Société statistique du Canada
Statistical Society of Canada

49th Annual Meeting of the Statistical Society of Canada

49e Congrès annuel de la Société statistique du Canada

May 29 – June 5, 2022
29 mai au 5 juin 2022

Virtual - Virtuel

Table of Contents • Table des matières

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

Welcome • Bienvenue **2**

Message from the SSC President • Message de la Presidente de la SSC **3**

Sponsors • Commanditaires **4**

Job Fair • Foire à l'emploi **4**

Organizers • Organisateurs **5**

The Conference • Le congrès **7**

Social Events • Événements sociaux **8**

Workshops • Ateliers **9**

Scientific Program • Programme scientifique **11**

Abstracts • Résumés **60**

Author List • Liste des auteurs **326**

Welcome • Bienvenue

Welcome to SSC2022! Due to the ongoing COVID-19 pandemic and its restrictions we are once again having a completely virtual conference. To accommodate the time zones of our vast country while providing a full conference experience, we have gone to a 5-day conference starting Monday May 30 to Friday June 3 and running from 11 a.m. to 7 p.m. EDT. Please note that online the timing of each session will automatically adjust to your time zone. Like last year, Workshops are again being held separately from this conference. Workshops are scheduled on May 29, June 4 and June 5 and you must register for them. Workshop registrants will receive separate Zoom links to use to attend their workshops.

The virtual organizing committee (VOC) has organized this year's conference along the lines of last year's and are again using Pathable as the platform host for virtual SSC2022. Since there may be some glitches in our second offering of a complex virtual conference; we hope you will be understanding and that your experience will overall be positive. In addition to scientific sessions, there are some social events, a virtual Job Fair, and an Awards ceremony. Registrants will also have access to all conference recorded sessions for several months following the conference. The VOC would like to thank all those whose efforts are making this second virtual conference happen. We especially thank the SSC Bilingualism Committee, the Scientific Program Committee, the SSC Program Committee and also Michelle Benoit and Marie-Pierre Nantel of the SSC Office and Larysa Valachko for their valuable support of all aspects including registration.

Bienvenue à SSC2022 ! En raison de la pandémie de COVID-19 en cours et de ses restrictions, nous organisons à nouveau une conférence entièrement virtuelle. Pour s'adapter aux fuseaux horaires de notre vaste pays tout en offrant une expérience de conférence complète, nous avons organisé une conférence de 5 jours du lundi 30 mai au vendredi 3 juin et se déroulant de 11 h à 19 h. EDT. Veuillez noter qu'en ligne, l'horaire de chaque session s'ajustera automatiquement à votre fuseau horaire. Comme l'année dernière, les ateliers se tiennent à nouveau séparément de cette conférence. Les ateliers sont prévus les 29 mai, 4 juin et 5 juin et vous devez vous y inscrire. Les personnes inscrites aux ateliers recevront des liens Zoom distincts à utiliser pour assister à leurs ateliers.

Le comité d'organisation virtuel (VOC) a organisé la conférence de cette année sur le modèle de l'année dernière et utilise à nouveau Pathable comme plate-forme hôte pour le SSC2022 virtuel. Puisqu'il peut y avoir quelques problèmes dans notre deuxième offre d'une conférence virtuelle complex ; nous espérons que vous serez compréhensif et que votre expérience sera globalement positive. En plus des sessions scientifiques, il y a des événements sociaux, un salon de l'emploi virtuel et une cérémonie de remise des prix. Les inscrits auront également accès à toutes les sessions enregistrées de la conférence pendant plusieurs mois après la conférence. Le COV tient à remercier tous ceux dont les efforts permettent à cette deuxième conférence virtuelle d'avoir lieu. Nous remercions particulièrement le comité du bilinguisme de la SSC, le comité du programme scientifique, le comité du programme de la SSC ainsi que Michelle Benoit et Marie-Pierre Nantel du bureau de la SSC et Larysa Valachko pour leur précieux soutien dans tous les aspects, y compris l'inscription.

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

Dear colleagues, students, friends, and participants:

On behalf of the Program Committee and Virtual Organizing Committee, I am thrilled to welcome you to the annual signature event of the Statistical Society of Canada (SSC) – the 49th SSC Annual Meeting (SSC2022).

2022 is undoubtedly a unique year. It is the 50th anniversary of the statistical community in Canada. We are excited to gather to exchange statistical ideas, disseminate new research results, share experiences, foster connections, and make new acquaintances.

The meeting will be held online from May 30 to June 3, 2022, preceded by a one-day student conference on May 28, 2022. The SSC2022 features a wide range of activities, primarily including a large number of scientific sessions accompanied by panel discussions, case studies, social networking opportunities, information sessions, a job fair, and the award ceremony night. As usual, plenary speeches will be delivered by the Presidential Invited Addressee and award winners. In addition, four invited sessions are dedicated to celebrating our society's 50th anniversary.

This conference would not have been possible without diligent work from many volunteers working behind the scene. Sincere thanks go to everyone who helped bring together this conference and the sponsors who supported the conference.

Finally, I extend my best wishes for a successful and fruitful conference!

Enjoy!

Grace Y. Yi
SSC President

Chers collègues, étudiants, amis et participants :

Au nom du comité du programme et du comité organisateur virtuel, je suis ravi de vous accueillir à l'événement phare annuel de la Société statistique du Canada (SSC) - le 49e congrès annuel de la SSC (SSC2022).

2022 est sans aucun doute une année unique. C'est le 50e anniversaire de la communauté statistique au Canada. Nous sommes ravis de nous réunir pour échanger des idées statistiques, diffuser de nouveaux résultats de recherche, partager des expériences, favoriser les relations et faire de nouvelles connaissances.

La réunion se tiendra en ligne du 30 mai au 3 juin 2022, précédée d'une conférence étudiante d'une journée le 28 mai 2022. Le SSC2022 propose un large éventail d'activités, comprenant principalement un grand nombre de sessions scientifiques accompagnées de tables rondes, des études de cas, des opportunités de réseautage social, des séances d'information, un foire de recrutement et la soirée de remise des prix. Comme d'habitude, les allocutions plénières seront prononcées par le destinataire invité du président et les lauréats. De plus, quatre sessions invitées sont consacrées à la célébration du 50e anniversaire de la société.

Cette conférence n'aurait pas été possible sans le travail assidu de nombreux bénévoles travaillant en coulisses. Des remerciements sincères vont à tous ceux qui ont aidé à organiser cette conférence et aux sponsors qui ont soutenu la conférence.

Enfin, je vous présente mes meilleurs vœux pour une conférence réussie et fructueuse !

Profitez !

Grace Y.Yi
Présidente de la SSC

Sponsors • Commanditaires

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

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

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

Job Fair • Foire à l'emploi

In order to assist job seekers, we have arranged a Job Fair during SSC2022. You are encouraged to visit the Job Fair tab at this website for more details about employers and available positions and ways to submit a c.v. You will be able meet individually during the conference with employers to discuss job opportunities. There is no registration fee to participate in the Job Fair.

Afin d'aider les demandeurs d'emploi, nous avons organisé une foire de l'emploi pendant SSC2022. Nous vous encourageons à visiter l'onglet Foire de l'emploi de ce site Web pour plus de détails sur les employeurs, les postes disponibles et les moyens de soumettre un c.v. Vous pourrez rencontrer individuellement pendant la conférence les employeurs pour discuter des possibilités d'emploi. Il n'y a pas de frais d'inscription pour participer à la Foire de l'emploi.

Organizers • Organisateurs

Virtual Organizing Committee • Comité d'organisation virtuel

- Shirley Mills (Co-Chair • Co-Président)
- Richard Lockhart (Co-Chair • Co-Président)
- Pengfei Li
- Angelo Canty
- Wendy Lou
- Asokan M. Variyath
- Michelle Benoit
- Marie-Pierre Nantel
- Larysa Valachko

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 Virtual Organizing Committee would like to thank all those who helped pull this event together.

The SSC Meetings Coordinator Nadia Ghazzali and other SSC executive members, and previous local arrangements chairs all shared their experience, offered useful advice and answered our numerous questions. Finally, Angelo Canty managed 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 d'organisation virtuel est très reconnaissant à tous ceux qui ont aidé à mettre sur pied cet événement.

Le coordonnateur des congrès Nadia Ghazzali 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, Angelo Canty a géré les services électroniques reliés au congrès et préparé la version PDF du programme.

Program Committee • Comité du programme

- Pengfei Li (Chair • Co-Président)
- Andrei Badescu
- Rob Deardon
- Jean-François Plante
- Nathaniel Stevens
- Éric Marchand
- Bruce Dunham
- Jean-François Beaumont
- Félix Camirand Lemyre
- Thérèse Stukel
- Beatrice Baribeau

The Conference • Le congrès

Time Zone • Fuseau horaire

All times in this program are given in Eastern Daylight Savings (Ottawa) Time. Note that this is different from the Pathable website where the times are adjusted to your local time zone.

Toutes les heures de ce programme sont exprimées en heure avancée de l'Est (Ottawa). Notez que ceci est différent du site Web Pathable où les heures sont ajustées à votre fuseau horaire local.

Language • Langue

All possible efforts have been made to incorporate bilingualism in the organization of this virtual conference. In all plenary and award sessions, attendees have the option of selecting English or French slides to view during the session. While the conference website supported by Pathable is built in English, maximum effort has been made to add French to each navigation tab. Also, a complete program schedule with abstracts in both official languages is available at the SSC2022 meeting site on ssc.ca.

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.

Tous les efforts possibles ont été déployés pour intégrer le bilinguisme dans l'organisation de cette conférence virtuelle. Dans toutes les sessions plénières et de remise des prix, les participants ont la possibilité de sélectionner des diapositives en anglais ou en français à visionner pendant la session. Bien que le site Web de la conférence pris en charge par Pathable soit construit en anglais, un maximum d'efforts a été fait pour ajouter le français à chaque onglet de navigation. De plus, un calendrier complet du programme avec des résumés dans les deux langues officielles est disponible sur le site de la réunion SSC2022 sur ssc.ca.

