne efficacité de

efficiency for the proposed direct nonparametric QR estimator relative to the regular QR estimator. The paper also investigates two real-world examples of applications by using the proposed method. Comparisons of the proposed method and existing methods are given.

l'estimateur direct non paramétrique de la régression quantile proposée par rapport à l'estimateur courant de la régression quantile. Dans le cadre de cet article, nous examinons également deux exemples concrets d'applications à l'aide de la méthode proposée, et nous présentons des comparaisons entre la méthode proposée et les méthodes existantes.

[Wednesday May 29/mercredi 29 mai, 14:30-15:00]

Mohammad Jafari Jozani (University of Manitoba)

More Efficient Quantile Regression Analysis Using Rank Information

Analyse plus efficace de la régression quantile en utilisant l'information sur le rang

In many medical applications, environmental studies. and ecological studies, it is easy to obtain observations that carry rank information. The rank information can be obtained using expert knowledge or easy-to-access covariates. In this talk, I will study the problem of quantile regression analysis using such data in order to estimate the conditional median or other quantiles of the response variable. I will mostly focus on estimating quantile regression using median ranked set samples, independent medians obtained form repeated sampling of the underlying population using relatively small samples. I introduce a new check function that can be used to incorporate the extra rank information and show for what ranges of quantiles the estimated quantile regression using median ranked set samples is more efficient than the one under simple random samples. Results will be evaluated by simulation studies and a real data application.

Dans plusieurs applications médicales, études environnementales et études en écologie, il est facile d'obtenir des observations qui comportent de l'information sur le rang. L'information sur le rang peut être obtenue en utilisant des connaissances provenant d'experts ou des covariables facilement accessibles. Dans cet exposé, j'étudierai le problème de l'analyse de la régression quantile en utilisant ce type de données pour estimer la médiane conditionnelle ou d'autres quantiles de la variable de réponse. Je me concentrerai principalement sur l'estimation de la régression quantile à l'aide de médianes d'échantillons ordonnés et des médianes indépendantes provenant d'échantillonnages répétés de la population sous-jacente en utilisant des échantillons relativement petits. Je présenterai une nouvelle fonction check qui peut être utilisée pour inclure l'information supplémentaire sur le rang et j'indiquerai pour quelles classes de quantiles la régression quantile estimée utilisant les médianes d'échantillons ordonnés est plus efficace que celle avec de simples échantillonnages aléatoires. Les résultats seront évalués à l'aide d'études de simulation et d'une application sur des données réelles.

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

Chair/Président: Gordon H Fick

Room/Salle: 144 (SB)

Abstract/Résumé

[Wednesday May 29/mercredi 29 mai, 13:30-14:35]

Peijun Sang (University of Waterloo)

Sparse Estimation for Functional Semiparametric Additive Models

Estimation éparse pour des modèles additifs semi-paramétriques fonctionnels

We propose a functional semiparametric additive model for the effects of a functional covariate and several scalar covariates and a scalar response. The effect of the functional covariate is modelled nonparametrically, while a linear form is adopted to model the effects of the scalar covariates. This strategy can enhance flexibility in modelling the effect of the functional covariate and maintain interpretability for the effects of scalar covariates simultaneously. We develop the method for estimating the functional semiparametric additive model by smoothing and selecting non-vanishing components for the functional covariate. Asymptotic properties of our method are also established. Two simulation studies are implemented to compare our method with various conventional methods. We demonstrate our method with two real applications.

Nous proposons un modèle additif semi-paramétrique fonctionnel pour les effets d'une seule covariable fonctionnelle, plusieurs covariables scalaires et une réponse scalaire. La modélisation de la covariable fonctionnelle est non paramétrique, tandis que nous adoptons une forme linéaire pour modéliser les effets des covariables scalaires. Cette stratégie peut rendre plus flexible la modélisation de l'effet de la covariable fonctionnelle, tout en maintenant simultanément l'interprétabilité des effets des covariables scalaires. Nous élaborons une méthode d'estimation d'un modèle additif semi-paramétrique fonctionnel afin de lisser et sélectionner les composantes qui demeurent pour la covariable fonctionnelle. Les propriétés asymptotiques de notre méthode sont aussi établies. Deux études de simulation sont mises en œuvre afin de comparer notre méthode à diverses autres méthodes conventionnelles. Nous démontrons notre méthode à l'aide d'applications réelles.

Methods for High-Dimensional and Large Data II
Méthodes pour traiter les données volumineuses et de grande dimension II

Chair/Président: Mohammad Ehsanul Karim

Room/Salle: 201 (ENA)

Abstract/Résumé

[Wednesday May 29/mercredi 29 mai, 13:30-13:45]

Patrick Fournier (Université du Québec à Montréal) , **Fabrice Larribe** (Université du Québec à Montréal)

Modelling Horizontal Gene Transfer in Bacteria via Conditional Ancestral Recombination Graphs

Modélisation du transfert horizontal de gènes dans les bactéries via tableau de recombinaison ancestral conditionnel

Horizontal gene transfer is the process by which organisms exchange genetic material other than by reproduction. This phenomenon is especially common among bacteria. In consequence, inference procedures must take it into account in order to achieve validity. A variety of methodologies has been developed with that goal in mind, ranging from (almost literally) looking for unexpected similarities between sequences to more sophisticated likelihood-based ones like MCMC sampling. However, the advent of high-throughput sequencing, among other things, calls for the development of fast, flexible and accurate inference methods. Our proposition is to introduce a slight modification to the ancestral recombination graph that allows the construction of an importance sampler able to build genealogies consistent with a given sample of bacterial genomic sequences.

Le transfert horizontal de gènes est le processus par lequel des organismes échangent des matériels génétiques autrement que par la reproduction. Ce phénomène est particulièrement commun chez les bactéries. Par conséquent, les procédures d'inférence doivent en tenir compte pour être valides. Diverses méthodes ont été développées à ces fins, de la recherche (presque littérale) de similitudes inattendues entre séquences à des méthodes plus sophistiquées fondées sur la vraisemblance, comme l'échantillonnage MCCM. Cependant, l'introduction notamment du séquençage à très haut débit exige le développement de méthodes d'inférences rapides, souples et précises. Nous proposons l'introduction d'une légère modification du tableau de recombinaison ancestral qui permet la construction d'un échantillonneur d'importance capable de créer des généalogies compatibles à un échantillon donné de séquences génomiques d'une bactérie.

[Wednesday May 29/mercredi 29 mai, 13:45-14:00]

Yuming Zhang (University of Geneva) , **Stéphane Guerrier** (University of Geneva) , **Maria-Pia Victoria-Feser** (University of Geneva) , **Mucyo Karemera** (Pennsylvania State University) , **Samuel Orso** (University of Geneva)

A Study of Simulation Based Estimators for High Dimensional Generalized Linear Models

Une étude d'estimateurs basés sur la simulation pour des modèles linéaires généralisés de grande dimension

The large number of variables, sample size, and complexity of statistical models has resulted in computational challenges to obtain estimators with desirable properties. In order to derive numerically efficient estimators with satisfying statistical properties, we propose a simulation-based method, the Iterative Bootstrap (IB), based on Guerrier et al. (2019). It is shown that the IB can be used, very generally, to remove the finite sample bias of (nearly) consistent estimators, when the number of parameters is allowed to increase with the sample size. This method can be applied to a wide class of models and we consider here, in particular, generalized linear models, including logistic and negative binomial regression models. Applications show that we can

Le nombre croissant de variables, la taille de l'échantillon et la complexité des modèles statistiques a entraîné des difficultés computationnelles pour obtenir des estimateurs dotés des propriétés désirables. Afin de dériver des estimateurs numériquement efficaces avec des propriétés statistiques satisfaisantes, nous proposons une méthode basée sur la simulation, le bootstrap itératif (BI), d'après Guerrier et coll. (2019). On a montré la possibilité d'utiliser le BI, de façon très générale, pour éliminer le biais d'échantillon de taille finie des estimateurs (presque) consistent, lorsque le nombre de paramètres peut augmenter en même temps que la taille de l'échantillon. Cette méthode peut s'appliquer à une vaste catégorie de modèles linéaires généralisés, y compris la régression logistique et la régression binomiale négative. Avec les applications, on peut voir qu'il est possible d'obtenir une efficacité

gain finite sample efficiency compared to standard estimators and that the algorithm converges exponentially fast, making it a realistic choice for high-dimensional.

de l'échantillon de taille finie, comparativement aux estimateurs normatifs, et une convergence de l'algorithme exponentiellement rapide, ce qui le rend réaliste pour les échantillons de grande dimension.

[Wednesday May 29/mercredi 29 mai, 14:00-14:15]

Wei Tu (University of Alberta) , **Linglong Kong** (University of Alberta) , **Zhihua Su** (University of Florida) , **Rohana Karunamuni** (University of Alberta)

Envelope-Based High-Dimensional Gaussian Copula Regression

Régression de copule gaussienne de grande dimension avec une méthode d'enveloppes

Envelopes were recently proposed by Cook, Li and Chiaromonte (2010) as a method for reducing estimative and predictive variations in multivariate linear regression. Currently most envelope-based models are essentially linear models based on the Gaussian assumption. We propose a nonlinear envelope model using high-dimensional Gaussian copula regression (Can and Zhang 2015). Using a Kendall's tau-based covariance matrix estimator, estimation based on the new method provides significantly better performance under unknown monotone marginal transformation.A new algorithm based on a matrix-wise instead of row- or column- wise update of the target matrix is proposed, and it has been shown to be much faster and usually more accurate compared with the current used algorithms. The proposed method is easy to implement and highly adaptive to other envelope-based models. We demonstrate the effectiveness of the proposed method using numerical simulations and real data analysis.

Cook, Li et Chiaromonte (2010) ont récemment proposé une méthode d'enveloppes pour réduire les variations estimatives et prédictives dans les régressions linéaires multivariées. Pour l'instant, les modèles à base d'enveloppes sont essentiellement linéaires et fondés sur une hypothèse gaussienne. Nous proposons un modèle d'enveloppe non linéaire utilisant une régression de copule gaussienne de grande dimension (Can et Zhang 2015). À l'aide d'un estimateur de matrice de covariance basé sur le tau de Kendall, la performance de l'estimation avec cette nouvelle méthode est notablement meilleure en présence de transformations marginales monotones inconnues. Nous proposons un nouvel algorithme basé sur une mise à jour de la matrice cible par la matrice plutôt que par rangs et colonnes, algorithme dont on a montré qu'il est plus rapide et généralement plus exact que les algorithmes actuellement utilisés. La mise en œuvre de la méthode proposée est facile et s'adapte très bien à d'autres modèles basés sur les enveloppes. Nous illustrons l'efficacité de la méthode proposée à l'aide de simulations numériques et d'analyses avec des données réelles.

[Wednesday May 29/mercredi 29 mai, 14:15-14:30]

Sonja Surjanovic (University of British Columbia) , **William J. Welch** (University of British Columbia)

Gaussian Process Regression with Large Datasets

Régression par processus gaussien avec de grands ensembles de données

Computer models are used as surrogates for physical computer experiments in a large variety of applications. Nevertheless, the number of evaluations of the computer model is often limited due to the complexity and cost of the model. Historically, Gaussian process regression has proven to be the almost ubiquitous choice of statistical surrogate for such a computer model, due to its flexible form and analytical expressions for measures of predictive uncertainty. However, even this statistical surrogate can be computationally intractable for large designs, due to computing time increasing with the cube of the design size. Multiple methods have been proposed for addressing this problem. We discuss several of them, and compare their predictive and computational performance in

Les modèles informatiques sont utilisés comme substituts à des expériences informatiques physiques dans bon nombre d'applications très diverses. Pourtant, le nombre d'évaluations du modèle informatique est souvent limité en raison de la complexité et du coût du modèle. Avec le temps, la régression par processus gaussien a été confirmée comme substitut statistique le plus généralisé pour un tel modèle informatique, en raison de sa forme flexible et des expressions analytiques des mesures de l'incertitude prédictive. Par contre, même ce substitut statistique peut être difficilement solvable pour des plans d'expérience de grande dimension, en raison de l'accroissement du temps de calcul correspondant au cube de la dimension de l'expérience. Nous aborderons plusieurs des nombreuses méthodes mises de l'avant pour la résolution de ce problème, en comparant leur rendement prédictif

several scenarios. We then propose a new method for solving this problem using a sequential approach.

et computationnel selon divers scénarios. Nous proposons ensuite une nouvelle méthode pour résoudre ce problème à l'aide d'une approche séquentielle.

[Wednesday May 29/mercredi 29 mai, 14:30-14:45]

Gyanendra Pokharel (University of Calgary) , **Paula Robson** (Alberta Health Services) , **Lorriane Shack** (University of Calgary) , **John Spinelli** (BC Cancer Agency) , **Karen Kopciuk** (Alberta Health Services)

Dimensionality Reduction and Stage Shifting by Modifying Determinants of Cancer at Diagnosis

Réduction de la dimensionnalité et dépistage par étapes en modifiant les déterminants du cancer au moment du diagnostic

This project utilizes estimates of distributions of risk factors identified from real data, and models cancer stage in ordinal scale using proportional or partial proportional odds models expecting to identify factors that can catch cancer at early stages. However, there are several issues when dealing with a large number of categorical covariates. To avoid biased inference, we propose to use a mixed principal component analysis to reduce data dimension instead of replacing ordered categorical variables by a dummy matrix of dichotomized categories or treating as continuous variables. We also develop a regularization method that penalizes categorical variables. The effects of groups of risk factors on shifting cancer stage at diagnosis using simulated data with the proposed feature selection methods are being investigated in simulation studies. The results will inform cancer screening and prevention programs to target modifiable risk factors with the greatest impact.

Dans l'espoir d'identifier des facteurs susceptibles de dépister précocement le cancer, ce projet fait appel à des estimations de distributions des facteurs de risque identifiés à partir de données réelles et modélise les stades du cancer selon une échelle ordinale, en utilisant des modèles de probabilité proportionnelle et partiellement proportionnelle. Plusieurs difficultés se présentent toutefois lorsque nous traitons un nombre élevé de covariables catégorielles. Pour éviter l'inférence biaisée, nous proposons l'utilisation d'une analyse en composantes principales de données mixtes afin de réduire la dimension des données plutôt que de remplacer les variables catégorielles ordinales par une matrice binaire de catégories dichotomisées ou de les traiter comme des variables continues. Nous élaborons aussi une méthode de régularisation qui pénalise les variables catégorielles. Par des études en simulation, nous examinons les effets des groupes de facteurs de risque sur le dépistage par étapes du cancer au moment du diagnostic, à l'aide de données simulées avec les méthodes de sélection de caractéristiques qui sont proposées. Les résultats permettront que les programmes de dépistage et de prévention ciblent avec le plus grand impact les facteurs de risque modifiables.

[Wednesday May 29/mercredi 29 mai, 14:45-15:00]

Min Zhang (University of South China) , **Xuewen Lu** (University of Calgary)

Intelligent Search of Radiation Source and Its Optimization

Recherche intelligente des sources de rayonnement et son optimisation

In the high radioactive field, how to quickly search and clear the nuclear radioactive sources is very important for developing the strategy of nuclear emergency response. Intelligent searches of the radiation source by computer can avoid radiation damage to human body. Designing intelligent algorithms will be the key to fast searches for radioactive sources. In this paper, a new method is proposed to study the inverse reduction of nuclear radiation field by mathematical structural method. A probability optimization model of stochastic search imitates the typical search experience and is realized by computer. A combination of the random search with heuristic search makes the random search algorithm adaptable to the actual search in a variety of complex

Dans un champ hautement radioactif, il est très important de savoir comment trouver et éliminer rapidement les sources de rayonnement nucléaire pour élaborer une stratégie d'intervention en cas d'urgence nucléaire. Des recherches intelligentes par ordinateur des sources de rayonnement peuvent prévenir les radiolésions. La conception d'algorithmes intelligents jouera un rôle-clé pour accélérer les recherches de sources de rayonnement. Dans cet article, nous proposons une nouvelle méthode pour étudier la réduction inverse d'un champ de rayonnement nucléaire au moyen d'une méthode structurelle mathématique. Un modèle d'optimisation de probabilité pour une recherche stochastique imite l'expérience de recherche typique et est réalisé par ordinateur. En combinant la recherche aléatoire à la recherche heuristique, l'algorithme de recherche aléatoire peut s'adapter à la

situations. At the end, a computer simulation is used to simulate the above complex conditions. It shows that the proposed random search algorithms can accurately locate the radioactive sources.

recherche en question dans de nombreuses situations complexes. Enfin, nous utilisons une simulation informatique pour simuler les conditions complexes mentionnées précédemment. Elle démontre que les algorithmes de recherche aléatoire proposés peuvent localiser précisément les sources de rayonnement.

Statistical Issues for Longitudinal and Time Series Analyses
Problèmes statistiques liés aux analyses longitudinales et de séries temporelles

Chair/Président: Lei Sun

Room/Salle: 146 (SB)

Abstract/Résumé

[Wednesday May 29/mercredi 29 mai, 13:30-13:45]

Melody Ghahramani (The University of Winnipeg) , **Scott White** (University of Manitoba)

Time Series Regression for Zero-Inflated and Overdispersed Count Data: A Functional Response Model Approach

Régression chronologique pour les données de dénombrement à surreprésentation de zéros et surdispersées : approche modélisée de réponse fonctionnelle

Count time series data feature prominently in epidemiology, business, and environmental sciences. Often, such data exhibit zero-inflation and overdispersion in addition to serial dependence. Parametric models such as the negative binomial or zero-inflated Poisson are employed to account for overdispersion and zero-inflation. In practice, the conditional variance structure may be unknown or may not be negative binomial. In this talk, a distribution-free approach for estimation of regression parameters of conditionally overdispersed and zero-inflated time series models is developed. Model parameters are optimal in the Godambe-information sense. Simulation studies and a case study comparing our method with fully parametric methods using weekly syphilis counts from 2007–2010 in Maryland, USA are used to illustrate the benefits of the method.