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.

Workshops • Ateliers

Workshops organized by the sections will be held on Sunday May 29 (Accreditation Committee, Business and Industrial Statistics Section, Statistical Education Section), Saturday June 4 (Data Science and Analytics Section, Survey Methods Section) and Sunday June 13 (Biostatistics Section, Probability Section). These workshops are not part of the main conference on Pathable but are being run separately using Zoom meetings. Registered participants will receive an email with the details of the meetings.

Les ateliers organisés par les groupes auront lieu le dimanche 6 juin (Comité d'accréditation, Groupe de statistique industrielle et de gestion, Groupe d'éducation en statistique), samedi 4 juin (Groupe de science des données et analytiques, Groupe des méthodes d'enquête) et dimanche 13 juin (Groupe de biostatistique, Groupe de probabilité). Ces ateliers ne font pas partie de la conférence principale sur Pathable mais sont organisés séparément à l'aide de réunions Zoom. Les participants inscrits recevront un courriel avec les détails des réunions.

Social Events • Événements sociaux

Monday May 30 — lundi 30 mai

12:30-13:30
Statistical Education Section Social Gathering/Groupe d'éducation en statistique rassemblement social

---

17:00-19:00
Student and Recent Graduate Social Evening/Soirée sociale des étudiants et diplômés récents

---

17:30-18:30
Women in Statistics Committee Social Gathering/Réunion sociale du Comité des femmes en statistique

Tuesday May 31 — mardi 31 mai

17:00-18:00
Accreditation Committee Social Gathering/Rassemblement social du Comité d'accréditation

---

17:00-19:00
SSC and Me/SSC et moi

Wednesday June 1 — mercredi 1 juin

17:00-18:00
McGill Biostatistics Social Gathering/Rassemblement social de biostatistique de McGill

---

17:00-18:00
University Gathering: University of Toronto Biostatistics/Rassemblement universitaire: University of Toronto biostatistique

Thursday June 2 — jeudi 2 juin

17:00-19:00
SSC Conference Award Ceremony/Cérémonie de remise des prix de la conférence SSC

Friday June 3 — vendredi 3 juin

17:00-19:00
Closing Ceremony/Cérémonie de cloture

Workshops • Ateliers

Sunday May 29 — dimanche 29 mai

11:00-18:00 — **Workshop / Atelier** (abstract/résumé 61)

Accreditation Workshop
Atelier du Comité d'accréditation

Chair/Président: Fernando Camacho
Organizer/Responsable: Fernando Camacho
Sponsor/Commanditaires: Accreditation Committee / Le Comité d'accréditation

11:00-18:00 **Peter Solymos** (E Source) **Khalid Lemzouji** (Worley)
Delivering applied statistics from concept to production / Fournir des statistiques appliquées, du concept à la production

---

13:00-16:00 — **Workshop / Atelier** (abstract/résumé 63)

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

Chair/Président: Jean-François Plante
Organizer/Responsable: Jean-François Plante
Sponsor/Commanditaires: Business and Industrial Statistics Section / Le Groupe de statistique industrielle et de gestion

13:00-16:00 **Sarah Legendre Bilodeau** (Videns Analytics) **Sébastien Duguay** (Videns Analytics)
Kubernetes, containers and the cloud: an overview of the tools and challenges to put models in production / Kubernetes, conteneurs et cloud : tour d'horizon des outils et défis pour mettre des modèles en production

---

13:30-17:00 — **Workshop / Atelier** (abstract/résumé 64)

Statistical Education Workshop
Atelier du Groupe d'éducation en statistique

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

13:30-17:00 **Tiffany A. Timbers** (The University of British Columbia) **Wesley Burr** (Trent University)
Reproducibility Workshop / Atelier sur la reproductibilité

Saturday June 4 — samedi 4 juin

12:30-17:00 — **Workshop / Atelier** (abstract/résumé 324)

Survey Methods Workshop
Atelier du Groupe des méthodes d'enquête

Chair/Président: Jean-François Beaumont
Organizer/Responsable: Jean-François Beaumont
Sponsor/Commanditaires: Survey Methods Section / Le Groupe des méthodes d'enquête

12:30-17:00 **Changbao Wu** (University of Waterloo)
From Sample Surveys to Missing Data and Causal Inference / Des enquêtes par sondage aux données manquantes et à l'inférence causale

---

13:00-16:30 — **Workshop / Atelier** (abstract/résumé 325)

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

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

13:00-16:30 **Rodolfo Lourenzutti** (University of British Columbia) **Arman Seyed-Ahmadi** (University of British Columbia) **Diego Ardila** (Shopify)
Intro to Databases in Industry: Data Cleaning, Querying, and Modeling at Scale / Introduction aux bases de données en industrie : nettoyage des données, interrogation et modélisation à grande échelle

Sunday June 5 — dimanche 5 juin

11:00-17:00 — **Workshop / Atelier** (abstract/résumé 67)

Probability Workshop
Atelier du Groupe de Probabilité

Chair/Président: Ting Kam Leonard Wong
Organizer/Responsable: Ting Kam Leonard Wong
Sponsor/Commanditaires: Probability Section / Le Groupe de Probabilité

11:00-17:00 **Ting Kam Leonard Wong** (University of Toronto) **Jun Zhang** (University of Michigan) **Paul Marriott** (University of Waterloo) **Guido Montufar** (University of California, Los Angeles) **Gabriel Khan** (Iowa State University) **Melvin Loek** (University of California, San Diego) **Tian Han** (Stevens Institute of Technology) **Wuchen Li** (University of South Carolina)
Information geometry and applications / Géométrie de l'information et applications

---

12:00-15:30 — **Workshop / Atelier** (abstract/résumé 68)

Biostatistics Workshop
Atelier du Groupe de biostatistique

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

12:00-15:30 **Jessica Gronsbell** (University of Toronto)
Electronic Health Records Phenotyping / Phénotypage des dossiers médicaux électroniques

Scientific Program • Programme scientifique

Monday May 30 lundi 30 mai

11:00-12:30 **Invited / Sur invitation** (abstract/résumé 70)

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

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

11:00-12:00 **Anthony Davison** (École polytechnique fédérale de Lausanne)
How Long Could a Human Live? / Y a-t-il une durée maximale de longévité humaine ? E EF

12:30-13:30 **Poster / Poster** (abstract/résumé 71)

Contributed Posters
Affiches contribuées

12:30-13:00 **Jizhou Tian** (Lady Davis Institute for Medical Research, Jewish General Hospital) **Yi Liu** (Lady Davis Institute for Medical Research, Jewish General Hospital, Montreal, Quebec, Canada) **Andrea Benedetti** (McGill University)
An Empirical Comparison of the Two-Stage and One-Stage Bootstrap Approaches in the Context of an Individual Participant Data Meta-Analysis / Comparaison empirique d'approches bootstrap à deux étapes et à une étape dans un contexte de méta-analyse de données individuelles de participants E E

12:30-13:00 **Xi Zhang** (McMaster University) **Orla Murphy** (Dalhousie University) **Paul D. McNicholas** (McMaster University)
Longitudinal Data Clustering with a Copula Kernel Mixture Model / Regroupement de données longitudinales avec modèle de mélange à noyau de copules E E

12:30-13:00 **Sidi Wu** (Simon Fraser University) **Cédric Beaulac** (Simon Fraser University) **Jiguo Cao** (Simon Fraser University)
Neural Networks with Functional Response / Réseaux neuronaux avec réponse fonctionnelle E E

12:30-13:00 **Ruwan C. Karunanayaka** (University of the Fraser Valley) **Boxin Tang** (Simon Fraser University)
On the Existence and Constructions of Orthogonal Designs / De l'existence et des constructions de plans orthogonaux E E

13:00-13:30 **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 d'importance séquentiel E E

13:00-13:30 **Jingjun (Victor) Chen** (McGill University) **Andrea Benedetti** (McGill University) **Zelalem F. Negeri** (McGill University; Lady Davis Institute for Medical Research, Jewish General Hospital) **Brett D. Thombs** (McGill University; Lady Davis Institute for Medical Research, Jewish General Hospital)
Individual Participant Data Meta-Analyses Using Bivariate Random Effect Models / Méta-analyses des données des participants individuels à l'aide de modèles à effets aléatoires bivariés E E

13:00-13:30 **Timofei Biziaev** (University of Calgary)
Comparison of Frequentist and Bayesian Approaches to Ordinal Regression Model Validation / Comparaison des approches fréquentiste et bayésienne pour la validation de modèle de régression ordinal E E

13:00-13:30 **Mengjie Bian** (McMaster University) **Angelo J. Canty** (McMaster University)
Identification of Invalid Genetic Variants in Mendelian Randomization / Identification de variants génétiques invalides dans une randomisation mendélienne E E

13:30-15:00 **Invited / Sur invitation** (abstract/résumé 76)

New Methods for Structured Variable Selection
Nouvelles méthodes de sélection structurée de variables

Chair/Président: Mireille Schnitzer
Organizer/Responsable: Guanbo Wang

13:30-13:52 **Marie Denis** (Centre de coopération internationale en recherche agronomique pour le développement) **Mahlet G. Tadesse** (Georgetown University)
Graph-Structured Variable Selection with Gaussian Markov Random Field Horseshoe Prior / Sélection de variables structurées en graphe avec un a priori en fer à cheval de champs aléatoires gaussiens de Markov

13:52-14:14 **Guanbo Wang** (McGill University) **Mireille Schnitzer** (Université de Montréal) **Tom Chen** (Harvard Pilgrim Health Care Institute and Harvard Medical School) **Rui Wang** (Harvard Pilgrim Health Care Institute and Harvard Medical School) **Robert Platt** (McGill University)
A general framework for identification of permissible variable subsets in structured model selection / Cadre général d'identification des sous-ensembles de variables admissibles dans la sélection structurée de modèles

14:14-14:36 **Yi Yang** (McGill University) **Yuwen Gu** (University of Connecticut) **Yue Zhao** (University of York) **Jun Fan** (McGill University)
Flexible Regularized Estimating Equations: Some New Perspectives / Nouvelles perspectives d'équations d'estimation régularisées souples

14:36-14:58 **Tingting Yu** (Harvard Medical School and Harvard Pilgrim Health Care Institute)
Variable selection in high dimensional linear regression accounting for heterogeneity in covariate effects across multiple data sources / Sélection de variables dans la régression linéaire de haute dimension tenant compte de l'hétérogénéité des effets des covariables dans les sources de données multiples

13:30-15:00 **Invited / Sur invitation** (abstract/résumé 79)

Statistical Disclosure Control Methods for Privacy
Méthodes de contrôle de la divulgation statistique et vie privée

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

13:30-13:55 **Fang Liu** (University of Notre Dame)
Utility Analysis of Differentially Private Gradient-based Optimization Algorithms / Analyse d'utilité d'algorithmes d'optimisation différentiellement confidentiels à base de gradients

13:55-14:20 **Hui Xie** (Simon Fraser University) **Yi Qian** (University of British Columbia)
Fast Distribution-free Statistical Control Methods to Construct Large-scale Privacy-preserving Databases / Méthodes statistique rapides et libre du contrôle de la distribution pour construire des bases de données à grande échelle préservant la confidentialité

14:20-14:45 **Liu Yi** (University of Alberta) **Ke Sun** (University of Alberta) **Bei Jiang** (University of Alberta) **Linglong Kong** (University of Alberta)
A Bridge to Gaussian Differential Privacy / Un pont vers la confidentialité différentielle gaussienne

14:45-15:00 **Bei Jiang** (University of Alberta)
Discussion / Discussion

13:30-15:00 **Invited / Sur invitation** (abstract/résumé 81)

Use of Machine Learning Methods for Handling Missing Survey Data
Utilisation de méthodes d'apprentissage automatique pour le traitement des données d'enquête manquantes

Chair/Président: Keven Bosa, Jean-François Beaumont
Organizer/Responsable: Keven Bosa
Sponsor/Commanditaires: Survey Methods Section / Le Groupe des méthodes d'enquête

13:30-14:00 **Mehdi Dagdoug** (Université de Bourgogne Franche-Comté) **Camelia Goga** (Université de Bourgogne Franche-Comté) **David Haziza** (University of Ottawa)
Model-Assisted Estimation with Machine Learning Methods in High-Dimensional Settings for Survey Data / Estimation assistée par modèle avec des méthodes d'apprentissage automatique dans des contextes de grande dimension pour les données d'enquête Ⓔ E

14:00-14:30 **Sixia Chen** (University of Oklahoma Health Sciences Center) **Chao Xu** (University of Oklahoma Health Sciences Center)
Handling High Dimensional Data with Missing Values by Modern Machine Learning Techniques / Traitement de données de haute dimension avec valeurs manquantes par des techniques modernes d'apprentissage automatique Ⓔ E

14:30-15:00 **Olanrewaju Michael Akande** (Duke University) **Zhenhua Wang** (University of Missouri) **Jason Poulos** (Harvard Medical School) **Fan Li** (Duke University)
Are Deep Learning Models Superior for Missing Data Imputation in Surveys? Evidence from an Empirical Comparison / Les modèles d'apprentissage profond permettent-ils une meilleure imputation des données manquantes dans les enquêtes ? Résultats d'une comparaison empirique Ⓔ E

13:30-15:00 **Invited / Sur invitation** (abstract/résumé 83)

Isobel Loutit Invited Address
Allocution Isobel-Loutit

Chair/Président: Hugh Chipman
Organizer/Responsable: Jean-François Plante
Sponsor/Commanditaires: Business and Industrial Statistics Section / Le Groupe de statistique industrielle et de gestion

13:30-14:30 **Jeff Wu** (Georgia Tech)
My Five Years in Canada: How it Impacted my Work and the Field of Experimental Design / Mes cinq années au Canada : leur impact sur mon travail et le domaine de la conception expérimentale Ⓔ E

13:30-15:00 **Invited / Sur invitation** (abstract/résumé 84)

Actuarial Applications in Finance
Applications actuarielles en finance

Chair/Président: Maciej Augustyniak
Organizer/Responsable: Maciej Augustyniak
Sponsor/Commanditaires: Actuarial Science Section / Le Groupe de science actuarielle

13:30-14:00 **Alexandru Badescu** (University of Calgary) **Maciej Augustyniak** (Université de Montréal) **Jean-François Bégin** (Simon Fraser University) **Sarath Kumar Jayaraman** (University of Calgary)
Long Memory in Option Pricing: A Fractional Discrete-Time Framework / Mémoire longue pour la tarification des options : un cadre fractionnaire en temps discret Ⓔ E

14:00-14:30 **Jean-François Bégin** (Simon Fraser University)
On Complex Economic Scenario Generators: Is Less More? / Sur les générateurs de scénarios économiques complexes : Moins, c'est mieux ? Ⓔ E

14:30-15:00 **Anne Mackay** (Université de Sherbrooke) **Michael A. Kouritzin** (University of Alberta)
On stochatic approximation and option pricing / Tarification d'options via un algorithme d'approximation stochastique Ⓔ E

13:30-15:00 **Contributed / Communications libres** (abstract/résumé 86)

Stochastic Processes, Monte Carlo Integration, and AFT Model
Processus stochastiques, intégration Monte Carlo et modèle de temps de défaillance accéléré

Chair/Président: Adam B. Kashlak

13:30-13:45 **Mamadou Yamar Thioub** (HEC Montréal) **Bouchra Nasri** (Université de Montréal) **Bruno Rémillard** (HEC Montréal)
Goodness-of-fit Tests and Robust Regime Selection Procedure for General Hidden Markov Models / Tests d'adéquation et procédure robuste de sélection de régimes pour les modèles de Markov cachés

13:45-14:00 **Sabrina Sixta** (University of Toronto)
Convergence Rate Bounds for Iterative Random Functions Using One-Shot Coupling / Limites de taux de convergence pour les fonctions aléatoires itératives utilisant le couplage à un coup

14:00-14:15 **Dinh-Toan Nguyen** (Université du Québec à Montréal)
Scaling Limit of the Collision Measures of Multiple Random Walks / Limite d'échelle des mesures de collision de plusieurs marches aléatoires

14:15-14:30 **Yanbo Tang** (University of Toronto)
Monte Carlo Integration in High Dimensions / L'intégration de Monte-Carlo en hautes dimensions

14:30-14:45 **Weinan Qi** (University of Waterloo) **Paul Marriott** (University of Waterloo) **Yi Shen** (University of Waterloo)
Excursion sets and critical points of Gaussian random fields over high thresholds / Ensembles d'excursions et points critiques des champs aléatoires gaussiens au-delà de seuils élevés

14:45-15:00 **Quinn Forzley** (University of Winnipeg) **Shakhawat Hossain** (University of Winnipeg)
Pretest and Shrinkage Estimation Strategies in Accelerated Failure Time Model / Stratégies d'estimation de rétrécissement et de prétests dans un modèle à temps d'échec accéléré

13:30-15:00 **Contributed / Communications libres** (abstract/résumé 89)

Identifying and Utilizing Group Structures in Heterogeneous Populations
Identification et utilisation des structures de groupe dans les populations hétérogènes

Chair/Président: Kevin McGregor

13:30-13:45 **Gyanendra Pokharel** (The University of Winnipeg)
Classification-Based Inference for Spatial Infectious Disease Models Incorporating Infection Time Uncertainty / Inférence basée sur une classification pour des modèles spatiaux de maladies infectieuses incorporant une incertitude du temps d'infection

13:45-14:00 **Wangshu Tu** (Carleton University) **Sanjeena Subedi** (Carleton University) **Ryan P. Browne** (University of Waterloo)
Mixtures of Logistic Skew-normal Multinomial Models / Mélanges de modèles logistiques multinomiaux asymétriques

14:00-14:15 **Dexen D.Z. Xi** (Western University) **Masoud Adelzadeh** (National Research Council Canada)
Finite Mixture Models and Shared Frailty Models for Fire Department Response Time in Building Fires / Modèles de mélanges finis et modèles à fragilités partagées du temps de réponse des services d'incendie en cas de feu dans un bâtiment

14:15-14:30 **Xiaoke Qin** (Carleton University) **Sanjeena Dang** (Carleton University)
Mixtures of Generalized Dirichlet-Multinomial Models for Microbiome Data / Mélanges de modèles multinomiaux généralisés de Dirichlet pour des données sur le microbiome

14:30-14:45 **Zihang Lu** (Queen's University) **Wendy Lou** (University of Toronto)
A Model-Based Approach for Clustering Developmental Trajectories with Complex Longitudinal Data / Approche basée sur un modèle pour le regroupement de trajectoires de développement avec des données longitudinales complexes

14:45-15:00 **Aida Eslami** (Université Laval) **Hervé Abdi** (School of Behavioral and Brain Sciences, The University of Texas at Dallas)
Integrating Group Structure in the Multiple Correspondence Analysis / Intégration de la structure de groupe dans l'analyse des correspondances multiples

---

13:30-15:00 **Contributed / Communications libres** (abstract/résumé 93)

Statistics Education
Éducation en statistique

Chair/Président: Suborna Shekhor Ahmed

13:30-13:45 **Douglas Whitaker** (Mount Saint Vincent University)
Investigation of Bivariate Grid-Type Items for Measuring Attitudes in Statistics Education: Preliminary Results / Items de type grille bivariée pour mesurer les attitudes en enseignement de la statistique : résultats préliminaires

13:45-14:00 **Tharshanna Nadarajah** (University of Toronto)
Teaching Through Collaboration / Enseigner par la collaboration

14:00-14:15 **Diana Katherine Skrzydlo** (University of Waterloo)
Designing Authentic Assessments for Learning / Concevoir des évaluations authentiques pour l'apprentissage