Les données chronologiques de dénombrement occupent une place importante dans l'épidémiologie, les affaires et les sciences de l'environnement. Souvent, ces données font état d'une surreprésentation de zéros et d'une surdispersion, ainsi que d'une dépendance sérielle. Des modèles paramétriques, comme le modèle binomial négatif ou le modèle de Poisson avec surreprésentation de zéros sont utilisés pour tenir compte de la surdispersion et de la surreprésentation de zéros. En pratique, la structure de variance conditionnelle peut être inconnue ou ne pas être binomiale négative. Dans cet exposé, nous mettons sur pied une approche d'estimation qui ne repose pas sur une distribution pour l'estimation des paramètres de régression des modèles chronologiques conditionnellement surdispersés et à surreprésentation de zéros. Les paramètres du modèle sont optimaux au sens de l'information de Godambe. Pour illustrer les avantages de la méthode, nous utilisons des études de simulation et une étude de cas comparant notre méthode à des méthodes entièrement paramétriques au moyen de dénombrements hebdomadaires de syphilis de 2007 à 2010 dans le Maryland, États-Unis.

[Wednesday May 29/mercredi 29 mai, 13:45-14:00]

Olawale Ayilara (University of Manitoba) , **Tolulope Sajobi** (University of Calgary) , **Lisa Lix** (University of Manitoba)

Goodness-Of-Fit Indices for Testing Longitudinal Measurement Invariance in Ordinal Data

Indices d'adéquation pour tester l'invariance de la mesure longitudinale dans les données ordinales

Statistical models to test for longitudinal change in patient-reported outcomes (PROs), such as health-related quality of life, rest on the assumption of longitudinal measurement invariance (MI), which implies that the latent variable is consistently measured over time. Goodness-of-fit indices, such as the 2 goodness-of-fit statistic, root-mean-square-error-of-approximation, comparative fit index, and root-mean-square-residual have been extensively studied for continuous outcomes. However there is limited investigation of the performance of these indices in ordinal data. We evaluate the sensitivity of model fit indices in testing MI in longi-

Les modèles statistiques pour tester le changement longitudinal des résultats rapportés par les patients (RRP), comme la qualité de vie relative à la santé, reposent sur l'hypothèse de l'invariance de la mesure (IM) longitudinale, qui indique que la variable latente est invariablement mesurée au fil du temps. Les indices d'adéquation, comme la statistique d'adéquation du 2, l'erreur quadratique moyenne de l'approximation, l'indice comparatif d'ajustement et la moyenne quadratique des résidus, ont fait l'objet d'études approfondies relatives aux résultats continus. Cependant, les études sont limitées en ce qui concerne la performance de ces indices sur les données ordinales. Nous évaluons la sensibilité des indices d'ajustement du modèle en testant la IM dans

tudinal ordinal data using computer simulation and a real-world Joint Replacement Registry from Manitoba, which collects pre- and post-surgery PROs. Our study will provide guidance on choosing a fit index based on sample size, number of latent variable indicators, association between time points, and magnitude of non-invariance.

des données ordinales longitudinales au moyen de la simulation informatique et de métadonnées d'un registre des remplacements articulaires du Manitoba, qui recueille des RRP avant et après une opération. Notre étude servira de référence pour savoir quel indice d'ajustement choisir en fonction de la taille de l'échantillon, du nombre d'indicateurs de variables latentes, de l'association entre les points dans le temps et de l'importance de la non-invariance.

[Wednesday May 29/mercredi 29 mai, 14:00-14:15]

Jia Li (University of Calgary) , **Alexander de Leon** (University of Calgary) , **Haocheng Li** (University of Calgary and Roche Canada)

Likelihood Analysis of Gaussian Copula Mixed Models for Multiple Correlated Disparate Longitudinal Non-Gaussian Continuous Outcomes

Analyse de vraisemblance des modèles mixtes à copules gaussiennes pour de multiples résultats disparates et corrélés, longitudinaux et non gaussiens continus

The talk concerns the analysis of longitudinal data on multiple correlated non-Gaussian continuous outcomes, where the outcomes may include rates/proportions and time-to-event endpoints, among others. Joint modelling of correlated outcomes in practice is usually carried out via linear and generalized linear mixed models (LMMs/GLMMs), where shared or correlated random effects are introduced to account for longitudinal correlations between longitudinal measurements on the same outcome as well as for those between measurements on different outcomes. In this talk, we adopt a Gaussian copula mixed model for the disparate non-Gaussian outcomes and employ the EM algorithm for restricted maximum likelihood estimation. A real-data application is also discussed to illustrate the methodology.

Cet exposé porte sur l'analyse de données longitudinales sur de multiples résultats continus non gaussiens corrélés, où les résultats peuvent entre autres inclure des taux et des proportions, ainsi que des paramètres de temps avant l'événement. Dans la pratique, la modélisation conjointe des résultats corrélés est normalement effectuée au moyen de modèles linéaires mixtes et de modèles linéaires généralisés mixtes. Des effets aléatoires partagés ou corrélés sont introduits dans ces modèles pour tenir compte des corrélations longitudinales entre les mesures longitudinales du même résultat, ainsi qu'entre celles des différents résultats. Dans cet exposé, nous adoptons un modèle mixte à copules gaussiennes pour les résultats non gaussiens disparates, et nous utilisons l'algorithme espérance-maximisation pour l'estimation du maximum de vraisemblance restreint. Nous discutons également d'une application de données réelles pour illustrer la méthodologie.

[Wednesday May 29/mercredi 29 mai, 14:15-14:30]

Amadou Diogo Barry (Université du Québec à Montréal) , **Karim Oualkacha** (Université du Québec à Montréal) , **Arthur Charpentier** (Université du Québec à Montréal)

Penalized Weighted Asymmetric Least Squares Regression for Longitudinal Data with Fixed-effects

Régression au moindre carré asymétrique pondérée et pénalisée pour les données longitudinales avec effets fixes

The fixed-effects (FE) model is a commonly used model in econometric to analyze longitudinal data. The FE model has the advantage to account for unobserved covariates. However, the FE model does not estimate the time-invariant effect and is affected by the incidental parameter problem. To mitigate this problem, we introduce the penalized weighted asymmetric least squares regression for longitudinal data with fixed-effects. We use the l1-penalty to shrink the fixed-effects incidental parameter and provide a sparse solution. In addition, the proposed method allows inference of time-invariant covariates. We propose a block-relaxation algorithm

Le modèle à effets fixes (EF) est couramment employé en économétrie pour analyser les données longitudinales. Le modèle EF a l'avantage de tenir compte des covariables non observées. Toutefois, ce modèle n'estime pas l'effet de temps invariant et est affecté par le problème de paramètre incident. Pour minimiser le problème, nous présentons la régression des moindres carrés asymétriques pondérée et pénalisée pour les données longitudinales avec effets fixes. Nous nous servons de la pénalité L1 pour réduire le paramètre incident à effets fixes et fournir une solution éparse. De plus, la méthode proposée permet l'inférence des covariables invariables avec le temps. Nous suggérons un algorithme de relaxation par blocs combiné à un algorithme de descente par

combined with a coordinate descent algorithm to compute efficiently the estimator. The exhaustive simulation results displayed its favourable qualities under various scenarios. The usefulness of the proposed estimator is illustrated through analysis of a real dataset.

coordonnée pour efficacement calculer l'estimateur. Les résultats tirés de simulations approfondies démontrent ses qualités avantageuses dans divers scénarios. L'utilité de l'estimateur proposé est illustrée par l'entremise d'analyses d'un vrai jeu de données.

[Wednesday May 29/mercredi 29 mai, 14:30-14:45]

Julan Al-Yassin (University of Windsor) , **Richard Caron** (University of Windsor) , **Robin Gras** (University of Windsor)

Time Series: Stochastic or Chaotic?

Séries chronologiques : stochastiques ou chaotiques ?

We present a new method to determine whether a time series was generated by a stochastic or chaotic process. This distinction has implications for further analysis and prediction of the time series. The method utilizes the Poincare Higuchi method (PH method) with quantitative criterion. Our method offers an improved heuristic for the selection of the Poincare section that has been developed through careful study of the solution space. We present some results about the fractal nature of the solution space. We present numerical experiments showing that there is a significant gap between the scores resulting from the application of our method to stochastic time series data versus chaotic. We consistently obtain scores smaller than a given threshold when our method is applied to stochastic time series, thus allowing us to distinguish them from chaotic time series.

Nous présentons une nouvelle méthode qui permet de déterminer si une série chronologique a été générée par un processus stochastique ou chaotique. Cette distinction a des implications pour l'analyse et la prévision de la série. Notre méthode utilise la méthode de Poincaré Higuchi (méthode PH) avec un critère quantitatif. Elle offre une meilleure heuristique pour la sélection de la section de Poincaré, développée grâce à une étude minutieuse de l'espace de solution. Nous présentons des résultats sur la nature fractale de l'espace de solution. Nous présentons des expériences numériques qui montrent l'écart important entre les résultats quand notre méthode est appliquée à une série chronologique stochastique ou chaotique. Nous obtenons régulièrement des résultats inférieurs à un seuil donné quand notre méthode est appliquée à une série chronologique stochastique, ce qui nous permet de distinguer celles-ci des séries chronologiques chaotiques.

[Wednesday May 29/mercredi 29 mai, 14:45-15:00]

Mohsen Soltanifar (University of Toronto)

A Time Series Based Point Estimation of Stop Signal Reaction Times (SSRT)

Estimation ponctuelle chronologique des temps de réaction du signal d'arrêt

Introduction: SSRT as a measurement of latency of the unobservable human brain stopping process has been formulated by Logan in 1994 without consideration of the order of trials in it. Asymptotically equivalent and larger indexes of mixture SSRT and weighted SSRT- proposed by the speaker in 2017- addressed this issue from longitudinal perspective, but an estimation based on the time series perspective was still missing. Methods: A subsample of 44 children age 6-17 each with 96 trials in Toronto, Canada was considered. State-Space missing data EM algorithm was applied for each subject data encompassing the order of trials in it and using Logan 1994 formula on ordered data, the new State-Space SSRT index was calculated. Results: State-Space SSRT is significantly larger than Logan 1994 SSRT, mixture SSRT, and weighted SSRT with paired t-test estimates (95%CI) = 21.9 (17.4, 26.3), 8.3 (0.2, 16.4), 7.7 (1.3,

Introduction : le temps de réaction du signal d'arrêt comme mesure de la latence du processus inobservable d'arrêt du cerveau humain a été formulée par Logan en 1994 sans tenir compte de l'ordre des essais effectués. En 2017, le conférencier a présenté des indices plus vastes et asymptotiquement équivalents du mélange du temps de réaction du signal d'arrêt et du temps de réaction du signal d'arrêt pondéré et a abordé cette question d'un point de vue longitudinal, mais une estimation basée sur la perspective chronologique faisait toujours défaut. Méthodes : on a pris en compte un sous-échantillon de 44 enfants âgés de 6 à 17 ans faisant chacun l'objet de 96 essais cliniques à Toronto, au Canada. L'algorithme d'espérance-maximisation des données manquantes d'espace d'états a été appliqué pour chaque donnée de sujet englobant l'ordre des essais et au moyen de la formule de Logan (1994) sur les données ordonnées, et le nouvel indice de temps de réaction du signal d'arrêt dans l'espace d'états a été calculé. Résultats : Le temps de réaction du signal d'arrêt dans l'espace d'états est

14.1), respectively. Simulations also confirmed these findings.

significativement plus grand que le temps de réaction du signal d'arrêt de Logan (1994), le temps de réaction du signal d'arrêt de mélange et le temps de réaction du signal d'arrêt pondéré avec des estimations de test t par paires (95 % IC) = 21,9 (17,4, 26,3), 8,3 (0,2, 16,4), 7,7 (1,3, 14,1), respectivement. Les simulations ont également confirmé ces résultats.

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

Chair/Président: Lisa M. Lix

Room/Salle: 142 (AD)

Abstract/Résumé

[Wednesday May 29/mercredi 29 mai, 13:30-13:45]

Selvakkadunko Selvaratnam (University of Alberta) , **Linglong Kong** (University of Alberta) , **Douglas Wiens** (University of Alberta)

The Impact of Scale Functions on the Construction of Robust Designs for Nonlinear Quantile Regression

L'incidence des fonctions d'échelle sur la construction de plans robustes de régression quantile non linéaire

We discuss kernel estimation of scale in nonlinear quantile regression. These estimates are used in the construction of designs for estimation of quantile regression functions. The designs in turn are robust against misspecifications of the form of the quantile function. We discuss two design methods: sequential local design and adaptive design. The sequential approach is a method to determine subsequent design points, given known values of the parameters and the variances. In the adaptive approach these 'known' values are replaced by the current estimates, and then the sequential method yields the next design point. In the context of a Michaelis-Menten response we demonstrate that, asymptotically, the adaptive approach performs as well as if the parameter values were known beforehand.

Nous discutons de l'estimation par noyau de l'échelle dans la régression quantile non linéaire. Cette estimation est utilisée dans la construction des plans d'estimation des fonctions de régression quantile. De leur côté, les plans sont robustes face aux erreurs de spécification de la forme de la fonction quantile. Nous discutons de deux méthodes de plan : le plan local séquentiel et la plan adaptatif. L'approche séquentielle permet de déterminer les points des plans subséquents, compte tenu des valeurs connues des paramètres et des variances. Dans l'approche adaptative, on remplace ces valeurs « connues » par les estimations actuelles, puis la méthode séquentielle livre le prochain point de plan. Dans le contexte d'une réponse de Michaelis-Menten, nous démontrons que, asymptotiquement, l'approche adaptative fonctionne aussi bien que si les valeurs des paramètres étaient connues au préalable.

[Wednesday May 29/mercredi 29 mai, 13:45-14:00]

Gun Ho Jang (Ontario Institute for Cancer Research)

Tail Probability of Maximal Chi-Squared Statistic for Association Studies

La probabilité de queue d'une statistique au chi carré maximale en études d'association

Associations between dichotomized and continuous variables often become a two sample test. For simple comparison, the continuous covariate is further dichotomized with a threshold value. An effective threshold value is easily found by maximizing the test statistic for various threshold values. The found maximal statistic no longer follows the distribution of the considered test statistic. In this manuscript, Pearson's chi-squared test statistics are used for a homogeneity test between two response groups. The exact probability and asymptotic distribution of the maximal chi-squared statistic are derived. The maximal chi-squared statistic is also applied for the two sample problem, like goodness-of-fit tests. The performance of maximal chi-squared statistic is compared to two sample tests such as the Kolmogorov-Smirnov and Anderson-Darling tests and

Les associations entre les variables dichotomisées ou continues deviennent souvent un test à deux échantillons. Afin d'obtenir une comparaison simple, la covariable continue est dichotomisée davantage avec une valeur limite. On peut facilement trouver une valeur limite efficace en maximisant la statistique de test pour différentes valeurs limite. La statistique maximale ainsi trouvée ne suit plus la distribution de la statistique de test prise en compte. Dans ce manuscrit, nous nous servons des statistiques de test du chi carré de Pearson pour réaliser un test d'homogénéité entre deux groupes de réponses. La probabilité exacte et la distribution asymptotique de la statistique maximale au chi carré sont dérivées. On applique aussi cette dernière au problème de deux échantillons, comme des tests d'adéquation. La performance de la statistique maximale au chi carré est comparée aux tests à deux échantillons, comme le test de Kolmogorov-Smirnov et d'Anderson-Darling et le test des rangs signés de Wilcoxon. Un exemple réel illustre l'as-

Wilcoxon rank sum test. A real example shows RNA gene expression association with tumour size change in pancreatic cancer.

sociation de l'expression génétique de l'ARN avec le changement de taille d'une tumeur dans le cas d'un cancer pancréatique.

[Wednesday May 29/mercredi 29 mai, 14:00-14:15]

Cindy Xin Feng (University of Saskatchewan)

Modelling Count Data with Excessive Zeros: Does the Choice Between Zero-Inflated Model and Hurdle Model Matter?

Modéliser des données de dénombrement avec surreprésentation de zéros : le choix entre un modèle à surreprésentation de zéros et un modèle «hurdle» est-il important ?

Counts data with excessive zeros are frequently encountered in practice. For example, the number of health services visits often includes many zeros representing patients with no utilization during a follow-up time. Zero-inflated or hurdle models are often used to fit such data. However, there is still a lack of comprehensive investigation of the differences between these two types of models. We review the zero-inflated and hurdle models and conduct extensive simulation studies to evaluate the performances of both types of models. Our simulation results show that the differences between these two types of models depend on the percentage of zero-deflated data points and the discrepancy in the data generating processes between the structural zeros and sampling zeros. The final choice of the regression model should be made after a careful assessment of goodness of fit and tailored to a particular dataset in question.

On rencontre fréquemment des données de dénombrement avec surreprésentation de zéros en pratique. Par exemple, le nombre de visites dans un établissement offrant des services de soin de santé comprend souvent beaucoup de zéros représentant des patients qui ne s'y sont pas présentés dans une période de temps suivie. Pour ajuster ces données, on adopte généralement soit un modèle à surreprésentation de zéros, soit un modèle «hurdle». Toutefois, il existe très peu d'études détaillées concernant les différences entre ces deux types de modèles. Nous avons donc passé en revue les modèles «hurdle» et à surreprésentation de zéros, puis avons mené une étude de simulation approfondie pour évaluer leur performance. Nos résultats en simulation démontrent que la différence entre les deux dépend du pourcentage de points de données à sous-représentation de zéros et du décalage dans les processus de génération de données entre les zéros structuraux et les zéros d'échantillonnage. Évaluer minutieusement l'adéquation et estimer quel modèle est mieux adapté au jeu de données en question sont de bons moyens pour choisir un modèle de régression.