14:15-14:30 **Nathalie Moon** (University of Toronto) **Liza Bolton** (University of Toronto) **Rebecca Christensen** (University of Toronto)
Reflective Writing in Statistics Courses / L'écriture réflexive dans les cours de statistiques

14:30-14:45 **Nooshin Khobzi Rotondi** (Ontario Tech University) **David Rudoler** (Ontario Tech University) **William Hunter** (Ontario Tech University) **Olayinka Sanusi** (Ontario Tech University) **Chris Collier** (Ontario Tech University) **Michael Rotondi** (York University)
Using a "midterm warning system" to improve student performance and engagement in an introductory statistics course: A randomized controlled trial / L'utilisation d'un « système d'avertissement à mi-parcours » pour améliorer les performances et l'engagement des étudiants dans un cours d'introduction aux statistiques : un essai comparatif aléatoire

14:45-15:00 **Melanie C. H. Gibbons** (University of Saskatchewan) **Marc T. Avey** (Canadian Council on Animal Care) **Phyllis G. Paterson** (University of Saskatchewan)
Experimental Design and Statistics Training in Select Canadian Graduate Programs at U15 Universities / Formation en planification d'expérience et statistique dans certains programmes canadiens d'études supérieures des universités U15

15:30-17:00 **Invited / Sur invitation** (abstract/résumé 97)

Statistical Analysis of Complex Large-Scale Health Data
Analyse statistique des données complexes à grande échelle sur la santé

Chair/Président: Peter X. Song
Organizer/Responsable: Peter X. Song

15:30-16:00 **Ji Zhu** (University of Michigan)
Fast Network Community Detection with Profile-Pseudo Likelihood Methods / Détection communautaire à réseau rapide avec des méthodes de pseudo vraisemblance profilée

16:00-16:30 **Bin Nan** (University of California, Irvine) **Yue Wang** (University of California) **Jack Kalbfleisch** (University of Michigan)
Kernel Estimation of Bivariate Time-varying Coefficient Model for Longitudinal Data with Terminal Event / Estimation par noyau d'un modèle bivarié à coefficients variant dans le temps pour données longitudinales avec événement terminal

16:30-17:00 **Annie Qu** (University of California, Irvine)
Optimal Individualized Omni-channel Treatment Decision Rule Under Budget Constraints / Règle décisionnelle de traitement omnicanal individualisé optimal selon des contraintes budgétaires

---

15:30-17:00 **Invited / Sur invitation** (abstract/résumé 100)

New Statistical Methods for Adaptive Clinical Trial Design
Nouvelles méthodes statistiques pour la conception d'essais cliniques adaptatifs

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

15:30-16:00 **Ying Yuan** (University of Texas MD Anderson Cancer Center)
Elastic Priors to Dynamically Borrow Information from Historical Data in Clinical Trials / Distributions a priori élastiques pour l'emprunt dynamique d'information tirée de données historiques d'essais cliniques

16:00-16:30 **Suyu Liu** (MD Anderson Cancer Center) **Beibei Guo** (Louisiana State University) **Elizabeth Garrett-Mayer** (American Society of Clinical Oncology)
A Bayesian Phase I/II Design for Cancer Clinical Trials Combining Immunotherapy and Chemotherapy / Plan bayésien de phase I/II pour les essais cliniques de traitement du cancer combinant l'immunothérapie et la chimiothérapie

16:30-17:00 **Depeng Jiang** (University of Manitoba) **Bosheng Li** (University of Manitoba) **Fangrong Yan** (China Pharmaceutical University)
A Bayesian Adaptive Design for an Immunotherapy with Heterogeneous Delayed Treatment Effect / Plan adaptatif bayésien pour une immunothérapie à effet tardif du traitement hétérogène

---

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

Leadership and Women in Statistics
Leadership et femmes en statistique

Chair/Président: Thérèse A. Stukel
Organizer/Responsable: Thérèse A. Stukel
Sponsor/Commanditaires: Women in Statistics Committee / Le Comité des femmes en statistique

15:30-17:00 **Charmaine B. Dean** (University of Waterloo) **Nadia Ghazzali** (Université du Québec à Trois-Rivières) **Amanda Golbeck** (University of Arkansas for Medical Sciences) **Lisa Strug** (University of Toronto)
Leadership and Women in Statistics / Leadership et femmes en statistique

---

15:30-17:00 **Invited / Sur invitation** (abstract/résumé 103)

Fifty Years of Statistics Teaching
Cinquante ans d'enseignement de la statistique

Chair/Président: Becky Wei Lin
Organizer/Responsable: Becky Wei Lin
Sponsor/Commanditaires: Statistical Education Section / Le Groupe d'éducation en statistique

15:30-16:00 **Kenneth Laurence Weldon** (Simon Fraser University)
Teaching the Big Ideas of Statistics / L'enseignement des grands concepts statistiques

16:00-16:30 **Bethany J.G. White** (University of Toronto)
Leveraging Technology in Statistics Education: A Look at Developments since 2000 / Tirer parti de la technologie dans l'enseignement des statistiques : un regard sur les développements depuis 2000 E E

16:30-17:00 **Jerald F. Lawless** (University of Waterloo)
Statistics Education 1972-2022: Some Past Developments and A Look Ahead / L'enseignement des statistiques de 1972 à 2022 : évolution passée et perspectives d'avenir E E

15:30-17:00 **Invited / Sur invitation** (abstract/résumé 105)

Change-point Detection
Détection des points de changement

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

15:30-16:00 **Bouchra Nasri** (Université de Montréal) **Bruno Rémillard** (HEC Montréal) **Tarik Bahroui**
Change-Point Problems for Multivariate Time Series Using Pseudo-Observations / Problèmes de points de rupture pour les séries chronologiques multivariées utilisant des pseudo-observations E E

16:00-16:30 **Sévérien Nkurunziza** (University of Windsor)
Some Inference Problems in Generalized Ornstein-Uhlenbeck Processes with Change-Points / Quelques problèmes d'inférence dans les processus d'Ornstein-Uhlenbeck généralisés avec des points de rupture E E

16:30-17:00 **Zhou Zhou** (University of Toronto) **Weichi Wu** (Tsinghua University)
Multiscale Jump Testing and Estimation Under Complex Temporal Dynamics / Test et estimation multi-échelles de sauts sous dynamique temporelle complexe E E

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

Graduate Research in Actuarial Science 1
Recherche aux cycles supérieures en science actuarielle 1

Chair/Président: Yi Lu
Sponsor/Commanditaires: Actuarial Science Section / Le Groupe de science actuarielle

15:30-15:45 **Sebastian F Calcetero** (University of Toronto)
A functional Severity Regression Model for Applications in General Insurance / Modèle de régression fonctionnel de la sévérité pour des applications en assurance générale E E

15:45-16:00 **Sébastien Jessup** (Concordia University) **Mélina Mailhot** (Concordia University) **Mathieu Pigeon** (Université du Québec à Montréal)
On the Impact of Model Combination Methods on Extreme Precipitation Projections / L'impact de combinaisons de modèles sur les projections de précipitations extrêmes E E

16:00-16:15 **Mingren Yin** (University of Waterloo)
Optimal Deductible Reinsurance with Model Uncertainty / Réassurance avec franchise optimale et incertitude du modèle E E

16:15-16:30 **Meng Sun** (Simon Fraser University) **Yi Lu** (Simon Fraser University)
Statistical Modeling of Data Breaches and its Application in Cyber Insurance / Modélisation statistique des fuites de données et son application dans la cyberassurance E E

16:30-16:45 **Christopher Blier-Wong** (Université Laval) **Hélène Cossette** (Université Laval) **Marceau Etienne** (Université Laval)
Risk Aggregation with FGM Copulas / Agrégation des risques avec copules FGM E E

16:45-17:00 **Pouya Faroughi** (Western University) **Shu Li** (Western university) **Jiandong Ren** (Western university)
Generalized Poisson Distribution and Its Application in Actuarial Science / Distribution de Poisson généralisée et application en sciences actuarielles E E

15:30-17:00 **Contributed / Communications libres** (abstract/résumé 110)

Causal Inference
Inférence causale

Chair/Président: Denis Talbot

15:30-15:45 **Shuo Sun** (McGill University) **Johanna G. Nešlehová** (McGill University) **Erica E.M. Moodie** (McGill University)
Principal Stratification for Quantile Causal Effects under Partial Compliance: an Analysis of COVID-19 Case Counts / Stratification principale des effets causaux sur les quantiles en cas de conformité partielle : analyse du nombre de cas de COVID-19

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

16:00-16:15 **Jingyue Huang** (University of Waterloo) **Leilei Zeng** (University of Waterloo) **Changbao Wu** (University of Waterloo)
Pseudo-Empirical Likelihood Approach for the Estimation of Average Treatment Effect / Approche de vraisemblance pseudo-empirique pour l'estimation de l'effet moyen du traitement

16:15-16:30 **Vanessa McNealis** (McGill University) **Erica E.M. Moodie** (McGill University) **Nema Dean** (University of Glasgow)
Doubly Robust Estimation of Causal Effects in the Presence of Network Interference / Estimation doublement robuste d'effets causaux en présence d'interférence réseau

16:30-16:45 **Ian E. Waudby-Smith** (Carnegie Mellon University) **David Arbour** (Adobe Research) **Ritwik Sinha** (Adobe Research) **Edward H. Kennedy** (Carnegie Mellon University) **Aaditya Ramdas** (Carnegie Mellon University)
Time-Uniform Central Limit Theory with Applications to Anytime-Valid Causal Inference / Théorie centrale limite uniforme dans le temps avec applications à l'inférence causale valide à tout moment

16:45-17:00 **Zeyu Bian** (McGill University) **Erica E.M. Moodie** (McGill University) **Susan Shortreed** (University of Washington) **Sahir R. Bhatnagar** (McGill University)
Variable Selection for Dynamic Treatment Regimes / Sélection de variables pour des régimes de traitement dynamique

15:30-17:00 **Contributed / Communications libres** (abstract/résumé 114)

Business and Industrial Statistics
Statistique industrielle et de gestion

Chair/Président: Yanglei Song

15:30-15:45 **Alexander Shestopaloff** (Memorial University of Newfoundland) **Radford M. Neal** (University of Toronto)
Bayesian Inference for Partially Observed Queueing Systems with Markov Chain Monte Carlo / Inférence bayésienne pour des systèmes de file d'attente partiellement observés avec la méthode Monte-Carlo par chaînes de Markov (MCMC)

15:45-16:00 **Po Yang** (Unviersity of Manitoba) **Shanika Basnayake** (University of Manitoba)
Bayesian Optimal Designs with High Prediction Efficiency / Plans optimaux bayésiens avec efficacité prédictive élevée

16:00-16:15 **Sean Hellingman** (Wilfrid Laurier University) **Zilin Wang** (Wilfrid Laurier University) **Mary E. Thompson** (University of Waterloo)
Markov Chain Models for Professional Soccer Tracking Data / Les modèles de chaîne de Markov pour les données de suivi du soccer professionnel

16:15-16:30 **Luke Hagar** (University of Waterloo) **Nathaniel T. Stevens** (University of Waterloo)
A More Computationally Tractable Approach to Bayesian Interval-Based Sample Size Determination / Méthode de calcul simplifiée pour la détermination du nombre de sujets nécessaires en fonction d'un intervalle bayésien

16:30-16:45 **Xiaohua Liu** (University of Manitoba) **Po Yang** (University of Manitoba)
A Bayesian Approach to Process Optimization On Data with Multi-Stratum Structure / Une approche bayésienne pour l'optimisation de processus pour des données avec structure à multistrates

16:45-17:00 **Hugh Chipman** (Acadia University) **Derek Bingham** (Simon Fraser University)
Let's practice what we preach: Planning and interpreting simulation studies with design and analysis of experiments / Mettre en pratique ce que l'on prêche : planification et interprétation d'études de simulation avec conception et analyse d'expériences

Tuesday May 31 mardi 31 mai

11:00-12:30 **Invited / Sur invitation** (abstract/résumé 117)

Recent Advances in Statistical Analysis of Event History Data
Progrès récents en analyse statistique des données d'historique des événements

Chair/Président: Leilei Zeng
Organizer/Responsable: Leilei Zeng

11:00-11:30 **Hua Shen** (University of Calgary)
Recurrent Event Analysis with Misclassified Covariate / Analyse d'événements récurrents avec covariable mal classée

11:30-12:00 **Yan Yuan** (University of Alberta) **Zhe Lu** (University of Alberta)
Age-Specific Risk Prediction, a Case Study of Early Menopause in Childhood Cancer Survivors / Prédiction du risque en fonction de l'âge : une étude de cas sur la ménopause précoce chez les survivants d'un cancer infantile

12:00-12:30 **Liqun Diao** (University of Waterloo) **Richard J. Cook** (University of Waterloo) **Ce Yang** (Harvard University)
Survival Trees for Current Status Data / Arbres de survie pour des données d'état actuel

---

11:00-12:30 **Invited / Sur invitation** (abstract/résumé 119)

Ensemble Learning via Diverse and Random Projections of Features
Apprentissage d'ensemble via des projections diverses et aléatoires de caractéristiques

Chair/Président: William J. Welch
Organizer/Responsable: Jabed Tomal

11:00-11:30 **S. Ejaz Ahmed** (Brock University)
Post Shrinkage Strategy in High Dimensional Data Analysis / Stratégie post-rétrécissement dans l'analyse de données de grande dimension

11:30-12:00 **Timothy I. Cannings** (University of Edinburgh) **Richard J. Samworth** (University of Cambridge)
Random-projection ensemble classification / Classification d'ensembles de projections aléatoires

12:00-12:30 **Jabed Tomal** (Thompson Rivers University) **William J. Welch** (University of British Columbia) **Ruben H. Zamar** (University of British Columbia)
Robust Ranking by Ensembling of Diverse Models and Assessment Metrics / Classement robuste par ensemble de divers modèles et métriques d'évaluation

---

11:00-12:30 **Invited / Sur invitation** (abstract/résumé 121)

Modelling Extreme Risks in Insurance
Modélisation des risques extrêmes en assurance

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

11:00-11:30 **Mélina Mailhot** (Concordia University) **Fatima Palacios Rodriguez** (Universidad de Sevilla) **Elena Di Bernardino** (Université Côte D'Azur)
Smooth Copula-based Generalized Extreme Value model / Modèle lisse d'extrémum généralisé à l'aide de copules

11:30-12:00 **Menglin Zhou** (The University of British Columbia) **Natalia Nolde** (University of British Columbia)
Reverse Stress Testing and Multivariate Extremes / Tests de résistance inversés et extrêmes multivariés

12:00-12:30 **Mathieu Boudreault** (Université du Québec à Montréal)
A Global Flood Risk Modeling Framework Built with Climate Models and Machine Learning / Cadre mondial de modélisation des risques d'inondation construit avec des modèles climatiques et l'apprentissage automatique

---

11:00-12:30 **Invited / Sur invitation** (abstract/résumé 123)

Improving Robust High-dimensional Causal Inference and Prediction Modelling
Amélioration de l'inférence causale robuste à haute dimension et modélisation de la prédiction

Chair/Président: Celia M.T. Greenwood
Organizer/Responsable: Celia M.T. Greenwood, Gabriela Cohen Freue
Sponsor/Commanditaires: Canadian Statistical Sciences Institute (CANSSI) / Institut canadien des sciences statistiques (INCASS)

11:00-11:25 **Sahir R. Bhatnagar** (McGill University)
Variable Selection in Parametric Hazard Models / Sélection de variables dans les modèles de risque paramétriques

11:25-11:50 **Xinyi Zhang** (University of Toronto)
Fighting Noise with Noise: Causal Inference with Many Candidate Instruments / Combattre le bruit par le bruit : inférence causale à l'aide de nombreux instruments candidats

11:50-12:15 **Eric Tchetgen Tchetgen** (Uniiversity of Pennsylvania)
Doubly Robust Calibration of Prediction Sets under Covariate Shift / Calibration doublement robuste d'ensembles de prédiction selon un décalage de covariables

12:15-12:30 **Gabriela Cohen Freue** (The University of British Columbia)
Discussion / Discussion

11:00-12:30 **Invited / Sur invitation** (abstract/résumé 126)

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

Sponsor/Commanditaires: Survey Methods Section / Le Groupe des méthodes d'enquête

11:00-12:00 **Mark S. Handcock** (University of California, Los Angeles) **Ian E. Fellows** **Krista J. Gile** **Henry F. Raymond**
Sampling Hard-to-Reach Populations / Échantillonnage de populations difficiles à rejoindre

---

11:00-12:30 **Contributed / Communications libres** (abstract/résumé 127)

New Developments and Applications of Machine-learning Methods
Nouveaux développements et applications des méthodes d'apprentissage automatique

Chair/Président: Alessandro Maria Maria Selvitella

11:00-11:15 **Gabriel Oppong Afriyie** (University of Calgary) **Meng Wang** (University of Calgary, Canada) **Na Li** (University of Calgary, Canada) **Chel Hee Lee** (University of Calgary, Canada) **Alberto Nettel Aguirre** (University of Wollongong, Australia) **Anita Brobbey** (University of Calgary, Canada) **David Hughes** (University of Liverpool, United Kingdom) **Tolulope Sajobi** (University of Calgary, Canada)
Longitudinal Discriminant Aanalysis for Dementia Risk Prediction / Analyse discriminante longitudinale pour prédire le risque de démence

11:15-11:30 **Gansen Deng** (Western University) **Ryan Koh** (Toronto Rehabilitation Institute) **Wenqing He** (Western University) **Samah Hassan** (Toronto Rehabilitation Institute) **Shoba Subramaniam** (Toronto Rehabilitation Institute) **Dinesh Kumbhare** (Toronto Rehabilitation Institute)
Self-reported Data Analysis of a Chronic Pain Study / Analyses de données autodéclarées d'une étude sur la douleur chronique

11:30-11:45 **Antonio Peruzzi** (Ca' Foscari University of Venice) **Roberto Casarin** (Ca' Foscari University of Venice)
Media Bias and Polarization via a Markov-Switching Latent Space Model: an Application to the Media Environment of France, Germany, and Italy. / Les biais et la polarisation médiatiques à l'aide d'un modèle latent à changement d'espaces de Markov : application à l'environnement médiatique en France, Allemagne et Italie Ⓔ E

11:45-12:00 **Amin Kharaghani** (University of Toronto: Dalla Lana School of Public Health) **Milos Milic** (Krembil Centre for Neuroinformatics, Centre for Addiction and Mental Health) **Earvin Tio** (Krembil Centre for Neuroinformatics, Centre for Addiction and Mental Health) **David A. Bennett** (Rush University Medical Center) **Philip L. De Jager** (Columbia University Medical Center) **Julie A. Schneider** (Rush University Medical Center) **Lei Sun** (University of Toronto) **Daniel Felsky** (Krembil Centre for Neuroinformatics, Centre for Addiction and Mental Health; University of Toronto)
Association of Novel Whole-person Eigen-polygenic Scores with Alzheimer's disease / Association à la maladie d'Alzheimer de nouveaux scores polygéniques eigen de soins intégraux de la personne Ⓔ E

12:00-12:15 **Mohammad Kaviul Anam Khan** (University of Toronto) **Rafal Kustra** (University of Toronto)
Understanding the Properties of Permutation Based Importance for Inputs of Black Box Machine Learning Methods / Comprendre les propriétés de l'importance basée sur des permutations pour les entrées des méthodes d'apprentissage automatique «boîte noire» Ⓔ E

12:15-12:30 **Mengying Lei** (McGill University) **Aurélie Labbe** (HEC Montreal) **Lijun Sun** (McGill University)
Scalable Spatiotemporally Varying Coefficient Modelling with Bayesian Kernelized Tensor Regression / Modélisation des coefficients extensibles à variation spatio-temporelle avec régression tensorielle noyautée bayésienne Ⓔ E