[Wednesday May 29/mercredi 29 mai, 14:15-14:30]

Yidan Shi (University of Waterloo) , **Leilei Zeng** (University of Waterloo) , **Mary Thompson** (Univeristy of Waterloo) , **Suzanne Tyas** (Univeristy of Waterloo)

Mixture Hidden Markov Model with Partially Observed Component Memberships

Modèle de Markov caché par mélange avec composantes d'appartenance partiellement observées

Multistate models are one of the most commonly used tools for investigating disease process data. The Hidden Markov Model (HMM) has drawn increasing attention since it utilizes the advantages of continuous-time Markov Chains while allowing the observed model to be more flexible, particularly when violation of the Markov property of the observed process is due to a mixture of different types of underlying disease development processes. We developed a hidden multistate model where the underlying model is a mixture of multiple time-homogeneous Markov models each corresponding to one disease type. Auxiliary information about the individuals' disease types, which correspond to the mixture component indicator, may be available for some of the subjects. Incorporation of this auxiliary information

Les modèles multiétats sont les outils les plus couramment utilisés pour examiner les données relatives à un processus pathogénique. Le modèle de Markov caché (MMC) retient de plus en plus l'attention depuis qu'il tire parti des avantages de chaînes de Markov à temps continu, tout en permettant une plus grande souplesse du modèle observé, en particulier quand une violation de la propriété de Markov du processus observé est causée par un mélange de divers types de processus sous-jacents de progression de la maladie. Nous avons élaboré un modèle multiétats caché dans lequel le modèle sous-jacent est un mélange de plusieurs modèles de Markov homogènes dans le temps, chacun correspondant à un type de maladie. Pour certains sujets, nous pouvons disposer de renseignements accessoires sur le type de maladie propre à chacun, qui correspond à l'indicateur des composantes du mélange. L'incorporation de cette information accessoire est envisagée. Un exemple

is considered. The method is illustrated by both a real data example and simulation studies.

avec des données réelles et des études en simulation illustrent la méthode.

[Wednesday May 29/mercredi 29 mai, 14:30-14:45]

Zheng Fan (University of Calgary) , **Hua Shen** (University of Calgary) , **Haocheng Li** (University of Calgary)

Causal Inference with Misclassification in Confounding Variables

Inférence causale avec classification erronée des variables de confusion

Casual inference pertains to statistical analyses for which researchers evaluate casual effects based on precisely measured data. In an observational study, interest often lies in estimating the casual effects that are more naturally interfered by potential confounding factors. In addition, some of the confounding variables may be measured with error or classified into an incorrect group or category. In the absence of a validation dataset, we investigated the consequences of naively ignoring the misclassification issue in confounding variables on the estimation of ATE and developed an EM algorithm through the latent variable model for parameter estimation and subsequent removal of the estimation bias of the ATE. We studied both continuous and discrete outcome variables, and the estimation methods we examined include regression, G-estimation, PS (Propensity Score) matching, PS stratification, IPW, AIPW. Simulation studies assess the performances of the proposed methods.

L'inférence causale se rapporte à des analyses statistiques où les chercheurs évaluent un effet causal sur une base de données mesurées avec précision. Dans les études observationnelles, il est souvent intéressant d'estimer des effets de causalité qui peuvent être affectés par d'éventuels variables de confusion naturels. De plus, certains des variables de confusion peuvent être mal mesurés ou mal classifiés. Pour les cas où aucune donnée de validation n'existe, nous étudions les conséquences de la non-prise en compte du problème de classification erronée des variables de confusion sur l'estimation de l'effet de traitement moyen et développons un algorithme EM en utilisant un modèle à variables latentes pour l'estimation des paramètres et l'élimination du biais d'estimation de l'effet de traitement moyen. Nous étudions des variables de résultat continus et discrets, ainsi que des méthodes d'estimation comme la régression, la G-estimation, l'appariement des scores de propension (SP), la stratification des SP, la pondération par l'inverse de la probabilité et la pondération par l'inverse de la probabilité augmentée. Nous présentons des études de simulation qui permettent d'évaluer les performances des méthodes proposées.

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

Chair/Président: Johanna G. Neslehova

Room/Salle: 113 (SS)

Abstract/Résumé

[Wednesday May 29/mercredi 29 mai, 13:30-13:45]

Lenin Arango-Castillo (Queen's University) , **Glen Takahara** (Queen's University)

Long-Range Dependence Parameter Estimation for Mixed Spectra Gaussian Processes

Estimation du paramètre de dépendance de longue portée pour processus gaussien à spectres mixtes

We present an approach to the problem of estimating the Hurst parameter in processes with mixed spectra in the context of the two most studied Gaussian Long-Range Dependent (LRD) processes: fractional Gaussian noise, fGn(H), and fractional autoregressive moving average processes, FARIMA(0, H, 0), with normal innovations. This method consists of a decomposition of signal and noise components of the processes under analysis. First, we use harmonic analysis for detection of line components. Second, we remove the line components mixed with the LRD stationary processes. Finally, we use a robust method for the estimation of the LRD parameter. We apply the method to synthetic periodic time series. Numerical results are presented. The technique is applied to stock-market trading volume time series.

Nous présentons une approche au problème de l'estimation du paramètre Hurst dans des processus avec spectres mixtes dans le cadre des deux processus gaussiens dépendants à longue portée (DLP) les plus étudiés : bruit gaussien fractionnaire, bGf(H), et les processus à moyenne mobile autorégressifs fractionnaires, FARIMA(0, H, 0), avec innovations normales. Cette méthode consiste en une décomposition des composantes du signal et du bruit des processus sous analyse. Premièrement, nous utilisons une analyse harmonique pour détecter les composantes ligne. Deuxièmement, nous retirons les composantes ligne mélangées avec les processus stationnaires DLP. Finalement, nous utilisons une méthode robuste pour l'estimation du paramètre DLP. Nous appliquons les méthodes à des séries chronologiques de périodiques synthétiques. Des résultats numériques sont présentés. La technique est appliquée à des séries chronologiques de volumes de transactions boursières.

[Wednesday May 29/mercredi 29 mai, 13:45-14:00]

Katherine Burak (University of Calgary) , **Alexander de Leon** (University of Calgary)

Cluster analysis of non-Gaussian data via mixtures of Gaussian copula distributions

Analyse de groupement de données mixtes au moyen de mélanges de distributions de copules gaussiennes

Model-based cluster analysis in non-Gaussian settings is not straightforward due to a lack of standard models for non-Gaussian data. In this talk, we adopt the class of Gaussian copula distributions (GCDs) to develop a flexible model-based clustering methodology that can accommodate a variety of non-Gaussian data, where variables may have different marginal distributions and come from different parametric families. Unlike conventional model-based approaches that rely on the assumption of conditional independence, GCDs model conditional dependence among the disparate variables using the matrix of so-called normal correlations. We outline a hybrid approach to cluster analysis that combines the method of inference functions for margins (IFM) and a parameter-expanded EM (PX-EM) algorithm. We then report simulation results to investigate

L'analyse modélisée de groupement dans des contextes non gaussiens n'est pas simple en raison de l'absence de modèles standards pour les données non gaussiennes. Dans cet exposé, nous adoptons la classe des distributions de copules gaussiennes pour élaborer une méthodologie souple de regroupement basée sur un modèle qui peut tenir compte de diverses données non gaussiennes en mode mixte, ainsi que des données comprenant des mélanges de variables discrètes et continues. Contrairement aux approches classiques modélisées qui reposent sur l'hypothèse de l'indépendance conditionnelle, les distributions de copules gaussiennes modélisent la dépendance conditionnelle parmi les variables disparates au moyen de la matrice des corrélations dites normales. Nous décrivons une approche hybride de l'estimation de vraisemblance qui combine la méthode des fonctions d'inférence pour les marges et un algorithme espérance-maximisation élargi aux paramètres (PX-EM). Nous présentons ensuite les résultats de

the performance of our methodology. Finally, we highlight the applications of this research by applying this methodology to a dataset.

la simulation afin d'examiner l'efficacité de notre méthodologie. Enfin, nous illustrons notre méthodologie sur des données mixtes recueillies pour distinguer qualitativement les remplissages de chenaux géologiques.

[Wednesday May 29/mercredi 29 mai, 14:00-14:15]

Marie-Pier Côté (Université Laval) , **Christian Genest** (McGill University)

Dependence in a Background Risk Model

La dépendance dans le modèle de risque contextuel

Many copulas, including the Archimedean and elliptical copulas, may be written as the survival copula of a random vector R(X,Y), where R is a strictly positive random variable independent of the random vector (X,Y). We present a unified framework for studying the dependence structure underlying this construction, which is called the background risk model. We obtain general formulas and interesting special cases. The usefulness of the construction for model building is illustrated with an extension of Archimedean copulas with completely monotone generators, based on the Farlie-Gumbel-Morgenstern copula. In particular, explicit expressions for the distribution and the Tail-Value-at-Risk of the aggregated risk RX+RY are available in a generalization of the widely used multivariate Pareto-II model.

Plusieurs copules, dont les archimédiennes et les elliptiques, peuvent être exprimées comme la copule de survie d'un vecteur aléatoire R(X,Y), où R est une variable aléatoire strictement positive indépendante du vecteur aléatoire (X,Y). On présente un cadre unifié pour étudier les propriétés de la structure de dépendance induite par cette représentation stochastique, appelée le modèle de risque contextuel. On obtient des expressions générales et d'intéressants cas particuliers. La construction permet de créer de nouveaux modèles tels qu'une extension des copules archimédiennes (avec générateurs complètement monotones) basée sur la copule de Farlie-Gumbel-Morgenstern. En particulier, on trouve des expressions explicites pour la distribution et la TVaR du risque agrégé RX+RY dans une généralisation du modèle Pareto-II multivarié.

[Wednesday May 29/mercredi 29 mai, 14:15-14:30]

Devan G Becker (University of Western Ontario) , **Douglas G. Woolford** (Western University) , **Charmaine B. Dean** (University of Waterloo)

A Joint-Modelling Framework for Inducing Dependence in Compound Poisson Models for Aggregate Losses with Application to Wildland Fire

Cadre de modélisation conjointe pour l'induction d'une dépendance dans les modèles de processus de Poisson composé pour les pertes globales, avec application à un feu de végétation

A common modelling framework for aggregate losses assumes that the frequency and severity distributions are independent. However, in the context of wildland fires, this may not be true. For example, when more fires are occurring the landscape is likely conducive to worse fire behaviour and hence, larger fires. We present a compound Poisson model that quantifies the dependence between frequency and severity by incorporating a shared random effect in the count and size distributions. The joint estimation of this random effect shares information between the models without assuming a causal structure. We explore spatial and temporal autocorrelation of the random effects to identify additional variation not explained by the inclusion of weather related covariates. Our model also contains hurdle and spline components to incorporate excess zeros and seasonal inhomogeneity.

Un cadre de modélisation courant pour les pertes globales présume que les distributions de fréquence et de gravité sont indépendantes. Pourtant, dans un contexte de feu de végétation, il est possible que ce soit faux. Par exemple, si plus d'un feu s'embrase, la zone est vraisemblablement plus susceptible de donner lieu à un comportement pire de l'incendie et par conséquent, d'en accroître l'ampleur. Nous présentons un modèle de processus de Poisson composé qui quantifie la dépendance entre la fréquence et la gravité en incorporant un effet aléatoire partagé dans les distributions de nombre et de dimension. L'estimation conjointe de cet effet aléatoire partage l'information entre les modèles sans présumer de structure causale. Nous explorons l'autocorrélation spatiale et temporelle des effets aléatoires afin d'identifier toute autre variation que n'explique pas l'inclusion de covariables relatives à la température. Notre modèle contient aussi des composantes « hurdle » et « spline » pour incorporer une

This type of dependence can be incorporated into aggregate insurance losses and longitudinal/time-to-event models.

surreprésentation de zéros et une inhomogénéité saisonnière. Ce type de dépendance peut être incorporé dans un modèle de limite globale d'assurance, longitudinal et de durées de vie.

[Wednesday May 29/mercredi 29 mai, 14:30-14:45]

Haoxin Zhuang (University of Waterloo) , **Liqun Diao** (University of Waterloo) , **Grace Yi** (University of Waterloo)

Composite Likelihood Methods for Analyzing Longitudinal Data with Periodic Patterns under Vine Copula Models

Méthodes de vraisemblance composée pour l'analyse de données longitudinales à structure périodique dans les modèles de copules en vignes

Longitudinal data with periodic patterns, such as longitudinal data of temperature and precipitation are common. Analysis of these data is often challenging due to the complexity of modeling and associated computational burden. We utilize a vine copula model to account for the dependence among the longitudinal responses. Such copula-based models provide rich choices for decompositions, the marginal distributions, and copula functions, which offers great flexibility for modeling longitudinal data with complex association structures. To release the computational burden and concentrate on the structure of interest, we propose composite likelihood methods, which divide the responses into periodic time blocks and leave the connecting structure between time blocks unspecified. We explore the efficiency, robustness, model selection and prediction of our proposed methods by simulation studies. The proposed model is applied to analyze a UK temperature dataset.

Il est fréquent que les données longitudinales (températures, précipitations) présentent des structures périodiques, or l'analyse de telles données est souvent difficile en raison de la complexité de la modélisation et des fardeaux de calcul associés. Dans cette présentation, nous utilisons un modèle de copules en vignes pour tenir compte de la dépendance entre les réponses longitudinales. Ces modèles fondés sur les copules offrent un vaste choix de moyens de décomposition, les distributions marginales, et les fonctions de copules, qui offrent la possibilité de modéliser des données longitudinales avec des structures d'association complexes. Pour réduire le fardeau de calcul et nous concentrer sur la structure d'intérêt, nous proposons des méthodes de vraisemblance composée, qui divisent les réponses en blocs horaires périodiques sans définir la structure de liaison entre les blocs horaires. Nous explorons l'efficacité, la robustesse, la sélection de modèle et la puissance prédictive de nos méthodes par des études de simulation. Nous appliquons le modèle proposé à l'analyse d'un jeu de données de température du Royaume-Uni.

[Wednesday May 29/mercredi 29 mai, 14:45-15:00]

Ce Zhang (University of Calgary) , **Xuewen Lu** (University of Calgary)

Efficient Estimation of the Additive Hazards Model with Bivariate Current Status Data

Estimation efficace du modèle à risques additifs avec données d'état actuel bivariées

In this paper, we present sieve maximum likelihood estimators of both finite and infinite dimensional parameters in the marginal additive hazards model with bivariate current status data. We use a copula to model the joint distribution of the bivariate survival times and constrained Bernstein polynomials to model the unknown baseline hazards functions. Compared with the existing methods for estimation of the additive hazards model, the proposed new method has two main advantages. First, our method does not need to specify the form of the copula model and can be easily implemented. Second, the proposed estimators have strong consistency, and the regression parameter estimator is asymptotically normal and semiparametrically efficient. Simulation studies reveal that the proposed estimators have good finite-sample properties. Finally, a real data appli-

Nous présentons des estimateurs de vraisemblance maximum filtre des paramètres de dimension finie et infinie dans le modèle marginal à risques additifs avec des données d'état actuel bivariées. Nous utilisons un modèle de copules pour modéliser la distribution conjointe des temps de survie bivariés et des polynômes de Bernstein sous contrainte pour modéliser les fonctions de risques de base inconnus. Par rapport aux méthodes existantes d'estimation du modèle à risques additifs, cette nouvelle méthode présente deux avantages principaux. D'abord, notre méthode n'exige aucune spécification de la forme du modèle de copules et est facile à mettre en œuvre. Ensuite, les estimateurs proposés sont très cohérents et l'estimateur du paramètre de régression est asymptotiquement normal et semiparamétriquement efficace. Des études de simulation révèlent que ces estimateurs présentent en outre de bonnes propriétés sur échantillon fini. Enfin, nous proposons en guise d'illustration une application sur données réelles.

cation is provided for illustration.

Integration of probability and non-probability samples
Intégration d'échantillons probabilistes et non probabilistes

Chair/Président: Susie Fortier

Organizer/Responsable: Jean-François Beaumont

Room/Salle: 144 (SB)

Abstract/Résumé

[Wednesday May 29/mercredi 29 mai, 15:30-16:00]

Changbao Wu (University of Waterloo) , **Yilin Chen** (University of Waterloo) , **Pengfei Li** (University of Waterloo)

Sample Matching and Double Robust Estimation with Non-Probability Samples

Comparaison d'échantillons et estimation doublement robuste avec échantillonnages non probabilistes

We provide an overview of recent developments for analyzing non-probability survey samples. Inferential frameworks and theoretical results on sample matching and double robust estimation are discussed, and finite sample performance of the estimators is examined through simulation studies. An application to analyzing a non-probability survey sample from the PEW Research Centre is presented. Some practical issues are also discussed.

Nous donnons un aperçu des progrès récents en matière d'analyse d'échantillons non probabilistes d'enquêtes. Nous discutons des cadres d'inférence et des résultats théoriques concernant la comparaison d'échantillons et l'estimation doublement robuste, et nous examinons l'efficacité des estimateurs sur un échantillon fini par des études de simulation. Nous présentons une application d'analyse d'un échantillonnage non probabiliste d'enquête du PEW Research Centre. Nous abordons également certaines questions pratiques.

[Wednesday May 29/mercredi 29 mai, 16:00-16:30]

Kenneth C.K. Chu (Statistics Canada) , **Jean-François Beaumont** (Statistics Canada)

Formation of Homogeneous Self-Selection Propensity Classes for Non-Probability Samples via Probability Samples

La formation de classes homogènes de la propension à l'autosélection pour des échantillons non probabilistes via des échantillons probabilistes

Due to rising costs, declining response rates, continual efforts to reduce response burden and today's relatively easy availability of data from non-probability survey samples, researchers have begun work on deriving methodologically sound estimators based on non-probability samples. One of the key obstacles to the utilization of non-probability sample data is self-selection bias. An emerging approach to address it is to make use of data from an auxiliary probability sample and adapt propensity score techniques to estimate self-selection propensity. In this talk, we present results of a feasibility study of applying classification trees to construct homogeneous self-selection propensity classes for a non-probability sample, where the tree construction procedure uses the combined data of the non-probability sample in question as well as a related auxiliary probability sample.

En raison de la hausse des coûts, du recul des taux de réponse, des efforts continuels pour alléger le fardeau de réponse et, de nos jours, de la disponibilité accrue de données d'échantillons non probabilistes, des chercheurs ont initié des travaux pour obtenir des estimateurs méthodologiquement solides fondés sur des échantillons non probabilistes. L'un des principaux obstacles à l'utilisation d'échantillons non probabilistes est le biais d'autosélection. Une nouvelle approche est d'utiliser les données d'un échantillon probabiliste auxiliaire et d'adapter des techniques du score de propension pour estimer la propension à l'autosélection. Dans cette communication, nous présenterons des résultats préliminaires d'une étude de faisabilité sur la construction de classes homogènes de la propension à l'autosélection, pour un échantillon non probabiliste, en appliquant des arbres de classification et en utilisant des données combinées de l'échantillon non-probabiliste en question et de l'échantillon probabiliste auxiliaire associé.

[Wednesday May 29/mercredi 29 mai, 16:30-17:00]

Marie-Hélène Felt (Bank of Canada) , **Heng Chen** (Bank of Canada) , **Christopher Henry** (Bank of Canada)

Calibration and Variance Estimation for Non-Probability Samples: An Application to the 2017 Bank of Canada Methods-Of-

Integration of probability and non-probability samples
Intégration d'échantillons probabilistes et non probabilistes

Payment Survey

Calibrage et estimation de la variance des échantillons non-probabilistes : le cas de l'enquête 2017 de la Banque du Canada sur les modes de paiement

This paper discusses weighting and variance estimation for a nonprobability quota sample, in the context of the Bank of Canada 2017 Methods-of-Payment survey. We apply the raking ratio approach to adjust for nonprobability sample selection with post-stratification and nonresponse weight adjustments. We assess the bias of various raking ratio approaches obtained with different initial weights, and with or without trimming. Finally, we estimate variances of weighted means and proportions using bootstrap replicate survey weights. Compared to probability sampling, we find that (1) reducing bias from unknown selection probabilities of nonprobability sampling requires strong assumptions, and (2) multiple weight adjustments inflate variance.

Ce document traite de la pondération et de l'estimation de la variance d'un échantillon non probabiliste établi selon la méthode des quotas, dans le cadre de l'enquête 2017 sur les modes de paiement de la Banque du Canada. Nous adoptons la méthode itérative du quotient, combinée avec de la stratification a posteriori et des ajustements pour la non-réponse, pour ajuster l'échantillon non probabiliste. Nous considérons plusieurs variantes de calibrage basées sur différents poids initiaux, et avec ou sans troncature des poids extrêmes, et les évaluons en termes de biais. Enfin, nous estimons les variances de moyennes et proportions pondérées en utilisant des poids bootstrap. Par rapport à l'échantillonnage probabiliste, nous constatons que (1) la réduction du biais à partir des probabilités de sélection inconnues de l'échantillonnage non probabiliste nécessite des hypothèses fortes ; et (2) les ajustements de poids multiples gonflent la variance.

Statistical Mining with Complex and Noisy Data
Exploitation statistique avec données complexes et bruitées

Chair/Président: Chen Xu

Organizer/Responsable: Chen Xu

Room/Salle: 201 (ENG)

Abstract/Résumé

[Wednesday May 29/mercredi 29 mai, 15:30-16:00]

Zhao Chen (Fudan University) , **Zhanxiong Xu** (Penn State University) , **Zhibiao Zhao** (Penn State University)

Efficient Estimation for Nonlinear Heteroscedastic Models Through Quantile Regression

Estimation efficace pour modèles hétéroscédastiques non-linéaires par la régression quantile

Motivated by the scarce literature on efficient estimation for nonlinear heteroscedastic models, we study efficient estimation for a general nonlinear heteroscedastic regression model. The proposed double-weighted CQR (DWCQR) method uses two weights: one weight accounts for heteroscedasticity and the other weight reflects different importance of quantiles, and optimal choices of weights are derived by minimizing the asymptotic covariance matrix. Under appropriate conditions, DWCQR with theoretical optimal weights can achieve the inverse Fisher information bound. To estimate the unknown theoretical optimal weights, we propose an adaptive procedure and show that DWCQR with estimated optimal weights can perform as well as if the optimal weights were known. The proposed method shows better empirical performance than existing methods.

Motivés par le peu de littérature sur l'estimation efficace pour les modèles hétéroscédatiques non-linéaires, nous avons étudié une procédure d'estimation efficace pour un modèle de régression hétéroscédatique non-linéaire général. La méthode doublement pondérée CQR (DWCQR) présentée utilise deux poids : un poids représente l'hétéroscédaticité et l'autre poids reflète différentes importances des quantiles, et les choix optimaux des pondérations sont obtenus par la minimisation de la matrice de covariance asymptotique. Dans des conditions appropriées, la DWCQR avec pondérations théoriques optimales peut atteindre la borne donnée par l'inverse de l'nformation Fisher. Pour estimer les pondérations théoriques optimales inconnues, nous proposons une procédure adaptive et nous démontrons que la DWCQR avec des pondérations optimales estimées peut performer aussi bien que si les pondérations optimales étaient connues. La méthode proposée présente une meilleure performance empirique que les méthodes existantes.

[Wednesday May 29/mercredi 29 mai, 16:00-16:30]

Yi Yang (McGill University)

Insurance Premium Prediction via Gradient Tree-Boosted Tweedie Compound Poisson Models

Prédiction d'une prime d'assurance au moyen de modèles de Poisson composés Tweedie avec boosting par arbre et par descente du gradient

The Tweedie GLM is a widely used method for predicting insurance premiums. However, the structure of the logarithmic mean is restricted to a linear form in the Tweedie GLM, which can be too rigid for many applications. As a better alternative, we propose a gradient tree-boosting algorithm and apply it to Tweedie compound Poisson models for pure premiums. We use a profile likelihood approach to estimate the index and dispersion parameters. Our method is capable of fitting a flexible nonlinear Tweedie model and capturing complex interactions among predictors. A simulation study confirms the excellent prediction performance of

Le modèle linéaire généralisé (MLG) Tweedie est fréquemment employé pour prédire les primes d'assurance. Toutefois, la structure de la moyenne logarithmique dans ce modèle est limitée à une forme linéaire qui peut parfois être trop stricte pour de nombreuses applications. En guise de meilleure option, nous proposons un algorithme avec boosting par arbre et par descente de gradient appliqué aux modèles de Poisson composés Tweedie pour obtenir les primes pures. Nous adoptons une approche de vraisemblance profilée pour estimer l'indice et les paramètres de dispersion. Notre méthode peut ajuster un modèle flexible non linéaire Tweedie et saisir les interactions complexes entre les prédicteurs. Une étude en simulation nous confirme que notre méthode procure d'ex-

our method. As an application, we apply our method to an auto-insurance claim data and show that the new method is superior to the existing methods in the sense that it generates more accurate premium predictions, thus helping solve the adverse selection issue. We have implemented our method in a user-friendly R package that also includes a nice visualization tool for interpreting the fitted model.

cellentes prédictions. En guise d'exemple, nous appliquons notre méthode à des données de réclamation d'assurance automobile et démontrons que la nouvelle méthode est supérieure aux méthodes actuelles, c'est-à-dire qu'elle produit des prédictions plus précises des primes, ce qui permet de mieux résoudre le problème d'antisélection. Nous avons implémenté notre méthode dans une librairie R conviviale qui comporte aussi un joli outil de visualisation pour interpréter le modèle ajusté.

[Wednesday May 29/mercredi 29 mai, 16:30-17:00]

Christina Dan Wang (New York University Shanghai) , **Zhao Chen** (Fudan University) , **Yimin Lian** (University of Science and Technology of China) , **Min Chen** (Academy of Mathematics and Systems Science)

Asset Selection Based on High Frequency Sharpe Ratio

Choix d'actifs en fonction d'un ratio de Sharpe à fréquence élevée

In portfolio choice problem, the classical Mean-Variance model in Markowitz (1952) relies heavily on the covariance structure among assets. To avoid the issue of estimating the covariance matrix with high or ultra-high dimensional data, we propose a fast procedure to reduce dimension based on a new risk/return measure constructed from intra-day high frequency data and select assets via Sure Explained Variability and Independence Screening (SEVIS). While most feature screening methods only copy with i.i.d.samples, by nature of our data, we make contributions to studying SEVIS for samples with serial correlation, specifically, for stationary mixing processes. We prove that SEVIS still satisfies sure screening property and ranking consistency property. More importantly, with the assets selected through SEVIS, we will build a portfolio that earns more excess return compared with several existing portfolio allocation methods with real data from the stock market.

Dans un problème portant sur le choix d'un portefeuille, le modèle classique moyenne-variance de Markowitz (1952) s'appuie fortement sur la structure de covariance entre les actifs. Pour éviter le problème d'estimation de la matrice de covariance avec des données de dimensions élevées ou très élevées, nous proposons une procédure rapide pour réduire la dimension en fonction d'une nouvelle mesure risque-rendement établie à partir de données de fréquence élevée intrajournalière et de certains actifs via Sure Explained Variability and Independence Screening (SEVIS). Même si la plupart des méthodes de sélection des caractéristiques copient seulement avec des échantillons indépendants et identiquement distribués, par la nature même de nos données, nous contribuons à l'étude de SEVIS pour des échantillons avec une corrélation sérielle, plus précisément pour des processus stationnaires de mélange. Dans un cas de mélange, nous prouvons que SEVIS satisfait encore à la propriété de sélection sûre et à celle de la cohérence de classement. Plus important encore, avec les actifs choisis à l'aide de SEVIS, nous établissons un portefeuille qui rapporte un excédent de fonds comparativement à plusieurs autres méthodes de répartition de portefeuille utilisant des données réelles provenant du marché boursier.

Recent developments in high-dimensional statistics
Progrès récents en statistique de grande dimension

Chair/Président: Kun Liang

Organizer/Responsable: Yingli Qin

Room/Salle: 122 (ICT)

Abstract/Résumé

[Wednesday May 29/mercredi 29 mai, 15:30-16:00]

Jun Li (Kent State University)

Change-Point Detection in High-Dimensional Time Series

Détection de points de changement dans des séries chronologiques de haute dimension

High-dimensional time series are characterized by a large number of measurements and complex dependence, and often involve abrupt changes at unknown time points. We will present some new procedures to detect change points from high-dimensional time series data. The developed procedures allow both sample size and dimensionality to diverge without constraint on the growth rate of dimensionality. Moreover, they incorporate both spatial and temporal dependence without assuming a Gaussian distribution and imposing restrictive structural assumptions on the data.

Les séries chronologiques de haute dimension sont définies par un grand nombre de mesures et une dépendance complexe et comportent souvent des changements abrupts à des points inconnus dans le temps. Nous présenterons de nouvelles procédures pour détecter les points de changement dans des données provenant de séries chronologiques de haute dimension. Les procédures développées permettent à la taille de l'échantillon et à la dimension de diverger sans contrainte sur le taux de croissance de la dimension. De plus, elles incorporent une dépendance spatiale et temporelle sans présumer d'une loi gaussienne et sans imposer des hypothèses structurales restrictives sur les données.

[Wednesday May 29/mercredi 29 mai, 16:00-16:30]

Pingshou Zhong (University of Illinois At Chicago) , **Shawn Santo** (Michigan State University)

Covariance Change Point Detection and Identification with Applications in the Brain's Dynamic Functional Connectivity

Détection et repérage de point de changement de covariance avec applications dans la connectivité cérébrale fonctionnelle et dynamique

This paper develops a new procedure to detect and identify change points among high dimensional covariances. We consider a high dimensional functional data setting where a large number of features are repeatedly measured at a large number of time points on a small number of experimental units. Functional MRI (fMRI) data are one example of high dimensional functional data. We develop a change point detection and identification procedure that can accommodate general temporal and spatial dependence. To address the computational challenging, we propose an efficient approximation algorithm to implement the proposed methods. Our proposed methods are applied to event segmentation in fMRI study through understanding brain's dynamic functional connectivity.

Cet article conçoit une nouvelle procédure pour détecter et trouver les points de changement parmi des covariances de haute dimension. Nous tenons compte d'un ensemble de données fonctionnelles de haute dimension dans lequel on mesure à répétition de nombreuses caractéristiques à de nombreux moments à partir de quelques unités expérimentales. En guise d'exemple de données fonctionnelles de haute dimension, on peut penser aux données de l'IRM fonctionnelles (IRMf). Nous créons une procédure de repérage et de détection qui s'adapte à la dépendance spatiotemporelle générale. Pour aborder ce défi computationnel, nous proposons un algorithme d'approximation efficace pour mettre en œuvre les méthodes proposées. Celles-ci sont appliquées à la segmentation d'événement dans les études sur l'IRMf par la compréhension de la connectivité cérébrale fonctionnelle et dynamique.

[Wednesday May 29/mercredi 29 mai, 16:30-17:00]

Yingli Qin (University of Waterloo) , **Yilei Wu** (University of Waterloo) , **Mu Zhu** (University of Waterloo)

Joint Estimation of Multiple High-Dimensional Covariance Matrices

Estimation conjointe de multiples matrices de covariance de haute dimension

When estimating covariance matrices for data from multiple related categories, such as subtypes of a certain disease, it is possible that these covariance matrices share some common structure. In this paper, we assume that the population precision matrix (the inverse of the covariance matrix) of each category can be decomposed into a common diagonal component, a common low-rank component, and a category-specific low-rank component. This decomposition can be motivated by a factor model, in which the effects of some latent factors are common across all categories, while those of others are category-specific. We propose a method to jointly estimate these precision (and therefore, also covariance) matrices, using a complexity penalty to encourage low rankness. Under moderate conditions, we show that our estimators are consistent. Numerical examples are provided.

Lorsqu'on estime les matrices de covariance pour des données provenant de plusieurs catégories qui y sont liées, comme les soustypes d'une certaine maladie, il est possible que ces matrices de covariances partagent une structure commune. Dans cet article, nous supposons que la matrice de précision de la population (l'inverse de la matrice de covariance) de chaque catégorie peut être décomposée en une composante diagonale commune, une composante de bas ordre commune et une composante de bas ordre à catégorie spécifique. Cette décomposition peut être motivée par un modèle factoriel, dans lequel les effets de certains facteurs latents sont nombreux dans toutes les catégories, alors que ceux des autres facteurs sont spécifiques à une catégorie. Nous proposons une méthode pour estimer conjointement ces matrices de précision (et donc aussi la covariance) au moyen d'une pénalité de complexité pour favoriser l'abaissement d'ordre. Selon des conditions modérées, nous démontrons que nos estimateurs sont convergents. Des exemples numériques sont fournis.

Biostatistics Section Presidential Address
Allocution de l'invité du Président du Groupe de biostatistique

Chair/Président: Patrick E. Brown

Organizer/Responsable: Patrick E. Brown

Room/Salle: 102 (ICT)

Abstract/Résumé

[Wednesday May 29/mercredi 29 mai, 15:30-16:35]

John Martin Bland (University of York)

Improving Statistical Quality in Published Research : The Clinical Experience

Améliorer la qualité statistique des recherches publiées : l'expérience clinique

Over the past 45 years, the quality of clinical research has improved greatly. I shall try to show how this has come about and identify key factors in this improvement. I shall go on to show that there is more to do. I shall look at the position in non-clinical biomedical research and see whether there are any lessons to be drawn from the clinical experience.

Depuis les 45 dernières années, la qualité de la recherche clinique s'est grandement améliorée. Je tâcherai de présenter comment cette amélioration s'est produite et quels ont été les éléments clés. Je démontrerai ensuite qu'il y a encore place à l'amélioration. J'examinerai la position de la recherche biomédicale non-clinique et je constaterai s'il y a des enseignements à tirer de l'expérience clinique.

New Developments in State-space Modeling Approaches for Ecology and Environmental Research
Nouvelles évolutions en méthodes de modèles d'espaces d'états pour l'écologie et la recherche environnementale

Chair/Président: Ying Zhang

Organizer/Responsable: Ying Zhang

Room/Salle: 146 (SB)

Abstract/Résumé

[Wednesday May 29/mercredi 29 mai, 15:30-16:00]

Hugh Chipman (Acadia University) , **Khurram Nadeem** (University of Guelph) , **Ying Zhang** (Acadia University)

A Hierarchical State-Space Approach for Modeling Population Indices Data

Une approche hiérarchique de l'espace d'état pour la modélisation des données sur les indices de population

A time series of animal abundance estimates, together with stochastic population dynamics (SPD) models, are key ingredients of a population viability analysis routinely used to estimate species extinction risk. Recently, distance sampling methodology has become popular for wildlife abundance estimation owing to its cost-effectiveness and ease of implementation. We use density dependent SPD models for long-term dynamics of wildlife populations with distance sampling data. We employ hierarchical state-space models to link the stochastic distance sampling model to the SPD model, while simultaneously allowing for valid inference on both model components. We illustrate the approach with an application to population reconstruction of Nova Scotia white-tailed deer.

Une série chronologique d'estimations de l'abondance animale, ainsi que des modèles stochastiques de la dynamique des populations sont les éléments clés d'une analyse de viabilité des populations couramment utilisée pour estimer le risque d'extinction des espèces. Récemment, la méthode d'échantillonnage à distance est devenue populaire pour l'estimation de l'abondance de la faune en raison de son rapport coût-efficacité et de sa facilité d'application. Nous utilisons des modèles stochastiques de la dynamique des populations dépendant de la densité pour déterminer la dynamique à long terme des populations fauniques avec des données d'échantillonnage à distance. Nous utilisons des modèles hiérarchiques d'espace d'état pour relier le modèle d'échantillonnage stochastique de distance aux modèles stochastiques de la dynamique des populations, tout en permettant simultanément une inférence valide sur les deux composantes du modèle. Nous illustrons cette approche par une application à la reconstitution de la population de cerfs de Virginie de la Nouvelle-Écosse.