11:00-12:30 **Contributed / Communications libres** (abstract/résumé 131)

Bayesian Inference and Modelling
Inférence bayésienne et modélisation

Chair/Président: Grace S. Chiu

11:00-11:15 **Pankaj Uttam Bhagwat** (University of Sherbrooke) **Eric Marchand** (University of Sherbrooke)
Bayesian Inference and Prediction for Mean-Mixtures of Normal Distributions / Inférence bayésienne et prédiction pour des mélanges de lois normales sur la moyenne Ⓔ E

11:15-11:30 **Ziming Chen** (University of Toronto) **Jeffrey Berger** (New York University School of Medicine) **Lana Castellucci** (The Ottawa Hospital) **Michael Farkouh** (Toronto General Hospital, University Health Network, Toronto) **Ewan Goligher** (Toronto General Hospital, University Health Network, Toronto) **Beverley Hunt** (King's College, London) **Lucy Kornblith** (University of California San Francisco) **Patrick Lawler** (Peter Munk Cardiac Centre, University Health Network, Toronto) **Eric Leifer** (National Heart, Lung, and Blood Institute) **Matthew Neal** (University of Pittsburgh Medical Center) **Ryan Zarychanski** (University of Manitoba) **Anna Heath** (The Hospital for Sick Children, Toronto, University of Toronto, University College London)
A Comparison of Methods for Bayesian Inference in Clinical Trials / Comparaison de méthodes d'inférence bayésienne pour les essais cliniques Ⓔ E

11:30-11:45 **James Willard** (McGill University)
Interim Analysis Covariate Adjustment for Bayesian Group Sequential Designs / Analyse intermédiaire avec covariables d'ajustement pour des plans bayésiens séquentiels de groupe Ⓔ E

11:45-12:00 **Shamsia Sobhan** (University of Manitoba) **Mahmoud Torabi** (University of Manitoba)
Spatial Survival Analysis in Presence of Semi-Competing Risks / Analyse de survie spatiale en présence de risques semi-concurrents Ⓔ E

12:00-12:15 **Victoire Michal** (McGill University) **Laís Picinini Freitas** (Fundação Oswaldo Cruz) **Alexandra M. Schmidt** (McGill University)
A Bayesian Hierarchical Model for Disease Mapping that Accounts for Scaling and Heavy-tailed Latent Effects / Un modèle hiérarchique bayésien en cartographie des maladies tenant compte de l'échelle et d'effets latents à queue épaisse Ⓔ E

12:15-12:30 **Nikola Surjanovic** (University of British Columbia) **Saifuddin Syed** (University of British Columbia) **Alexandre Bouchard-Côté** (University of British Columbia) **Trevor Campbell** (University of British Columbia)
Parallel Tempering With a Variational Reference / Atténuation parallèle avec référence variationnelle

---

11:00-12:30 **Contributed / Communications libres** (abstract/résumé 135)

Statistical Analysis of Covid-19 Data
Analyse statistique des données Covid-19

Chair/Président: Lengyi Spectrum Han

11:00-11:15 **Haoyu Wu** (McGill University)
Estimating COVID-19 Incidence using Deaths, Cases, Tests, Surveys, and Vaccinations / Estimation de l'incidence de la COVID-19 en répertoriant les décès, cas, tests, sondages et taux de vaccination

11:15-11:30 **Justin James Ian Slater** (University of Toronto) **Ayuish Bansal** (Centre for Global Health Research) **Jeffrey S. Rosenthal** (University of Toronto) **Harlan Campbell** (University of British Columbia) **Paul Gustafson** (University of British Columbia) **Patrick E. Brown** (University of Toronto)
An Almost Bayesian Approach to Estimating COVID-19 Incidence and Infection Fatality Rates / Approche quasi-bayésienne pour l'estimation de l'incidence de la COVID-19 et des taux de mortalité par infection

11:30-11:45 **Yasin Khadem Charvadeh** (University of Western Ontario) **Grace Y. Yi** (University of Western Ontario)
Comparing the Effectiveness of Virus Control Policies for COVID-19 with the Q-Learning Method / Comparer l'efficacité des politiques antivirus pour la COVID-19 avec la méthode d'apprentissage par renforcement (Q-learning)

11:45-12:00 **Yijia Weng** (Western University)
Meta-Analysis for Estimating the COVID-19 Average Incubation Time / Méta-analyse pour l'estimation du temps moyen d'incubation de la COVID-19

12:00-12:15 **Yuan Bian** (University of Western Ontario) **Yasin Khadem Charvadeh** (University of Western Ontario) **Grace Y. Yi** (University of Western Ontario) **Wenqing He** (University of Western Ontario)
Is 14-Days a Sensible Quarantine Length for COVID-19? A Case Study of COVID-19 Incubation Times / Une période de quarantaine de 14 jours pour la COVID-19 est-elle raisonnable ? Étude de cas sur les périodes d'incubation de la COVID-19

12:15-12:30 **Jingxue Feng** (Simon Fraser University) **Jie Wang** (Simon Fraser University) **Jiarui Zhang** (Simon Fraser University) **Liangliang Wang** (Simon Fraser University)
Clustering and Identification of SARS-CoV-2 Mutations Associated with Clinical Severity / Regroupement et identification des mutations du SRAS-CoV-2 associées à la sévérité clinique

---

12:30-13:30 **Poster / Poster** (abstract/résumé **??**)

Case Study 1: Developing a physician performance model in critical care – Assessing quality and value
Étude de cas 1 : Développer un modèle de performance des médecins en soins intensifs – Évaluation de la qualité et de la valeur

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

12:30-13:00 **Xiao Yang** (Carleton University) **Panxi Chen** (University of Michigan)
Carleton University/University of Michigan / Carleton University/University of Michigan

12:30-13:00 **Pablo Andres Lopera** (Conestoga College) **Martha Tellez** (Conestoga College) **Eduardo Gutierrez** (Conestoga College) **Rohit Thakur** (Conestoga College)
Conestoga College / Connestoga College

12:30-13:00 **Eralda Gjika** (Carleton University) **Belal Hossain** (University of British Columbia) **Amarildo Ceka** (University of British Columbia)
Carleton University/University of British Columbia / Carleton University/University of British Columbia E E

12:30-13:00 **Vadim Tyuryaev** (York University) **Aleksandr Tsybakin** (York University)
York University 1 / York University 1 E E

12:30-13:00 **Muditha Lakmali** (University of Manitoba) **Md Ashiqul Haque** (University of Manitoba) **Azizur Rahman** (University of Manitoba) **Samuel Quan** (University of Manitoba)
University of Manitoba / University of Manitoba E E

12:30-13:00 **Jingyu Cui** (Western University) **Gansen Deng** (Western University) **Chenqian Xian** (Western University) **Yijia Weng** (Western University)
Western University 1 / Western University 1 E E

12:30-13:00 **Yuan Bian** (University of Western Ontario) **Hui Guo** (Western University) **Yu Shi** (Western University)
Western University 2 / Western University 2 E E

13:00-13:30 **Wonje Choi** (University of Calgary) **Sungki Park** (University of Calgary)
University of Calgary / University of Calgary E E

13:00-13:30 **Jing Guo** (York University) **Octavia Wong** (York University) **Yongwen Pan** (York University) **Vi Nguyen** (York University)
York University 2 / York University 2 E E

13:00-13:30 **Yue Gu** (University of Waterloo) **Rebecca Tang** (University of Waterloo) **Christina Feng** (University of Waterloo) **Daniel Mao** (University of Waterloo)
University of Waterloo 1 / University of Waterloo 1 E E

13:00-13:30 **Shiheng Huang** (McMaster University) **Hanwen Ju** (McMaster University) **Yiren Tan** (McMaster University) **Hainan Xu** (McMaster University)
McMaster University / McMaster University E E

13:00-13:30 **Alexandra Mossman** (University of Waterloo) **Jessica Saini** (University of Waterloo)
University of Waterloo 2 / University of Waterloo 2 E E

13:00-13:30 **Shreena Nisha Kalaria** (BC Cancer Research Centre) **Jianping Yu** (University of Victoria)
University of Victoria / University of Victoria E E

13:00-13:30 **Sébastien Jessup** (Concordia University) **Emily Wright** (Concordia University)
Concordia University / Université Concordia E E F

13:30-15:00 **Invited / Sur invitation** (abstract/résumé 139)

SSC 2021 Gold Medal Address
Allocution du récipiendaire de la Médaille d'or 2021 de la SSC

Chair/Président: Bruce Smith

13:30-14:30 **Art Owen** (Stanford University) **Hal Varian** (Google) **Dan Kluger** (Stanford University) **Harrison Li** (Stanford University) **Tim Morrison** (Stanford University)
Tie-breaker Designs / Plans d'échantillonnage de bris d'égalité E E

15:30-17:00 **Invited / Sur invitation** (abstract/résumé 140)

New Advances in Microbiome Data Science
Nouvelles avancées en science des données sur le microbiome

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

15:30-16:00 **Longhai Li** (University of Saskatchewan) **Wei Bai** (University of Saskatchewan) **Mei Dong** (University of Toronto) **Longhai Li** (University of Saskatchewan) **Wei Xu** (University of Toronto)
Randomized Quantile Residuals for Diagnosing Zero-Inflated Generalized Linear Mixed Models with Applications to Microbiome Count Data / Résidus quantiles randomisés pour le diagnostic des modèles linéaires généralisés mixtes avec excès de zéros et applications aux données de comptage du microbiome

16:00-16:30 **Wei Xu** (Princess Margaret Cancer Centre)
Model Development on Longitudinal Microbiome Sequencing Data using Machine-learning Methodology / Développement de modèle pour les données longitudinales de séquençage de microbiome au moyen d'une méthodologie d'apprentissage automatique

16:30-17:00 **Pingzhao Hu** (University of Manitoba)
Computational meta-analyses of oral microbiome studies / Méta-analyses computationnelles des études sur le microbiome buccal