[Wednesday May 29/mercredi 29 mai, 16:00-16:30]

Guohua Yan (University of New Brunswick) , **Xingde Duan** (Guizhou University of Finance and Economics) , **Xiaolei Zhang** (Yunnan Normal University) , **Renjun Ma** (University of New Brunswick) , **Ying Zhang** (Acadia University)

A Simultaneous Trajectory Modelling of Weather-Related Natural Disasters in Canada

Modélisation des trajectoires simultanées des catastrophes naturelles liées aux intempéries au Canada

As the temperature of the earth surface increases, the number of natural disasters taking place each year is on the rise. Although this issue has attracted much attention in the literature, previous studies have mainly focused on examining the links between single-typed disasters and global warming. In this work we study the long-term trends of multiple types of disasters in Canada and explore their possible link to global warming. By using a multivariate state-space model for count time series, we investigate the association between five major weather-related natural disasters and two significant

À mesure que la température de la surface de la Terre augmente, le nombre de catastrophes naturelles progresse chaque année. Bien que ce problème ait fait couler beaucoup d'encre, jusqu'ici les études ont surtout porté sur les liens entre les catastrophes de type unique et le réchauffement climatique. Dans cette présentation, nous étudions les tendances à long terme de multiples types de catastrophes au Canada et explorons leurs liens éventuels avec le réchauffement climatique. En utilisant un modèle espace d'états multivarié pour les séries chronologiques, nous étudions l'association entre cinq grandes catastrophes naturelles liées aux intempéries et deux prédicteurs statistiquement

predictors of global warming, temperature and precipitation. This multivariate trajectory analysis not only allows us to capture the heterogeneity of each type of disasters, but also helps characterize the interrelationship among different types of disasters over time.

significatifs du réchauffement climatiques, la température et les précipitations. Cette analyse de trajectoire multivariée permet non seulement de saisir l'hétérogénéité de chaque type de catastrophe, mais elle aide aussi à caractériser les interdépendances dans le temps entre divers types de catastrophes.

[Wednesday May 29/mercredi 29 mai, 16:30-17:00]

Connie Stewart (University of New Brunswick Saint John) , **Shelley Lang** (Fisheries and Oceans Canada)

Measuring Repeatability in the Diet of Grey Seals (Halichoerus Grypus)

Mesure de la répétabilité dans l'alimentation des phoques gris (Halichoerus grypus)

The diet of grey seals are frequently presented as averages with individuals of a given age, sex or morphology treated as ecologically equivalent. However, individuals can vary substantially in their resource use which, in turn, has the potential to significantly affect the structure and dynamics of populations and their communities. The degree of individual specialization can be measured by repeatability. Given estimates of the diet of individual grey seals over time, we estimate repeatability from the mean squares obtained from a non-parametric multivariate analysis of variance, as well as an appropriate measure of distance for the data at hand which, in this instance, are compositional vectors. Confidence intervals via bootstrapping are constructed; they take into account both sampling and measurement error, where the latter error arises because the diet of the seals is estimated. Results of a simulation study and two real-life data examples will be presented.

Le régime alimentaire des phoques gris est souvent présenté sous forme de moyennes (individus d'un âge, d'un sexe ou d'une morphologie donnés) qui sont traitées comme étant écologiquement équivalentes. Toutefois, l'utilisation des ressources peut varier considérablement d'un individu à l'autre, ce qui peut avoir des répercussions importantes sur la structure et la dynamique des populations et de leurs collectivités. On peut mesurer le degré de spécialisation individuelle au moyen de la répétabilité. Grâce aux estimations du régime alimentaire des phoques gris individuels au fil du temps, nous estimons la répétabilité à partir des carrés moyens obtenus à partir d'une analyse non paramétrique multivariée de la variance, ainsi que d'une mesure appropriée de la distance pour les données disponibles qui, dans ce cas, sont des vecteurs de composition. On créé des intervalles de confiance bootstrap : ils tiennent compte à la fois de l'erreur d'échantillonnage que de l'erreur de mesure (on peut tenir compte de cette dernière, car l'alimentation des phoques est estimée). Nous présenterons les résultats d'une étude de simulation et deux exemples de données réelles.

Causal Inference: Applications and Case Studies
Inférence causale : applications et études de cas

Chair/Président: Sanjeena Dang

Room/Salle: 142 (AD)

Abstract/Résumé

[Wednesday May 29/mercredi 29 mai, 15:30-15:45]

Thai-Son Tang (University of Toronto) , **Keith A. Lawson** (University Health Network) , **Antonio Finelli** (University Health Network) , **Olli Saarela** (University of Toronto)

Causal Inference Methods for Quality-Of-Care Comparisons Involving Small Institutions

Méthodes d'inférence causale pour la comparaison de la qualité des soins entre petits établissements

In comparisons of hospital quality-of-care, the adaptation of a causal modeling framework can be used to answer whether a patient would receive a different level of care had they been treated at a different hospital. To ensure such comparisons are fair, adjusting for case-mix differences between hospitals is necessary. For instance, direct standardization through inverse probability weighting, where the weights are derived from a multinomial hospital assignment model, can be used. However, fitting such a model is challenging in the presence of small hospitals due to practical positivity violations. To benefit from model extrapolation, we propose constrained estimation of the assignment model, combined with a doubly robust estimator involving a mixed effects model for the treatment received. We demonstrate that the resulting method is doubly robust for large institutions and single robust for the small ones, and illustrate this for kidney cancer surgical care in Ontario.

Lorsqu'il s'agit de comparer la qualité des soins hospitaliers, on peut adapter un cadre de modélisation causal pour savoir si un patient aurait reçu un niveau de service différent dans un autre hôpital. Pour garantir une comparaison juste, il faut tenir compte des types de cas traités dans chaque hôpital, par exemple en procédant à une standardisation directe par pondération par probabilité inverse où les poids sont dérivés d'un modèle d'attribution multinomial. Cependant, il est difficile d'ajuster ce type de modèle pour de petits hôpitaux en raison de la violation pratique de la positivité. Afin de pouvoir néanmoins extrapoler le modèle, nous proposons une estimation sous contrainte du modèle d'attribution, combinée à un estimateur doublement robuste avec un modèle à effets mixtes pour le traitement reçu. Nous montrons que la méthode qui en résulte est doublement robuste pour les grands établissements et robuste pour les petits et l'illustrons pour l'exemple des soins chirurgicaux en cancer du rein en Ontario.

[Wednesday May 29/mercredi 29 mai, 15:45-16:00]

Mohammad Ehsanul Karim (The University of British Columbia) , **Menglan Pang** (McGill University) , **Robert Platt** (McGill University)

Can We Train Machine Learning Methods to Outperform the High-Dimensional Propensity Score Algorithm?

Pouvons-nous entraîner des méthodes d'apprentissage machine pour surpasser l'algorithme de scores de propension de dimension élevée ?

The use of retrospective health care claims datasets is frequently criticized for the lack of complete information on potential confounders. Utilizing patient's health status-related information from claims datasets as proxies for unobserved confounders, the high-dimensional propensity score algorithm enables us to reduce bias. Using a cohort study of postmyocardial infarction statin use, we compare the performance of the algorithm with a number of popular machine learning approaches for confounder selection in high-dimensional covariate spaces: random forest, LASSO, and elastic net. Our

L'utilisation des ensembles des données sur les réclamations médicales rétrospectives est fréquemment critiquée pour l'absence totale d'information sur les facteurs de confusion potentiels. En utilisant l'information sur l'état de santé des patients provenant des ensembles de données sur les réclamations comme procurations pour les facteurs de confusion non-observés, l'algorithme de scores de propension de dimension élevée permet de réduire le biais. À l'aide d'une étude de cohorte sur l'utilisation de la statine pour un infarctus post-myocardique, nous comparons la performance de l'algorithme avec des approches populaires d'apprentissage machine pour la sélection de facteurs de

results suggest that, when the data analysis is done with epidemiologic principles in mind, machine learning methods perform as well as the high-dimensional propensity score algorithm. Using a plasmode framework that mimicked the empirical data, we also showed that a hybrid of machine learning and high-dimensional propensity score algorithms generally perform slightly better than both.

confusion dans des espaces de covariables de dimension élevée : forêt aléatoire, LASSO et elastic net. Nos résultats indiquent que lorsque l'analyse de données est effectuée en ayant à l'esprit des principes épidémiologique, les méthodes d'apprentissage machine performent aussi bien que l'algorithme de scores de propension de dimension élevée. En utilisant un cadre plasmode qui reproduit les données empiriques, nous avons aussi démontré qu'un algorithme hybride à la jonction de l'apprentissage machine et de l'algorithme de scores de propension de dimension élevée performe généralement légèrement mieux que chacun d'eux pris séparément.

[Wednesday May 29/mercredi 29 mai, 16:00-16:15]

Shomoita Alam (McGill University) , **Shomoita Alam** (,) **Erica Moodie** (McGill University) , **David Stephens** (McGill University)

Should a Propensity Score Model be Super? The Utility of Ensemble Procedures for Causal Adjustment

Un modèle de scores de propension doit-il être « Super » ? Utilité des procédures de prévision d'ensemble pour l'ajustement causal

In investigations of the effect of treatment on outcome, the propensity score (PS) is a tool to eliminate imbalance in the distribution of confounding variables between treatment groups. Recent work has suggested that Super Learner (SL), an ensemble method, outperforms logistic regression (LR) in nonlinear settings; however, experience with real-data analyses tends to show overfitting of the PS model using this approach. We investigated a wide range of simulated settings of varying complexities including simulations based on real data to compare the performance of LR, generalized boosted models (GBM) and SL in providing balance and for estimating the average treatment effect via PS regression, PS matching, and inverse probability of treatment weighting. We found that SL and LR are comparable in terms of covariate balance, bias and mean squared error and both outperform GBM; however, SL is computationally very expensive thus leaving no clear advantage to the more complex approach.

Dans les études sur l'effet d'un traitement sur le résultat, le score de propension (SP) est un outil qui permet d'éliminer les déséquilibres dans la distribution des variables confusionnelles entre groupes de traitement. De récents travaux suggèrent que Super Learner (SL), une méthode de prévision d'ensemble, surpasse la régression logistiques (RL) dans les situations non linéaires ; cependant, les analyses de données réelles tendent à montrer un surajustement du modèle de SP avec cette approche. Nous avons étudié un éventail de situations simulées plus ou moins complexes, y compris des simulations fondées sur des données réelles, pour comparer les performances de la LR, des modèles boostés généralisés (MBG) et de SL en ce qui concerne l'équilibre et l'estimation de l'effet de traitement moyen par régression du SP, appariement des SP et pondération par l'inverse de la probabilité de traitement. Nous avons conclu que SL et la RL sont comparables en termes d'équilibre des covariables, de biais et d'erreur quadratique moyenne et qu'elles surpassent toutes deux les MBG ; cependant, la méthode SL demande beaucoup de calculs, si bien que cette approche plus complexe ne présente aucun avantage clair.

[Wednesday May 29/mercredi 29 mai, 16:15-16:30]

Sudipta Saha (University of Toronto) , **Olli Saarela** (University of Toronto) , **Amy Liu** (Princess Margaret Cancer Centre)

A Causal Model for Simulating Subgroup Effects in Randomized Screening Trials

Un modèle de causalité pour simuler les effets des sous-groupes dans des essais de dépistage randomisés

The primary analysis of randomized screening trials for cancer typically adheres to the intention-to-screen principle, and is powered for this purpose. However, such trials often collect high-quality data that can be utilized for secondary analyses. In particular, we are interested in the causal effect of early versus delayed treatments

L'analyse principale des essais de dépistage randomisés du cancer respecte généralement le principe de l'intention de dépistage et est utilisée à cette fin. Par contre, de tels essais recueillent souvent des données de haute qualité qui peuvent être utilisées pour des analyses secondaires. Nous nous intéressons notamment à l'effet causal des traitements précoces versus des traitements tardifs sur

on mortality among the early detectable subgroup, measured as a proportional or absolute reduction in case fatality. Our objective is to discuss methods for determining the power to test such an effect under a given trial design. For this purpose, we propose a simple causal multi-state model with two design parameters, the probability of early detection of cancer in the presence of screening, and the subsequent effect of early versus delayed treatment, quantified either through a hazard ratio or acceleration factor. We illustrate the methods in the context of screening for lung cancer, and point how they could be used to design new trials.

la mortalité parmi le sous-groupe au dépistage précoce, mesuré en réduction absolue ou proportionnelle des mortalités. Notre but est de discuter de méthodes pour déterminer la capacité de tester un tel effet dans un cadre d'essais donné. À cette fin, nous proposons un simple modèle de causalité multi-états avec deux paramètres de conception, la probabilité de détection précoce du cancer en présence de dépistage et l'effet subséquent du traitement précoce versus tardif, quantifiés soit par un rapport de risque ou un facteur d'accélération. Nous illustrons les méthodes dans le cadre du dépistage du cancer du poumon et nous indiquons comment elles peuvent être utilisées pour concevoir de nouveaux essais.

[Wednesday May 29/mercredi 29 mai, 16:30-16:45]

Yasin Khadem Charvadeh (Memorial University of Newfoundland) , **Candemir Cigsar** (Memorial University of Newfoundland)

The Use of Propensity Score Matching Methods for Estimating Treatment Effects in Recurrent Events

Utilisation des méthodes d'appariement des scores de propension pour estimer l'effet du traitement lors d'événements récurrents

Observational studies are often used to investigate the effects of treatments on a specific outcome. In many observational studies, the event of interest can be a recurrent type, which means that subjects may experience the event of interest more than one time during follow-up. The lack of random allocation of treatments to subjects in observational studies may induce treatment selection bias leading to systematic differences in observed and unobserved baseline characteristics between treated and untreated subjects. Propensity score matching is a popular technique to address this issue. It is based on the estimation of the conditional probability of treatment assignment given the measured baseline characteristics. In this study, we discuss the efficiency of methods based on propensity scores to estimate the treatment effects in recurrent event rates through simulations. We consider various scenarios under the settings of time-fixed and time-dependent treatment indicators.

Des études observationnelles sont souvent utilisées pour évaluer les effets d'un traitement sur un résultat spécifique. Dans de nombreuses études observationnelles, l'événement d'intérêt peut être récurrent, ce qui veut dire que les sujets peuvent vivre l'événement plus d'une fois durant le suivi. L'absence d'une allocation aléatoire des traitements aux sujets dans les études observationnelles peut introduire un biais de sélection du traitement, ce qui donne lieu à des différences systématiques dans les caractéristiques de référence observées et non observées entre les sujets traités et non traités. L'appariement des scores de propension est une technique populaire pour traiter ce problème. Elle est basée sur l'estimation de la probabilité conditionnelle de l'allocation des traitements selon les caractéristiques de référence. Dans cette étude, nous discutons de l'efficacité des méthodes qui se basent sur les scores de propension pour estimer les effets des traitements sur les taux d'événements récurrents par la simulation. Nous considérons différents scénarios dans le cadre d'indicateurs de traitement à temps fixes et à temps dépendants.

[Wednesday May 29/mercredi 29 mai, 16:45-17:00]

Steve Ferreira Guerra (McGill University) , **Michal Abrahamowicz** (McGill University) , **Robert Platt** (McGill University)

A Novel Bootstrap Algorithm for Estimating the Variance of Longitudinal Propensity Score Matching Estimators

Un nouvel algorithme bootstrap pour estimer la variance du score de propension longitudinal des estimateurs d'appariement

Matching methods have become abundantly used to conduct causal inference for drug effects. In longitudinal settings comparing alternative sequences of treatment, Longitudinal Propensity Score Matching (LPSM)

Les méthodes d'appariement sont de plus en plus utilisées pour réaliser une inférence causale relative aux effets des médicaments. Dans un contexte longitudinal comparant les séquences alternatives d'un traitement, l'appariement du score de propension lon-

is a framework that removes time-dependent confounding by sequentially matching subjects. Given the complex nature of the matching, an important issue is the estimation of standard errors which usually do not account for the estimation of the propensity score or rely on the naïve bootstrap, which has been shown to be inconsistent in this context, resulting in biased estimation of standard errors. We propose to develop a novel bootstrap procedure to approximate the sampling distribution of LPSM estimators that accounts for the matching nature of the estimator. We use a simulation study to evaluate the performance of the proposed procedure and to compare it to other standard inferential procedures such as the traditional variance estimator, and the naïve and M-out-of-N bootstraps.

gitudinal (ASPL) est un cadre qui élimine les facteurs de confusion dépendants du temps en appariant séquentiellement les sujets. Compte tenu de la nature complexe de l'appariement, un important problème à résoudre est celui de l'estimation des erreurs standards qui ne tiennent habituellement pas compte de l'estimation du score de propension ou sinon se fie au bootstrap naïf, qui a été prouvé comme étant incohérent dans ce contexte, ce qui produit une estimation biaisée des erreurs standards. Nous proposons la conception d'une nouvelle procédure bootstrap qui tient en compte l'appariement de l'estimateur pour estimer la distribution d'échantillonnage des estimateurs ASPL. Nous adoptons une étude par simulations pour évaluer la performance de la procédure proposée et pour la comparer à d'autres procédures d'inférence standards telles que l'estimateur de variance standard ainsi que les bootstraps naïfs et ≪m out of n≫.

Modeling Time-to-Event Data
Modélisation de données de durées de vie

Chair/Président: Gyanendra Pokharel

Room/Salle: 119 (SA)

Abstract/Résumé

[Wednesday May 29/mercredi 29 mai, 15:30-15:45]

Xiaoming Lu (Memorial University of Newfoundland) , **Zhaozhi Fan** (Memorial University of Newfoundland)

A Joint Model of Longitudinal Quantiles and Multiple-Censored Survival Data

Un modèle conjoint de quantiles longitudinaux et des données de survie multiples et censurées

We propose a new joint modelling method to combine right- and interval-censored survival times and longitudinal quantiles to capture the underlying effects among them. Three models are linked by manifest variable and latent random variables to capture the dependence between and within those three models. We assume an asymmetric Laplace distribution (ALD) in the longitudinal model and Cox models for survival in order to accommodate the data structure. A Monte Carlo expectation maximization strategy is applied for the estimation, which can be directly used under any distributional assumptions for longitudinal measurements and random effects.

Nous proposons une nouvelle méthode de modélisation conjointe pour combiner les temps de survie censurés à droite et par intervalles et les quantiles longitudinaux pour saisir les effets sous-jacents parmi ceux-ci. Trois modèles sont reliés par des variables manifestes et des variables aléatoires latentes pour saisir la dépendance entre ces trois modèles et à l'intérieur de ceux-ci. Nous supposons une loi de Laplace asymétrique (LLA) dans le modèle longitudinal et les modèles de Cox relatifs à la survie pour tenir compte de la structure des données. Nous appliquons une stratégie Espérance-Maximisation de Monte Carlo pour l'estimation, qui peut être employée telle quelle selon n'importe quelles hypothèses de mesures longitudinales et d'effets aléatoires.

[Wednesday May 29/mercredi 29 mai, 15:45-16:00]

Shahedul A. Khan (University of Saskatchewan)

A Flexible Proportional Hazards Model for Joint Analysis of Longitudinal and Time-To-Event Data

Un modèle de risques proportionnels flexible pour analyse conjointe de données longitudinales et de temps jusqu'à événement

In many studies, a longitudinal response is observed along with an observation of the time to the occurrence of an event; the event can be timed from the beginning of an observation period, resulting in time-to-event data. A typical goal in such studies is to investigate the effects of the longitudinal response on the development of the event. The modern approach to handle an analysis when both the longitudinal and survival responses are collected is to jointly model the longitudinal response and the time-to-event outcome through shared random effect(s). The standard approach is to consider a linear mixed-effects model for the longitudinal response and a proportional hazards (PH) model for the association analysis. We propose a flexible PH model for joint analysis, and develop a Bayesian approach for statistical inference, implemented through the MCMC algorithm. This study suggests that the proposed model can

Dans plusieurs études, une réponse longitudinale est observée ainsi qu'une observation du temps jusqu'à ce qu'un événement se produise ; l'événement peut être chronométré depuis le début d'une période d'observation ce qui donne des données de temps jusqu'à événement. Un objectif courant dans de telles études est d'examiner les effets de la réponse longitudinale sur le développement de l'événement. L'approche moderne pour faire une analyse quand des réponses longitudinales et de survie sont recueillies est de modéliser la réponse longitudinale et le résultat du temps jusqu'à événement conjointement par des effets aléatoires partagés. L'approche standard consiste à considérer un modèle linéaire à effets mixtes pour la réponse longitudinale et un modèle de risques proportionnels (PH) pour l'analyse d'association. Nous proposons un modèle PH flexible pour analyse conjointe et nous développons une approche bayésienne pour inférence statistique mis en place par l'algorithme MCCM. Cette étude indique que la méthode proposée peut être bénéfique dans une analyse conjointe

be valuable in joint analysis of longitudinal and time-to-event data.

de données longitudinales et de temps jusqu'à événement.

[Wednesday May 29/mercredi 29 mai, 16:00-16:15]

Tingxuan Wu (University of Saskatchewan)

Randomized Survival Probability Residual for Assessing Parametric Survival Models

Résidu de probabilité de survie randomisé pour l'évaluation des modèles de survie paramétriques

Traditional residuals for diagnosing accelerated failure time models in survival analysis, such as Cox-Snell, martingale and deviance residuals, have been widely used. However, the residuals are often only examined visually, which can be subjective. The lack of objective measures to examine model adequacy has been a long-standing issue that needs to be addressed for survival analysis. A new type of residual is proposed called Normal-transformed Randomized Survival Probability (NRSP) residual. Simulation studies were conducted to compare the performance of NRSP residuals with traditional residuals. Our simulation studies demonstrated that NRSP residuals are approximately normally distributed when the fitted model is correctly specified, and have great statistical power to detect model inadequacies. We also apply NRSP residuals to a real dataset to check the goodness-of-fit of three plausible models.

Les résidus traditionnels (Cox-Snell, martingale ou déviance) sont souvent utilisés pour diagnostiquer les modèles du temps de défaillance accéléré en analyse de survie. Cependant, l'examen de ces résidus est souvent visuel, et donc subjectif. Le manque d'une mesure objective pour examiner la pertinence au modèle est un problème de longue date en analyse de survie. Nous proposons un nouveau type de résidu, appelé résidu de probabilité de survie randomisé transformé normal (Normal-transformed Randomized Survival Probability ou NRSP). Nous effectuons des études de simulation pour comparer la performance des résidus NRSP aux résidus traditionnels. Ces études démontrent que les résidus NRSP présentent une distribution approximativement normale lorsque le modèle ajusté est correctement spécifié et qu'ils ont une grande puissance statistique pour détecter les insuffisances du modèle. Nous appliquons aussi ces résidus NRSP à un jeu de données réelles pour vérifier la qualité de l'ajustement de trois modèles plausibles.

[Wednesday May 29/mercredi 29 mai, 16:15-16:30]

Rebecca A. Clark (University of Alberta) , **Yan Yuan** (University of Alberta)

Evaluating Model Accuracy under Sampling Frame for Time-To-Event Data

Évaluation de la précision d'un modèle dans un cadre d'échantillonnage de données de temps d'événement

Both sampling design and censoring of participants can impact the analysis of time-to-event outcomes in cohort studies, and weights are required to account for these features during model development and evaluation. Using simulation studies, we investigated how to appropriately weight observations, varying the relationship between the sampling design, censoring distribution and event time distribution. We assessed different weight formulations for the estimation of model accuracy measures by using various combinations of sampling weights and inverse probability-of-censoring weights. Depending on the sampling design and censoring distribution, one or more estimators gave consistent estimates for the accuracy measures. Inadequately accounting for the weights during evaluation can result in biased estimates of accuracy measures. Investigators need to consider these features in order to properly evaluate risk prediction models.

Dans les études de cohorte, le plan d'échantillonnage et la censure de participants peuvent avoir un impact sur l'analyse du temps d'événement, si bien qu'il faut, lors du développement et de l'évaluation du modèle, appliquer une pondération pour tenir compte de ces facteurs. Nous examinons ici, par des études de simulation, comment pondérer les observations au mieux, en variant la relation entre plan d'échantillonnage, distribution de la censure et distribution des temps d'événement. Nous évaluons différentes formules de poids pour l'estimation de mesures de la précision du modèle avec diverses combinaisons de poids d'échantillonnage et de pondération par l'inverse de la probabilité de censure. Selon le plan d'échantillonnage et la distribution de la censure, un ou plusieurs estimateurs produisent des estimations convergentes pour les mesures de précision. La non-prise en compte de cette pondération au moment de l'évaluation peut fausser les estimations des mesures de la précision. Les chercheurs doivent tenir compte de ces facteurs pour bien évaluer les modèles de prévision du risque.

[Wednesday May 29/mercredi 29 mai, 16:30-16:45]

Modeling Time-to-Event Data
Modélisation de données de durées de vie

Fahmida Yeasmin (University of Calgary) , **Alexander de Leon** (University of Calgary) , **Hua Shen** (University of Calgary)

Conditional Dependence in Joint Modelling of Time-To-Event and Longitudinal Outcomes

Dépendance conditionnelle appliquée à la modélisation conjointe des résultats de temps d'événements et longitudinaux

Modeling longitudinal and event-time outcomes separately has been shown to yield biased and inefficient effect size estimates for correlated outcomes. Current methodologies for joint modeling typically rely on the assumption of conditional independence of the longitudinal and event-time outcomes given subject-specific random effects. We investigated the impact of failure to account for underlying dependencies between outcomes for this assumption and showed that estimates are biased and less efficient if the assumption fails. We developed a flexible joint modeling methodology that incorporates conditional dependence between the joint outcomes in a copula-based framework, and used the method of inference function for margins and data cloning for estimation. Simulation studies were conducted to investigate the properties and evaluate the performance of the proposed methodology.

Il est démontré que modéliser les résultats de temps d'événements et longitudinaux séparément produit des estimations biaisées et inefficaces de la taille de l'effet en fonction des résultats corrélés. Les méthodologies actuelles relatives à la modélisation conjointe se fient généralement à l'hypothèse d'indépendance conditionnelle des résultats longitudinaux et de temps d'événements en fonction des effets aléatoires précis d'un sujet déterminé. Nous avons étudié quelles étaient les conséquences lorsque cette hypothèse est incapable de tenir compte des dépendances sous-jacentes entre les résultats, et nous avons démontré que les estimations sont biaisées et moins efficaces en cas d'échec de l'hypothèse. Nous avons développé une méthodologie de modélisation conjointe flexible qui intègre la dépendance conditionnelle entre les résultats conjoints dans un cadre fondé sur les copules, puis avons employé la méthode de fonction d'inférence pour les marges et le clonage de données pour l'estimation. Nous avons aussi mené des études portant sur les propriétés et l'évaluation des performances de la méthodologie proposée.

[Wednesday May 29/mercredi 29 mai, 16:45-17:00]

Mingchen Ren (University of Calgary) , **Ying Yan** (Sun Yat-sen University) , **Alexander de Leon** (University of Calgary)

Causal Mediation Analysis of a Survival Outcome with Multiple Mediators Subject to Measurement Error

Analyse de la médiation causale d'un résultat de survie avec plusieurs médiateurs faisant l'objet d'une erreur de mesure

Although recent advances in causal mediation analysis have focused on the evaluation of a treatment effect on a survival outcome with multiple mediators, they have generally relied on the tenuous assumption that the mediators are measured without error. We study causal mediation analysis of a survival outcome via an additive hazard model in the presence of multiple mismeasured mediators (and covariates). Using counterfactuals to understand relevant causal pathways, we identify the treatment's direct and indirect effects, and obtain their corrected estimators based on the classical additive error model. We use simulations to study the finite-sample performance of our corrected estimators vis-a-vis naive estimators that ignore measurement errors in the mediators. Finally, we illustrate our methodology on CD4 count data from a clinical trial comparing mono- and combination therapies in HIV-infected adults, and provide a causal interpretation of the estimated mediated effects.

Bien que les progrès récents dans l'analyse de la médiation causale se soient concentrés sur l'évaluation de l'effet d'un traitement sur un résultat de survie avec plusieurs médiateurs, ils se sont généralement appuyés sur l'hypothèse ténue que les médiateurs sont évalués sans erreur. Nous étudions l'analyse de médiation causale d'un résultat de survie à l'aide d'un modèle de risque additif en présence de multiples médiateurs (et covariables) mal mesurés. Au moyen de contrefactuels pour comprendre les voies causales pertinentes, nous déterminons les effets directs et indirects du traitement et obtenons leurs estimateurs corrigés à l'aide du modèle d'erreur additive classique. Nous utilisons des simulations pour étudier l'efficacité pour échantillons finis de nos estimateurs corrigés par rapport aux estimateurs naïfs qui ne tiennent pas compte des erreurs de mesure des médiateurs. Enfin, nous illustrons notre méthodologie sur les données de dénombrement des cellules CD4 provenant d'un essai clinique comparant les monothérapies et les combinaisons thérapeutiques chez les adultes infectés par le VIH, et nous fournissons une interprétation causale des effets médiats estimés.

Innovations in Statistical and Data Science Education
Innovations en enseignement de la statistique et de la science des données

Chair/Président: Joel A. Dubin

Room/Salle: 113 (SS)

Abstract/Résumé

[Wednesday May 29/mercredi 29 mai, 15:30-15:45]

Sohee Kang (University of Toronto Scarborough) , **Sotirios Damouras** (University of Toronto Scarborough)

Effective Online Tool for Mathematics Communication

Outil en ligne efficace pour la communication mathématique

Lack of democracy in the classroom is a major challenge, especially in the STEM fields. It is often observed that relative to their male peers, women are less likely to engage in both in- and out-of-class discussion in post-secondary mathematics and statistics courses. We developed an online real-time communication tool, Mathematics Classroom Collaborator (MC2) (http://mc2.trentu.ca) as a remedying technical tool to develop more inclusive classrooms. It makes the entry of mathematics as easy and as intuitive as possible, including an option for anonymity, and it works on a variety of platforms – smartphones, tablets, and notebook computers. In this talk, we share our experience with employing the MC2 in a statistics and introductory probability course. We then present how these ideas can be extended to develop new communication models for the technologically-enhanced class, including increased participation by women or English language learners.

Le « manque de démocratie » en salle de classe est un enjeu de taille, notamment dans le domaine des STIM. On a souvent remarqué que, par rapport à leurs homologues masculins, les femmes ont moins tendance à participer dans les discussions en cours et en dehors de la salle de classe, s'agissant des cours de mathématiques et statistique post-secondaires. Nous avons élaboré un outil en ligne de communication en temps réel, Mathematics Classroom Collaborator (MC2) (http ://mc2.trentu.ca), comme outil technique visant à rendre les salles de classe plus inclusives. Il rend la saisie mathématique aussi facile et conviviale que possible, avec une option anonymat, et fonctionne sur diverses plateformes – smartphones, tablettes et ordinateurs portatifs. Dans cette présentation, nous parlons de notre expérience du MC2 dans un cours d'introduction à la statistique et à la probabilité. Nous expliquons ensuite comment étendre ces idées au développement de nouveaux modèles de communication pour une salle de classe soutenue par la technologie et qui favorise la participation accrue des femmes et des apprenants de l'anglais.

[Wednesday May 29/mercredi 29 mai, 15:45-16:00]

Bethany J.G. White (University of Toronto) , **Lilin Tong** (University of Toronto) , **Ming Zhao** (University of Toronto)

Exploring Students' Readiness to Engage with Statistics in Life Science Research

Examen de l'intention des étudiants à se lancer en statistique dans les sciences de la vie

One way to help address inappropriate use and interpretation of statistics in life sciences research is to improve the statistics training of the researchers. A scholarship of teaching and learning project was conducted at the University of Toronto this year to explore students' perceptions about statistical practice in the life sciences and their preparedness to appropriately use statistics in research. Student attitudes and self-efficacies for statistics, as well as their abilities to recognize and handle issues related to statistical practice in life sciences research was assessed by way of surveys administered at the beginning and end of their mandatory statistics course. In this talk, I will briefly describe the course, share results of this study and consider how these find-

L'une des façons d'aborder le problème de mauvaise interprétation ou utilisation des statistiques en sciences de la vie est d'améliorer la formation en statistique des chercheurs. Nous avons mené un projet de bourse en enseignement et apprentissage à l'Université de Toronto cette année pour examiner l'opinion des étudiants concernant la pratique statistique dans les sciences de la vie et leur disposition à se servir des statistiques de façon appropriée en recherche. Au moyen d'une enquête distribuée aux étudiants au début et à la fin de leur cours obligatoire en statistique, nous avons pu évaluer leur attitude et leur sentiment d'auto-efficacité par rapport aux statistiques, ainsi que leurs aptitudes à reconnaître et à gérer les problèmes liés à la pratique statistique dans les sciences de la vie. Lors de cet exposé, j'illustrerai brièvement le cours, partagerai les résultats de cette enquête et envisagerai comment ces

ings can inform future course offerings to better prepare our students to engage with statistics, both as consumers and producers, in life science research.

résultats peuvent nous renseigner pour mieux adapter les cours offerts à l'avenir et préparer les étudiants à se lancer en statistique, en tant que consommateur ou producteur, dans les sciences de la vie.

[Wednesday May 29/mercredi 29 mai, 16:00-16:15]

Nicholas Mitsakakis (University of Toronto)

Teaching Machine Learning in the Health Sciences: Learning Experiences from Developing a New Graduate Course

Enseigner l'apprentissage machine en sciences de la santé : des expériences d'apprentissage tirées de l'élaboration d'un nouveau cours de cycle supérieur

Machine Learning and data science are increasingly popular in health sciences, which creates a demand for developing and delivering relevant courses. Here I will present my experiences from developing and teaching a new graduate course on Applied Machine Learning for Health Data at the University of Toronto. I will discuss syllabus development issues pertaining to decisions on prerequisite statistical and computing background, textbooks and other materials, format of delivery (lecturing vs. tutorial), formative (through weekly exercises) and summative (through assignments and final exams) assessment, among others. Efforts were made to place the course within the broader context of data science and in close connection with statistics, focusing on concepts and interpretation but using math (and geometry) when needed. I will discuss the students' reception of the course, achieved learning outcomes, and overall motivation, along with areas for improvement and plans for a sequel.

L'apprentissage machine et les sciences des données sont de plus en plus populaires en sciences de la santé, ce qui entraîne la création et l'offre de cours pour combler cette demande. Je présenterai ici mes expériences tirées de l'élaboration et de l'enseignement d'un nouveau cours de cycle supérieur portant sur l'apprentissage machine appliqué aux données sur la santé à l'Université de Toronto. J'aborderai les problèmes rencontrés lors de l'élaboration du plan de cours concernant l'établissement des prérequis en statistique et en informatique, la documentation et les manuels, le format du cours (magistral ou dirigé) et l'évaluation formative (grâce à des exercices hebdomadaires) et globale (au moyen de travaux pratiques et d'examens), pour en nommer quelques-uns. Nous nous sommes efforcés de situer le cours dans un contexte plus élargi des sciences des données et très étroitement lié aux statistiques, en se concentrant sur les concepts et l'interprétation, mais au moyen des mathématiques (et de la géométrie) au besoin. J'examinerai l'opinion des étudiants relative au cours, à leur apprentissage et à leurs motivations, ainsi que des points à améliorer et des plans pour la suite.

[Wednesday May 29/mercredi 29 mai, 16:15-16:30]

Tharshanna Nadarajah (St. Francis Xavier University) , **Asokan Variyath** (Memorial University of Newfoundland)

A Systematic Approach for Effective Assignment Problem Solving

Approche systématique pour une résolution efficace d'un problème d'affectation

Undergraduate teaching has always been facing the challenge of improving the quality of teaching and learning on a continuous basis. An effective way to improve the learning process is to systematically solve assignment problems. We introduce the Plan-Do-Check-Act (PDCA) approach for better problem-solving strategies in the teaching and learning process. It incorporates the statistical thinking process. We developed some case examples for understanding the methodology in different areas and that have been implemented in a few statistics undergraduate courses.

L'enseignement universitaire a toujours été confronté au problème de l'amélioration continue de la qualité de l'enseignement et de l'apprentissage. Un moyen efficace d'améliorer le processus d'apprentissage consiste à résoudre systématiquement les problèmes d'affectation. Nous présentons une approche de gestion de la qualité Planifier-Développer-Ajuster-Contrôler [Plan-Do-Check-Act (PDCA)] pour une meilleure stratégie de résolution de problème dans le processus d'enseignement et d'apprentissage, incorporant le processus de la pensée statistique. Nous avons conçu des études de cas servant à mieux comprendre la méthodologie dans divers domaines et sa mise en œuvre dans quelques cours universitaires en statistique.