---

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

Functional Data Analysis for Complex Data
Analyse fonctionnelle de données complexes

Chair/Président: Liangliang Wang
Organizer/Responsable: Liangliang Wang

15:30-16:00 **Jiguo Cao** (Simon Fraser University) **Shu Jiang** (University of Waterloo) **Graham Colditz Bernard Rosner**
Predicting the Onset of Breast Cancer using Mammogram Imaging Data with Irregular Boundary / Prévision de l'apparition du cancer du sein à l'aide de données d'imagerie mammaire à limites irrégulières

16:00-16:30 **Tianyu Guan** (Brock University)
Exploring Pre-launch Movie Electronic Word of Mouth Time Series by Functional Data Analysis / Exploration des critiques de films avant sortie par analyse de données fonctionnelles sur séries chronologiques

16:30-17:00 **Jinhan Xie** (University of Alberta)
Optimal Functional Logistic Regression Under Case-Control Design / Régression logistique fonctionnelle optimale planifiée sous un cas-témoins

---

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

Reflection and Outlook of Statistical Sciences – Celebration of the 50th Anniversary of the Canadian Statistical Community
Réflexion et perspectives des sciences statistiques – Célébration du 50e anniversaire de la communauté statistique canadienne

Chair/Président: Grace Y. Yi
Organizer/Responsable: Grace Y. Yi
Sponsor/Commanditaires: 50th Anniversary Committee / Le Comité du 50e anniversaire de la SSC

15:30-17:00 **Charmaine B. Dean** (University of Waterloo) **Richard Lockhart** (Simon Fraser University) **Johanna G. Nešlehová** (McGill University) **Bruno Rémillard** (HEC Montréal) **Thérèse A. Stukel** (ICES/ University of Toronto) **Lei Sun** (University of Toronto)
Reflection and Outlook of Statistical Sciences – Celebration of the 50th Anniversary of the Canadian Statistical Community / Réflexion et perspectives des sciences statistiques - Célébration du 50e anniversaire de la communauté statistique canadienne

---

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

Data Fairness and Ethics
Équité des données et éthique

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

15:30-16:00 **Fanny Chevalier** (University of Toronto)
Don't Look. See! Are we Blinded by Data (Visualization)? / Sommes-nous aveuglés par les (visualisations de) données ? E E

16:00-16:30 **Lauren Klein** (Emory University)
Data Feminism / Féminisme et données E E

16:30-17:00 **Nathan A. Taback** (University of Toronto)
Discussion / Discussion E E

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

Current Challenges in Genomic Epidemiology
Défis actuels en épidémiologie génomique

Chair/Président: Jinko Graham
Organizer/Responsable: Jinko Graham
Sponsor/Commanditaires: Biostatistics Section / Le Groupe de biostatistique

15:30-16:00 **Brad McNeney** (Simon Fraser University) **Pulindu Ratnasekera** (Simon Fraser University) **Jinko Graham** (Simon Fraser University)
Robust Inference of Gene-Environment Interaction from Heterogeneous Samples of Case-Parent Trios / Inférence robuste de l'interaction gène-environnement à partir d'échantillons hétérogènes E E

16:00-16:30 **Qingrun Zhang** (University of Calgary)
cLD: Rare-Variant Disequilibrium Between Genomic Regions Identifies Novel Genomic Interactions / Déséquilibre de liaison cumulatif (cLD) : un déséquilibre lié à un variant rare entre des régions génomiques pour identifier de nouvelles interactions génomiques E E

16:30-17:00 **Lloyd T. Elliott** (Simon Fraser University)
Brain Imaging Genetics with 40,000 Subjects and 3,000 Phenotypes / Génétique de l'imagerie cérébrale avec 40 000 sujets et 3 000 phénotypes E E

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

Graduate Research in Actuarial Science 2
Recherche aux cycles supérieurs en science actuarielle 2

Chair/Président: Fangda Liu
Sponsor/Commanditaires: Actuarial Science Section / Le Groupe de science actuarielle

15:30-15:45 **Ramin Eghbalzadeh** (Concordia University)
A discrete-time version of the arbitrage-Free Nelson-Siegel term structure model / Version en temps discret du modèle de structure des termes Nelson-Siegel sans arbitrage E E

15:45-16:00 **Mathilde Bourget** (UQAM)
Statistical Modeling of Flood Risk in Climate Change / Modélisation statistique du risque d'inondation en contexte de changements climatiques E F E

16:00-16:15 **Emma Kroell** (University of Toronto) **Silvana Manuela Pesenti** (University of Toronto) **Sebastian Jaimungal** (University of Toronto)
Reverse Sensitivity Testing for Stochastic Processes / Test de sensibilité inverse des processus stochastiques E E

16:15-16:30 **Spark Tseung** (University of Toronto) **Tsz Chai Fung** (Georgia State University) **Ian Weng Chan** (University of Toronto) **Andrei L. Badescu** (University of Toronto) **X. Sheldon Lin** (University of Toronto)
Modelling Heterogeneous Risks with Random Effects in the Mixture-of-Experts Model / Modélisation des risques hétérogènes à l'aide d'effets aléatoires dans le modèle de mélange d'experts Ⓔ E

16:30-16:45 **Liyuan Lin** (University of Waterloo) **Hirbod Assa** (Kent Business School) **Ruodu Wang** (University of Waterloo)
On technical properties and calibrations of PELVE / Propriétés techniques et calages du PELVE Ⓔ E

16:45-17:00 **Zhenzhen Huang** (University of Waterloo) **Pengyu Wei** (Nanyang Technological University) **Chengguo Weng** (University of Waterloo)
Statistical Classification Methods for the Combining Portfolio Strategy / Méthodes de classification statistique pour la stratégie de combinaison de portfolios Ⓔ E

15:30-17:00 **Contributed / Communications libres** (abstract/résumé 153)

Control Chart and Statistical Methods for Clinical Trials
Carte de contrôle et méthodes statistiques pour les essais cliniques

Chair/Président: Luke Hagar

15:30-15:45 **Armando Turchetta** (McGill University) **Nicolas Savy** (Institut de Mathématiques de Toulouse) **Erica E.M. Moodie** (McGill University) **David A. Stephens** (McGill University)
A Time-Dependent Poisson-Gamma Model for Recruitment Forecasting in Multicenter Clinical Trials / Modèle Poisson-Gamma dépendant du temps pour la prévision du recrutement dans les essais cliniques multicentriques Ⓔ E

15:45-16:00 **Fatemeh Mahmoudi** (University of Calgary) **Xuewen Lu** (University of Calgary)
Variable Selection in Semiparametric Shared Frailty Illness-death Models for Semi-competing Risks Data / Sélection des variables dans des modèles maladie-décès semi-paramétriques à fragilité partagée pour des données de risques semi-concurrents Ⓔ E

16:00-16:15 **Qi Lyu** (University of Regina)
Modified Economic Model of Hotelling's T^2 Control Chart with Variable Sampling Interval / Modèle économique modifié de la carte de contrôle T^2 d'Hotelling avec intervalle d'échantillonnage variable Ⓔ E

16:15-16:30 **Apsara Pathum Jayasooriya** (Memorial University of Newfoundland) **Asokan Mulayath Variyath** (Memorial University of Newfoundland) **Yanqing Yi** (Memorial University of Newfoundland)
Statistical Inference for Multiple Stage Randomized Clinical Trials with Binary Responses / Inférence statistique pour des essais cliniques randomisés en plusieurs étapes avec des réponses binaires Ⓔ E

16:30-16:45 **Junwei Shen** (McGill University) **Shirin Golchi** (McGill University) **Erica E.M. Moodie** (McGill University) **David Benrimoh** (McGill University, Aifred Health)
New designs for Bayesian adaptive cluster-randomized trials / Nouveaux plans pour les essais adaptatifs bayésiens randomisés en grappes Ⓔ E

16:45-17:00 **Hira Nadeem** (University of Regina)
Special Case of Direct-Inverse Sampling Scheme for the Cross Product Ratio - Participant Enrollment in Clinical Trials / Cas spécial du plan d'échantillonnage direct inversé pour le rapport de produits croisés – recrutement de participants à des essais cliniques Ⓔ E

15:30-17:00 **Contributed / Communications libres** (abstract/résumé 157)

Statistical Analysis of Dependent Data and Enviromental Data
Analyse statistique des données dépendantes et des données environnementales

Chair/Président: Hon-Yiu So

15:30-15:45 **Alex Stringer** (University of Waterloo)
New Results in Modelling Dependent Data / Nouveaux résultats dans la modélisation des données dépendantes

15:45-16:00 **Glen McGee** (University of Waterloo) **Alex Stringer** (University of Waterloo)
Flexible Marginal Models for Dependent Data / Modèles marginaux flexibles pour données dépendantes

16:00-16:15 **Mohamad Elmasri** (McGill University) **Aurélie Labbe** (HEC Montreal) **Denis Larocque** (HEC Montreal) **Laurent Charlin** (HEC Montreal)
Predictive Inference for Travel Time on Transportation Networks / Inférence prédictive pour le temps de trajet sur les réseaux de transport

16:15-16:30 **Kexin Luo** (Western University) **Myriam Brossard** (Lunenfeld-Tanenbaum Research Institute, Sinai Health) **Shelley B. Bull** (Lunenfeld-Tanenbaum Research Institute, Sinai Health; Dalla Lana School of Public Health, University of Toronto)
Estimation of Genome-wide Significance Thresholds for Multi-variant Region-level Genetic Association Testing of Complex Traits / Estimation des niveaux de signification pangénomiques pour tester l'association de régions génétiques à multi-variant pour des caractères complexes