Author List • Liste des auteurs

Abedin, Tasnima, 20, 68
Abrahamowicz, Michal, 57, 259
Acar, Elif, 45, 196
Adamic, Peter, 41, 176
Adegoke, Adeola, 23
Ademola, Ayoola, 27, 99
Afful, Annshirley, 25, 89
Afonso, Filipe, 22, 78
Ahmed, Ejaz Syed, 35, 139
Al-Yassin, Julan, 53, 239
Alam, Shomoita, 56, 258
Alemayehu, Wendimagegn, 50, 220
Alvandi, Amirhossein, 39, 161
Anderson, Robert, 19, 63, 64
Andrew, Paterson, 20, 67
Andrews, Jeffrey L., 30, 114
Andrulis, Irene, 38, 160
Ansari, Usama Zafar, 30, 117
Araiza Iturria, Carlos Andres, 32, 128
Arango-Castillo, Lenin, 54, 244
Augusta, Carolyn, 37, 149
Avusuglo, Wisdom S., 39, 165
Ayilara, Olawale, 24, 53, 237
Azoulay, Laurent, 29, 110

Ba, Ismaila, 50, 223
Babul, Arif, 35, 138
Bader, Gary, 46, 200
Badescu, Andrei, 32, 127
Bae, Taehan, 38, 162
Baek, Seungchul, 40, 169
Bagmar, Md. Shaddam Hossain, 27, 103
Balbuena, Lloyd, 46, 200
Balion, Cynthia, 42, 181
Barry, Amadou Diogo, 53, 238
Basnayake, Shanika, 23
Bassim, Carol, 42, 181
Beaulieu, Martin, 25, 87
Beaumont, Jean-François, 55, 248
Becker, Devan G, 54, 245
Bégin, Jean-François, 36, 144
Berkowitz, Matthew, 24
Bilayi-Biakana, Clemonell Lord Baronat, 48, 209
Bingham, Derek, 36, 39, 148, 160
Blais, Lucie, 29, 109
Bland, John Martin, 18, 56, 254
Bornbaum, Catherine, 22, 76
Bornn, Luke, 43, 187
Boughal, Hanaa, 20, 68
Boulanger, Laurence, 20, 67
Braun, John, 25, 29, 91, 110
Brennan, Andrew, 39, 166
Brenner, Bluma, 45, 195
Brossard, Myriam, 38, 159
Brown, Patrick, 21, 34, 43, 74, 135, 184
Bull, Shelley, 20, 38, 46, 50, 67, 159, 160, 201, 220
Burak, Katherine, 54, 244
Burkett, Kelly, 46, 50, 199, 219
Buro, Karen, 31, 35, 121, 140
Burr, Wesley, 27, 48, 102, 211

Cadigan, Noel, 21, 72
Cahill, Farrell, 38, 162
Cai, Kaida, 46, 203
Cai, Rutong, 38, 163
Cai, Song, 31, 120
Camirand Lemyre, Félix, 30, 114
Campbell, David A., 30, 40, 118, 168
Campbell, Harlan, 50, 224
Campbell, Pearl, 26, 96
Cantoni, Eva, 42, 184
Canty, Angelo, 20, 67
Cao, Chen, 44, 190
Cao, Jiguo, 18, 34, 40, 136, 169
Caron, Richard, 53, 239
Carriere, Keumhee Chough, 27, 97

Carroll, Raymond J., 30, 114
Castel, Sophie, 27, 102
Chakraborty, Shubhadeep, 36, 149
Chaoubi, Ihsan, 32, 129
Charles, Colin, 21, 74
Charpentier, Arthur, 53, 238
Chatrchi, Golshid, 31, 120
Che, Menglu, 28, 103
Chekouo, Thierry, 41, 49, 174, 219
Chen, Anqi, 38, 158
Chen, Bo, 30, 118
Chen, Gemai, 35, 140
Chen, Hanning, 37, 153
Chen, Heng, 55, 248
Chen, Li-Pang, 38, 157
Chen, Min, 55, 251
Chen, Sixia, 42, 184
Chen, Su, 51, 228
Chen, Ting-Huei, 20, 68
Chen, Yan, 24
Chen, Yilin, 22, 55, 78, 248
Chen, Zhao, 55, 250, 251
Chepita, Ryan, 25, 87
Chipman, Hugh, 44, 56, 189, 255
Chow, Benjamin, 31, 124
Chowdhury, Mashfiqul, 52, 230
Chu, Kenneth C.K., 55, 248
Cigsar, Candemir, 57, 259
Ciro de Oliveira, Deive, 41, 178
Clark, Rebecca A., 57, 262
Coache, Anthony, 23, 81
Coblenz, Maximilian, 37, 150
Coeurjolly, Jean François, 50, 223
Cohen Freue, Gabriela, 35, 140
Cossette, Hélène, 32, 41, 129, 176
Côté, Marie-Pier, 54, 245
Coulombe, Janie, 45, 196
Cowen, Laura L.E., 21, 73
Craiu, Radu, 38, 46, 48, 159, 201, 213
Cui, Hengjian, 25, 89
Cui, Jingyu, 46, 203
Czaplicki, Nicole, 39, 166

Dai, Xiongtao, 34, 136
Daignault, Katherine, 32, 125
Damouras, Sotirios, 58, 264
Dampf, Hana, 24
Dang, Jessica Ou, 32, 128
Davis, Karelyn, 29, 112, 113
Davison, Matt, 36, 146
de Leon, Alexander, 26, 53, 54, 57, 94, 238, 244, 263
de Tibeiro, Jules J. S., 22, 78
Dean, Charmaine B., 54, 245
Deardon, Rob, 34, 37, 42, 135, 149, 180
Deeth, Lorna, 42, 180
Delaigle, Aurore, 30, 114
Deng, Dianliang, 43, 52, 185, 230
Descary, Marie-Hélène, 45, 197
Deshaies-Moreault, Catherine, 39, 166
Dey, Rajib, 50, 221
Dharmasena, Isuru, 23
Diao, Liqun, 54, 246
Diday, Edwin, 22, 78
Ding, Keyue, 27, 98
Ding, Xin, 23
Dinh, Vu, 24, 85
Dong, Larry, 23
Dong, Mei, 46, 200
Dongmo Jiongo, Valéry, 39, 167
Doung, Mylinh, 42, 181
Duan, Xingde, 56, 255
Duanmu, Haosui, 19, 63
Duchesne, Thierry, 43, 184
Dumbacher, Brian, 39, 166
Dumitrescu, Laura, 31, 120
Dunham, Bruce, 18
Durand, Madeleine, 29, 109
Duval, Francis, 32, 127
Dyck, Justin Wayne, 42, 180

Elliott, Lloyd T, 44, 190
Eltinge, John, 25, 87
Epasinghege Dona, Nirodha Mihirani, 23, 82
Escobar-Anel, Marcos, 36, 145, 146
Espin-Garcia, Osvaldo, 38, 46, 159, 201
Ewusie, Joycelyne E, 46, 199

Falk, Carl F., 41, 173
Fan, Bo, 26, 92
Fan, Zhaozhi, 57, 261
Fan, Zheng, 54, 243
Fang, Junhan, 26, 29, 95, 109
Fani, Shabnam, 46, 202
Felt, Marie-Hélène, 55, 248
Feng, Cindy Xin, 42, 54, 181, 242
Feng, Mingbin, 32, 128

Feng, Runhuan, 47, 207
Feng, Zeny, 42, 180
Ferland, René, 49, 218
Ferreira Guerra, Steve, 57, 259
Fétique, Ninon, 24, 85
Feuerstahler, Leah M., 41, 173
Finelli, Antonio, 30, 32, 56, 118, 125, 257
Finselbach, Hannah, 25, 88
Fontaine, Simon, 26, 92
Fop, Michael, 44, 192
Fortier, Susie, 25, 87
Fournier, Patrick, 52, 233
Fung, Tsz Chai, 32, 127
Furman, Edward, 19, 61

Gaillardetz, Patrice, 39, 164
Gambino, Jack, 43, 186
Gao, Yu, 44, 191
Gao, Yuxiang, 22, 77
Gao, Zheng, 36, 47, 148, 209
Garrido, Jose, 41, 178
Gauthier, Geneviève, 47, 205
Gavanji, Roya, 42, 181
Ge, Xinyi, 23, 27, 97
Genest, Christian, 54, 245
Ghahramani, Melody, 53, 237
Gibbs, Alison L., 48, 211
Gillis, Darren, 21, 74
Godin, Frédéric, 32, 128
Gong, Zhenxian, 36, 145
Gonzalez, Alejandro, 29, 112
Gos, Gessica, 38, 160
Gould, Robert, 48, 212
Graham, Jinko, 20, 38, 67, 156
Gras, Robin, 53, 239
Greco, Anthony, 27, 99
Griffith, Lauren, 42, 181
Grothe, Oliver, 37, 150
Gu, Xing, 36, 146
Gu, Yuqi, 44, 193
Gu, Yuwen, 26, 92
Guérin, Hélène, 24, 85
Guerrier, Stéphane, 52, 233
Guo, Beibei, 49, 215
Gustafson, Paul, 27, 98
Gutoskie, Joshua, 22, 76

Hachem, Saeb, 39, 164
Hadley, Daniel, 41, 177
Hamm, Naomi, 24
Hamzeh Hosseini, Seyed, 29, 112
Han, Peisong, 27, 28, 102, 103
Hardy, Isabelle, 45, 195
Hardy, Mary, 32, 128
Haris, Asad, 26, 93
Hasan, M. Tariq, 46, 203
Hatefi, Armin, 22, 39, 81, 161
Haziza, David, 18, 42, 184
He, Wenqing, 29, 109
He, Xuming, 49, 216
Heng, Jiani, 23
Henry, Christopher, 55, 248
Herrmann, Klaus, 37, 150
Hill, Michael D., 27, 99
Ho, Lam, 24, 85
Hofert, Marius, 31, 37, 121, 150
Hong, Hanping, 47, 207
Hoque, Md Erfanul, 24, 45, 196
Hossain, Alomgir, 31, 124
Hossain, Shakhawat, 42, 183
Hosseini, Zeinab, 29, 112
Hsu, Grace Guan, 36, 148
Hu, Beijia, 38, 163
Hu, Joan, 29, 43, 110, 187
Hu, Pingzhao, 37, 152
Hu, Yaozhong, 24, 86
Huang, Bo, 49, 214
Huang, Guowen, 21, 74
Huang, Longlong, 46, 202
Huang, Mei Ling, 52, 230
Huang, Whitney K., 45, 194
Hwang, Heungsun, 40, 172
Hyndman, Cody, 28, 107

Ibanescu, Ilinca-Ruxandra, 45, 195
Ibañez, Dominique, 29, 112, 113
Islam, Naorin, 29, 112
Isserlin, Ruth, 46, 200
Ivanoff, Gail, 48, 209

Jafari Jozani, Mohammad, 22, 37, 52, 81, 152, 231
Jalbert, Jonathan, 30, 114
Jang, Gun Ho, 53, 241
Jayasinghe, Pramoda Sachinthana, 37, 152
Ji, Yunqi, 38, 42, 162, 179
Jiang, Bei, 44, 49, 192, 216

Jiang, Cong, 31, 123
Jiang, Fei, 40, 169
Jiang, Wenjun, 47, 207
Jiang, Yidi, 24
Jirasek, Andrew, 25, 91
Joe, Harry, 41, 177
Johnson, Brad, 23, 82
Jung, Hyejung, 23

Kamso, Mohammed Mujaab, 24
Kang, Sohee, 58, 264
Kaputa, Stephen, 39, 166
Karemera, Mucyo, 52, 233
Karim, Mohammad Ehsanul, 27, 56, 98, 257
Karimnezhad, Ali, 26, 96
Karmaus, Wilfried, 51, 228
Karunamuni, Rohana, 52, 234
Kashlak, Adam, 37, 45, 153, 197
Kawaguchi, Eric, 21, 70
Kebbe, Nadine, 29, 113
Kennedy, Lauren, 22, 77
Kenny, Natasha, 18
Keshavarz, Pardis, 29, 112
Khadem Charvadeh, Yasin, 57, 259
Khan, Shahedul A., 57, 261
Kharoubi, Rachid, 37, 154
Kim, Albert Y., 51, 225
Kim, Su Hwan, 27, 97
Kokoszka, Piotr, 40, 168
Kolkiewicz, Adam, 45, 197
Kong, Linglong, 40, 44, 49, 52, 53, 171, 189, 216, 217, 234, 241
Kopciuk, Karen, 46, 52, 202, 235
Kordi, Behzad, 37, 152
Kornas, Kathy, 22, 76
Kouritzin, Michael, 41, 177
Kratsios, Anastasis, 28, 107
Kulik, Rafal, 48, 209, 210
Kulperger, Reg, 36, 144
Kuznetsov, Alexey, 19, 47, 61, 207

Labbe, Aurélie, 45, 195
Lac, Le An, 42, 183
Lakens, Daniel, 50, 224
Lamidi, Mubasiru, 24
Lang, Shelley, 56, 256
Larribe, Fabrice, 52, 233
Lawless, Jerry, 28, 45, 103, 194
Lawson, Andrew, 34, 134
Lawson, Keith A., 30, 32, 56, 118, 125, 257
Le, Khoa, 24, 86
Leblanc, Alexandre, 42, 183
Léger, Christian, 22, 77
Lekivetz, Ryan, 33, 131
Lele, Subhash, 40, 168
Lemzouji, Khalid, 18
Li, Daniel, 48, 214
Li, Gang, 21, 70
Li, Haocheng, 53, 54, 238, 243
Li, Jia, 53, 238
Li, Juan, 40, 172
Li, Jun, 55, 252
Li, Longhai, 46, 200
Li, Na, 23
Li, Pengfei, 22, 26, 55, 78, 94, 248
Li, Shuanming, 47, 208
Li, Yifan, 36, 144
Li, Yunjing, 23, 37, 161
Li, Zhigang, 21, 71
Lian, Yi, 49, 217
Lian, Yimin, 55, 251
Liang, Kun, 44, 191
Liang, You, 51, 229
Liao, Fangming, 23
Lin, Ling, 23
Lin, Ruitao, 48, 214
Lin, Sheldon, 32, 127
Lin, Shih-Kuei, 36, 145
Lin, Wei-Hsiang, 36, 145
Liu, Alexander, 38, 163
Liu, Amy, 57, 258
Liu, Coco, 24
Liu, Dongmeng, 38, 156
Liu, Jeff, 46, 200
Liu, Juxin, 25, 89
Liu, Meng, 27, 101
Liu, Peng, 49, 216
Liu, Suyu, 49, 215
Liu, Xiaohua, 38, 162
Liu, Xiufang, 43, 185
Liu, Yunfang, 23
Liu, Zhenqiu, 21, 70
Liu, Zhihui (Amy), 31, 123
Lix, Lisa, 53, 237
Lo, Bryan, 26, 96
Loader, Dylan, 23

Lockhart, Richard, 44, 189
Loeb, Peter A, 19, 63
Long, Quan, 44, 190
Lou, Wendy, 37, 44, 154, 189
Loukine, Lidia, 29, 112
Lovblom, Leif Erik, 23, 42, 179
Lowerison, Mark, 23, 27, 99
Lu, Hua, 23
Lu, Shengjie, 44, 190
Lu, Xiaoming, 57, 261
Lu, Xuewen, 38, 44, 46, 53, 54, 163, 190, 202, 203, 235, 246
Lu, Yi, 47, 208
Lu, Zihang, 37, 154
Luo, Liping, 38, 163
Luo, Xiaodong, 49, 214
Lysy, Martin, 22, 31, 80, 120

Ma, Jinhui, 42, 181
Ma, Renjun, 25, 46, 56, 91, 203, 255
Ma, Xiayi, 24
Ma, Yanyuan, 40, 169
Macdonald, Peter D.M., 31, 121
Mackay, Anne, 41, 177
MacKinnon, James G., 50, 222
Mahsin, Md, 34, 135
Mailhot, Mélina, 19, 32, 61, 128
Malrieu, Florent, 24, 85
Mamon, Rogemar, 36, 146
Mandal, Saumen, 37, 50, 150, 223
Manuel, Christopher M., 35, 142
Marceau, Étienne, 32, 41, 129, 176
Marshall, François A, 22, 80
Matsen, Frederick A., 24, 85
McDougall, Janet Elizabeth, 18
McLeod, Bob, 37, 152
McNealis, Vanessa, 22, 77
Meagher, Karen, 33, 132
Menon, Bijoy K., 27, 99
Metzler, Adam, 39, 165
Michal, Victoire, 42, 184
Miles, Justin, 19, 61
Mills Flemming, Joanna Elizabeth, 34, 134
Mills, Shirley, 51, 225
Miron, Julien, 42, 184
Mitsakakis, Nicholas, 37, 42, 58, 161, 179, 265
Mohsin, Faizan, 23
Monahan, Adam, 45, 194
Moodie, Erica, 29, 45, 56, 110, 196, 258
Morgan, Joanne, 21, 72
Morgan, Joseph, 32, 33, 131
Morris, Max D., 51, 227
Mortensen, Jacob, 43, 187
Mukherjee, Himadri, 49, 219
Müller, Hans-Georg, 34, 136
Munaweera, Inesh, 21, 74
Murphy, Thomas Brendan, 44, 192
Muthukumarana, Saman, 21, 42, 74, 183

Nadarajah, Tharshanna, 58, 265
Nadeem, Khurram, 56, 255
Naik, Shanoja, 41, 176
Naqvi, Ali, 23
Nathoo, Farouk, 30, 35, 116, 138, 139
Negrea, Jeffrey, 38, 158
Neslehova, Johanna G., 47, 206
Nie, Lei, 48, 214
Nie, Yunlong, 35, 138
Nielsen, Morten O., 50, 222
Niu, Di, 49, 216
Nolde, Natalia, 38, 41, 48, 157, 177, 210

Oldford, Wayne, 30, 117
Omar, Zayd, 32, 125
Onifade, Maryam Yetunde, 50, 219
Opathalage, Sachithra, 23
Opoku, Eugene, 35, 139
Orso, Samuel, 52, 233
Oualkacha, Karim, 37, 53, 154, 238
Ozturk, Omer, 22, 81

Panarella, Michela, 38, 160
Pang, Menglan, 56, 257
Park, Peter, 23
Parker, Matthew R., 21, 73
Paterson, Andrew, 38, 159
Pattison, Vivian, 21, 73
Pecku, Margaret, 23
Peng, Yingwei, 27, 97
Pensky, Marianna, 26, 92
Peragaswaththe Liyanage, Janaka, 21, 71
Perkins, Theodore J., 26, 96
Perreault, Luc, 30, 114
Pham, Song, 24
Phelan, Gabriel C., 30, 118
Picka, Jeffrey D., 22, 81
Pickard, Darcy, 18
Pietrosanu, Matthew, 23, 49, 217

Pigeon, Mathieu, 32, 127
Plante, Jean-François, 43, 184
Platt, Robert, 26, 29, 45, 49, 56, 57, 93, 109, 110, 196, 217, 257, 259
Pohl, Andrew, 23
Poilane, Benjamin, 42, 184
Pokharel, Gyanendra, 52, 235
Porth, Lysa, 51, 229
Prakash Patil, Rashmi, 29, 112
Prasad, Avinash, 31, 121
Punzalan, Joyce Raymond, 26, 94
Punzo, Antonio, 28, 105

Qian, Wei, 26, 92
Qiao, Cunye, 29, 112
Qin, Yingli, 55, 252
Qiu, Jinniao, 28, 108
Quan, Hui, 49, 214

Raftery, Adrian E., 44, 192
Rahman, Azizur, 23
Raimondo, Roberto, 19, 64
Raina, Parminder, 42, 181
Rajapakshage, Rasika, 26, 92
Rajapakshe, Rasika, 25, 91
Ramsay, James O., 40, 172
Randell, Edward, 38, 162
Rao, JNK, 31, 120
Rastegari, Javad, 36, 146
Reesor, Mark, 28, 39, 107, 165
Reid, Nancy, 38, 43, 156, 184
Ren, Jerry, 42, 179
Ren, Jiandong, 47, 207
Ren, Mingchen, 57, 263
Renaud, Jean-François, 39, 164
Rice, Greg, 40, 45, 168, 197
Richard, Dan, 31, 121
Richards, Kate, 21, 73
Richardson, Sylvia, 19, 60
Rivest, Louis-Paul, 44, 189
Robson, Paula, 52, 235
Roger, Michel, 45, 195
Romanescu, Razvan, 38, 160
Rondeau, Isabelle, 29, 113
Rosella, Laura, 22, 76
Rosen, Dan, 19, 61
Rostamiforooshani, Mehdi, 38, 156
Rosychuk, Rhonda, 18
Roy, Mili, 24, 50, 222
Roy-Gagnon, Marie-Hélène, 46, 199

Saad, Salma, 49, 218
Saarela, Olli, 30–32, 38, 56, 57, 118, 123, 125, 156, 257, 258
Sadeghpour, Alireza, 29, 112, 113
Safo, Sandra, 41, 174
Saha Chaudhuri, Paramita, 50, 221
Saha, Sudipta, 57, 258
Sajobi, Tolulope, 27, 53, 99, 237
Salahub, Christopher, 30, 117
Sanders, Eric, 27, 98
Sandholtz, Nathan, 43, 187
Sang, Peijun, 34, 38, 52, 136, 159, 232
Santo, Shawn, 55, 252
Saunders, David, 19, 61
Schmidt, Alexandra M., 32, 125
Schnitzer, Mireille E., 29, 109
Schuckers, Michael E., 43, 187
Schultz, Geoff, 42
Schulz, David, 23
Scrucca, Luca, 44, 192
Selvaratnam, Selvakkadunko, 53, 241
Sendova, Kristina, 47, 208
Shack, Lorriane, 52, 235
Shamloo, Arash, 29, 112
Shang, Han Lin, 40, 168
Shen, Hua, 26–28, 44, 46, 54, 57, 94, 103, 106, 189, 202, 203, 243, 263
Shi, Shan, 30, 116
Shi, Yidan, 54, 242
Silva, Rajitha, 43, 187
Simoneau, Gabrielle, 29, 110
Simpson, Daniel, 22, 77
Singh, Gurbakhshash, 27, 99
Skentelbery, Rachel, 25, 88
Smith, Aaron, 19, 63
So, Hon Yiu, 42, 181
Soave, David, 45, 194
Sobhan, Shamsia, 24
Soltanifar, Mohsen, 53, 239
Solymos, Peter, 40, 168
Song, Peter X, 51, 228
Song, Yin, 35, 138
Soulier, Philippe, 48, 210
Spicker, Dylan, 31, 124
Spinelli, John, 52, 235
Stallard, Jim, 18, 20, 66

Stanley, Anu, 42, 180
Steele, Russell J., 44, 192
Stentoft, Lars, 36, 146
Stephens, David, 32, 45, 56, 125, 195, 258
Stewart, Connie, 56, 256
Stewart, David J., 26, 96
Stingo, Francesco Claudio, 41, 174
Stoev, Stilian, 36, 47, 148, 209
Stryhn, Henrik, 25, 91
Su, Wanhua, 26, 31, 95, 121
Su, Zhihua, 52, 234
Suchard, Marc, 21, 24, 70, 85
Sun, Guang, 38, 162
Sun, Yuan, 49, 216
Surjanovic, Sonja, 52, 234
Swartz, Tim B., 43, 187
Sylvestre, Marie-Pierre, 20, 67
Sze, Connie, 38, 163

Taback, Nathan A., 51, 225
Tadayon, Vahid, 36, 142
Takahara, Glen, 54, 244
Tang, Boxin, 38, 158
Tang, Thai-Son, 56, 257
Taylor, Graham W., 37, 149
Thiessen, David Luke, 27, 101
Thind, Barinder, 24
Thompson, John R.J., 21, 72
Thompson, Katherine Jenny, 39, 166
Thompson, Mary, 31, 54, 123, 242
Thomson, Trevor, 29, 110
Tian, Jiahao, 24
Tieu, Jenny, 52, 230
Timbers, Tiffany A, 20, 65
Tong, Lilin, 58, 264
Torabi, Mahmoud, 36, 42, 45, 142, 180, 196
Tortora, Cristina, 28, 105
Tran, Anh Nam, 37, 150
Trask, Catherine, 42, 181
Trotz-Williams, Lise, 42, 180
Trufin, Julien, 41, 176
Tsai, Cary Chi-Liang, 36, 145
Tu, Dongsheng, 27, 97
Tu, Wei, 23, 52, 234
Turner, Rolf, 21, 73
Tyas, Suzanne, 54, 242

Van Bussel, Melissa, 27, 102
van den Heuvel, Edwin, 42, 181
Variyath, Asokan, 58, 265
Vatanparast, Hassan, 29, 112
Verschoor, Chris, 42, 181
Vézina, Geneviève, 39, 166
Victoria-Feser, Maria-Pia, 52, 233
Vigneault, Michel, 29, 112, 113
Villandré, Luc, 43, 45, 184, 195
Voisin, Veronique, 46, 200
Volodin, Andrei, 49, 218

Wallace, Michael, 31, 123, 124
Wang, Christina Dan, 55, 251
Wang, Dehui, 43, 185
Wang, Jane-Ling, 34, 137
Wang, Jingyu, 23
Wang, Kuan Chiao, 29, 112
Wang, Liqun, 25, 89
Wang, Meng, 24
Wang, Naisyin, 44, 192
Wang, Qiongbin, 24
Wang, Ruodu, 32, 129
Wang, Xiao, 40, 171
Wang, Xikui, 44, 51, 189, 229
Wang, Xu (Sunny), 20, 65
Wang, Yingqi, 25, 91
Wang, Yue, 23
Ward, Madeline, 42, 180
Warren, Katherine, 48, 214
Watier, François, 23, 49, 81, 218
Watkinson, Douglas, 21, 74
Watson, Tristan, 22, 76
Weaver, Colin, 23
Webb, Matthew D., 50, 222
Wei, Yunran, 32, 129
Welch, William J., 52, 234
Westerhout, Cynthia, 50, 220
White, Bethany J.G., 58, 264
White, Scott, 23, 53, 237
Wiberg, Marie, 40, 172
Wickramasinghe, Lahiru R., 42, 183
Wiens, Douglas, 34, 53, 133, 241
Williamson, Daniel, 39, 160
Wirjanto, Tony, 40, 168
Woolford, Douglas G., 54, 245
Wu, Changbao, 22, 26, 27, 55, 78, 94, 102, 248
Wu, Haoyu, 24
Wu, Jingjing, 20, 37, 46, 68, 153, 202

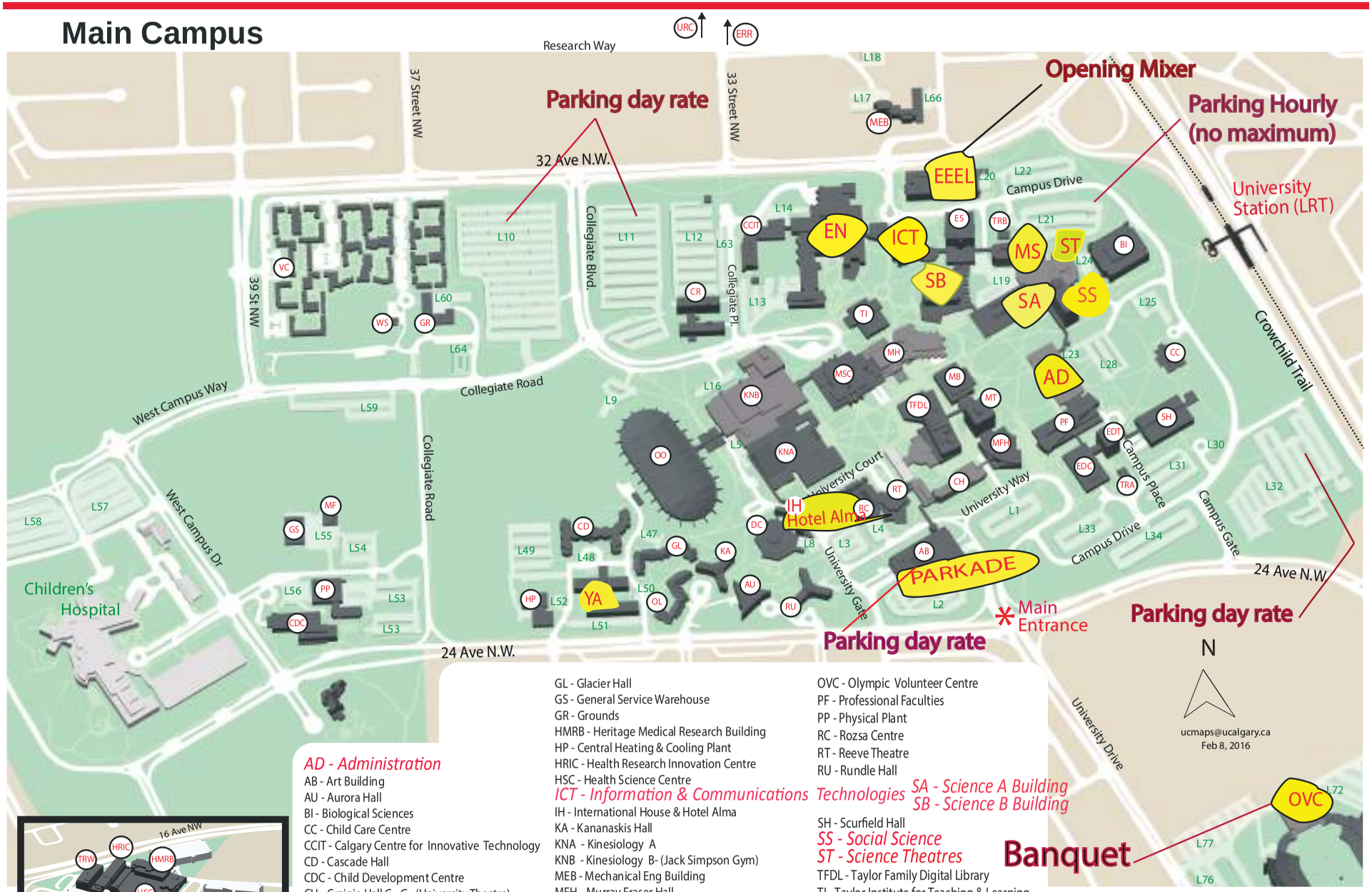
Wu, Lucas, 43, 187
Wu, Mingkuan, 23
Wu, Steven, 30, 116
Wu, Tingxuan, 57, 262
Wu, Weichi, 36, 143
Wu, Yilei, 55, 252

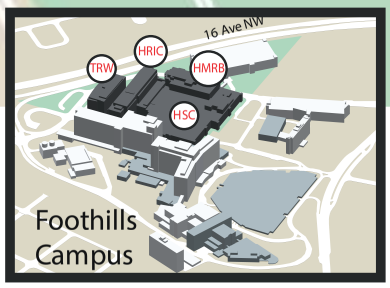
Xie, Yijun, 45, 197
Xiong, Zhi, 44, 190
Xu, Changchang, 50, 220
Xu, Changjiang, 46, 200
Xu, Gongjun, 44, 193
Xu, Wei, 31, 123
Xu, Xiaojian, 27, 99
Xu, Zhanxiong, 55, 250
Xue, Lin, 25, 89

Yan, Guohua, 56, 255
Yan, Jingyi, 23
Yan, Ying, 57, 263
Yang, Bowei, 23
Yang, Daniel, 23
Yang, Dominik Zhongda, 24
Yang, Dongyang, 24
Yang, Fenghao, 47, 207
Yang, Jun, 31, 122
Yang, Xiande, 23
Yang, Yi, 26, 49, 55, 92, 217, 250
Yao, Yao, 38, 163
Yazdi, Faezeh, 39, 160
Yeasmin, Fahmida, 57, 263
Yi, Grace, 25, 26, 29, 38, 45, 54, 89, 95, 109, 157, 199, 246
Yi, Liu, 49, 216
Yi, Yanqing, 38, 162
You, Jiaying, 37, 152
Yu, Dengdeng, 49, 217
Yu, Hao, 36, 144, 146
Yu, Menggang, 21, 70
Yu, Qianhui, 23
Yuan, Meng, 26, 94
Yuan, Yan, 57, 262
Yuan, Ying, 48, 49, 214, 215

Zeng, Leilei, 54, 242
Zeng, Yanni, 24
Zhang, Ce, 54, 246
Zhang, Chengkai, 30, 117
Zhang, Jiaxin, 23, 37, 153
Zhang, Jinyuan, 48, 210
Zhang, Kai, 23, 40, 170
Zhang, Lisu, 23
Zhang, Min, 53, 235
Zhang, Qihuang, 45, 199
Zhang, Qiong, 23
Zhang, Shixiao, 27, 102
Zhang, Xianyang, 36, 149
Zhang, Xiaoke, 34, 137
Zhang, Xiaolei, 56, 255
Zhang, Ying, 56, 255
Zhang, Yuming, 52, 233
Zhang, Zhengwu, 40, 171
Zhang, Zhiyue, 23
Zhao, Lihui, 21, 71
Zhao, Ming, 58, 264
Zhao, Yang, 27, 101
Zhao, Yang YZ, 101
Zhao, Yuqian, 40, 168
Zhao, Zhibiao, 55, 250
Zheng, Nan, 21, 72
Zhong, Pingshou, 55, 252
Zhou, Chen, 38, 157
Zhou, Christina Zihan, 24
Zhou, Menglin, 38, 157
Zhou, Wenzhuo, 40, 170
Zhou, Xinghua, 28, 107
Zhou, Zhiyang, 38, 159
Zhou, Zhou, 31, 36, 122, 143
Zhu, Feiyu, 22, 80
Zhu, Hongtu, 40, 171
Zhu, Mu, 31, 55, 121, 252
Zhu, Rui, 49, 216
Zhu, Ruoqing, 40, 170
Zhu, Yang, 24
Zhu, Zimo, 23
Zhuang, Haoxin, 54, 246
Zuyderhoff, Pierre, 41, 176
Zwiers, Francis, 45, 194

Main Campus





AD - Administration

AB - Art Building
AU - Aurora Hall
BI - Biological Sciences
CC - Child Care Centre
CCIT - Calgary Centre for Innovative Technology
CD - Cascade Hall
CDC - Child Development Centre
CH - Craigie Hall C - G (University Theatre)
CR - Crowsnest Hall
DC - Dining Centre
EDC - Education Classroom Block
EDT - Education Tower
EEEL - Energy. Environment. Experiential Learning

EN - Schulich School of Engineering A - G

ERR - Energy Resource Research
ES - Earth Science
GL - Glacier Hall
GS - General Service Warehouse
GR - Grounds
HMRB - Heritage Medical Research Building
HP - Central Heating & Cooling Plant
HRIC - Health Research Innovation Centre
HSC - Health Science Centre

ICT - Information & Communications Technologies

IH - International House & Hotel Alma
KA - Kananaskis Hall
KNA - Kinesiology A
KNB - Kinesiology B- (Jack Simpson Gym)
MEB - Mechanical Eng Building
MFH - Murray Fraser Hall
MH - MacEwan Hall
MB - MacKimmie Library Block
MT - MacKimmie Library Tower

MS - Math Science

MSC - MacEwan Student Centre
MF - Materials Handling Facility
OL - Olympus Hall
OO - Olympic Oval
OVC - Olympic Volunteer Centre
PF - Professional Faculties
PP - Physical Plant
RC - Rozsa Centre
RT - Reeve Theatre
RU - Rundle Hall

SA - Science A Building
SB - Science B Building

SH - Scurfield Hall

SS - Social Science
ST - Science Theatres

TFDL - Taylor Family Digital Library
TI - Taylor Institute for Teaching & Learning
TRA - Trailer A
TRB - Trailer B
TRW - Teaching Research & Wellness
URC - University Research Centre
VC - Varsity Courts (Family Housing)
WS - Weather Station

YA Yamnuska Hall

UNIVERSITY OF CALGARY