16:30-16:45 **Kevin Granville** (University of Western Ontario) **Douglas G. Woolford** (University of Western Ontario) **Charmaine B. Dean** (University of Waterloo) **Colin B. McFayden** (Ontario Ministry of Northern Development, Mines, Natural Resources and Forestry, Aviation, Forest Fire and Emergency Services) **Den Boychuk** (Ontario Ministry of Northern Development, Mines, Natural Resources and Forestry, Aviation, Forest Fire and Emergency Services)
On the Selection of an Interpolation Method with an Application to the Fire Weather Index in Ontario, Canada / Sélection d'une méthode d'interpolation avec application à l'Indice forêt-météo en Ontario, au Canada

16:45-17:00 **François A. Marshall** (Boston Univeristy)
Inferring Driver Nonlinearity in Physical Systems using Cyclostationary Signal Processing / Déduire la nonlinéarité du moteur dans les systèmes physiques à l'aide du traitement du signal cyclostationnaire

Wednesday June 1 — mercredi 1 juin

11:00-12:15 **Invited / Sur invitation** (abstract/résumé 161)

2021 SSC Impact Award Address
Allocution du récipiendaire du prix pour impact de la SSC 2021

Chair/Président: Tolu Sajobi

11:00-12:00 **Thérèse A. Stukel** (ICES/ University of Toronto)
Innovative Uses of Health Administrative Data for Health Policy Research / Utilisation innovatrice de données administratives de santé pour la recherche sur les politiques en santé Ⓔ EF

11:00-12:30 **Invited / Sur invitation** (abstract/résumé 162)

Input Privacy Preserving Technologies for Official Statistics
Technologies de préservation de la confidentialité des données pour les statistiques officielles

Chair/Président: Abel C. Dasylva
Organizer/Responsable: Abel C. Dasylva
Sponsor/Commanditaires: Survey Methods Section / Le Groupe des méthodes d'enquête

11:00-11:30 **Teresa Scassa** (University of Ottawa)
Legal Dimensions of Privacy Preserving Technologies for Official Statistics / Dimensions légales de technologies protégeant la confidentialité pour les statistiques officielles Ⓔ E

11:30-12:00 **Saeid Molladavoudi** (Statistics Canada)
Privacy Enhancing Technologies at Statistics Canada / Technologies d'amélioration de la confidentialité à Statistique Canada Ⓔ E

12:00-12:30 **Jerome Reiter** (Duke University) **Chengxin Yang** (Duke University)
Formally Private Verification of Statistical Analyses / Vérification formellement privée d'analyses statistiques Ⓔ E

11:00-12:30 **Invited / Sur invitation** (abstract/résumé 164)

A Memorial Session for Hélène Massam
Séance commémorative pour Hélène Massam

Chair/Président: Christian Genest
Organizer/Responsable: Xin Gao, Christian Genest
Sponsor/Commanditaires: Probability Section / Le Groupe de probabilité

11:00-11:06 **Christian Genest** (McGill University)
Introductory Remarks / Remarques introductives Ⓔ E

11:06-11:34 **Laurent Briollais** (University of Toronto/Lunenfeld-Tanenbaum Research Institute) **Nanwei Wang** (University of New Brunswick) **Xin Gao** (York University) **Helene Massam** (York University)
The Scalable Birth-death MCMC Algorithm for Mixed Graphical Model Learning with Application to Genomic Data Integration / L'Algorithme MCMC de naissance et de mort pour l'apprentissage de modèles graphiques mixtes appliqué à l'intégration de données génomiques Ⓔ E

11:34-12:02 **Yanyan Wu** (University of Hawaii at Manoa)
Saddle Point Approximation for the Test of Equality of Covariance Matrices from Decomposable Graphical Gaussian Models / Méthode du point col pour le test de l'égalité des matrices de covariance découlant de modèles graphiques gaussiens décomposables Ⓔ E

12:02-12:30 **Gerard Letac** (Université de Toulouse)
Scale Mixtures of Gaussian Laws: the Quasi-Kolmogorov-Smirnov and Logistic Laws. / Mélanges de variance de lois gaussiennes : lois quasi Kolmogorov-Smirnov et quasi logistiques Ⓔ E

11:00-12:30 **Invited / Sur invitation** (abstract/résumé 166)

The Business of Sports Analytics
Le commerce de l'analyse sportive

Chair/Président: Jean-François Plante
Organizer/Responsable: Shirley E. Mills
Sponsor/Commanditaires: Business and Industrial Statistics Section / Le Groupe de statistique industrielle et de gestion

11:00-11:30 **Michael Jung** (Maple Leaf Sports and Entertainment)
Creating Actionable Insights in the Business of Sports / Création de perspectives exploitables dans l'industrie du sport Ⓔ E

11:30-12:00 **Luke C. Bornn** (Zelus Analytics) **Jacob Mortensen** (Zelus Analytics)
From Pixels to Points: Using Tracking Data to Measure Performance in Professional Sports / Des pixels aux points : l'utilisation de données de suivi pour mesurer la performance dans les sports professionnels Ⓔ E

12:00-12:30 **Shane Malloy** (University of New Brunswick)
The Future of Statistics in NHL Hockey Operations / L'avenir de la statistique dans les opérations de hockey de la LNH Ⓔ E

11:00-12:30 **Contributed / Communications libres** (abstract/résumé 168)

Insurance, Reinsurance, and Finance
Assurance, réassurance et finance

Chair/Président: Golara Zafari

11:00-11:15 **Shi Zhang** (University of New Brunswick) **Renjun Ma** (University of New Brunswick) **Guohua Yan** (University of New Brunswick)
Cox Survival Models with Partially Crossed Random Effects: an Application to Car Accident Data Cross-Classified by Location and Agent / Modèles de survie de Cox avec effets aléatoires partiellement croisés : application aux données d'accidents de voiture classées de manière croisée par lieu et par agent Ⓔ E

11:15-11:30 **Ye Wang** (University of Calgary) **Wenjun Jiang** (University of Calgary)
Optimal Reinsurance Under Vajda Condition and Range-Value-at-Risk / Réassurance optimale sous condition de Vajda et plage de valeur à risque Ⓔ E

11:30-11:45 **Louis Arsenault-Mahjoubi** (Simon Fraser University) **Jean-François Bégin** (Simon Fraser University)
On the Bayesian Estimation of Jump-Diffusion Models in Finance / Sur l'estimation bayésienne des modèles de diffusion avec sauts en finance Ⓔ E

11:45-12:00 **Dechen Gao** (Western University) **Jiandong Ren** (Western University)
Fuzzy credibility / Crédibilité floue Ⓔ E

12:00-12:15 **Tingting Chen Peter Adamic** (Laurentian University) **Anthony F. Desmond** (University of Guelph)
Generalized Additive Modelling for the Accurate Estimation of Insurance Claims / Modélisation additive généralisée pour l'estimation précise des réclamations d'assurance Ⓔ E

12:15-12:30 **Si Chen** (Wilfrid Laurier University) **Zilin Wang** (Wilfrid Laurier University) **David Soave** (Wilfrid Laurier University) **Mary Kelly** (Wilfrid Laurier University)
Fitting Left Truncated Data using Aggregate Loss Model with Poisson-Tweedie Loss Frequency / Ajustement de données tronquées à gauche en utilisant le modèle de perte agrégée avec fréquence de perte Poisson-Tweedie Ⓔ E

11:00-12:30 **Contributed / Communications libres** (abstract/résumé 172)

Statistics Education, Efficient Computation, and Studies Related to Covid-19
Éducation en statistique, calcul efficace et études sur la Covid-19

Chair/Président: Yifan Li

11:00-11:15 **Jack Davis** (University of Waterloo)
Gambling and Games of Chance – A Course Proposal / Jeux d'argent et jeux de hasard – Une proposition de cours

11:15-11:30 **Suborna Shekhor Ahmed** (University of British Columbia) **Michelle Zeng** (University of British Columbia) **Patrick Culbert** (University of British Columbia) **Yangqian Qi** (University of British Columbia)
Survey data analysis of engagement and self-efficacy in a concurrent hybrid modality / Analyses de données d'enquête sur l'engagement et l'autoefficacité dans un modèle concurrent hybride

11:30-11:45 **Samuel Perreault** (University of Toronto)
Efficient Computation for Inference with Kendall's Tau / Calcul efficace pour l'inférence avec sur le tau de Kendall

11:45-12:00 **Federico Severino** (Université Laval) **Marzia Angela Cremona** (Université Laval) **Éric Dadié** (Université Laval)
COVID-19 effects on the Canadian Term Structure of Interest Rates / Effets de la COVID-19 sur la structure à terme des taux d'intérêt au Canada

12:00-12:15 **William Ruth** (Simon Fraser University) **Richard Lockhart** (Simon Fraser University)
Simulated Epidemic Spread in University Classes / Simulation d'une propagation épidémique pendant les cours dans une université

12:15-12:30 **Surani Matharaarachchi** (University of Manitoba) **Mike Domaratzki** (University of Western Ontario) **Alan Katz** (University of Manitoba) **Saman Muthukumarana** (University of Manitoba)
Discovering Symptom Patterns of Long COVID Patients in Tweets using Association Rule Mining / Découverte de modèles de symptômes chez des patients atteints de COVID longue dans des tweets à l'aide de l'extraction de règles d'association

12:30-13:30 **Poster / Poster** (abstract/résumé **??**)

Case Study 2: Towards a clear understanding of rural internet – What statistical measure can be used to assess, compare and forecast internet speed for rural Canadian communities
Vers une compréhension claire de l'Internet rural – Quelles mesures statistiques peuvent être utilisées pour évaluer, comparer et prévoir les vitesses d'Internet pour les communautés rurales canadiennes ?

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

12:30-13:00 **Jonghoon Park** (York University) **Ravish Kamath** (York University) **Sonny Dinh** (York University)
York University 1 / York University 1

12:30-13:00 **Philipp Schroeppel** (University of Waterloo) **Christian Mitrache** (University of Waterloo)
University of Waterloo / University of Waterloo

12:30-13:00 **Wensha Zhang Son Luu** (Dalhousie University) **Jingyu Li** (Dalhousie University)
Dalhousie University / Dalhousie University

12:30-13:00 **Sylvester Ranjith Francis** (Conestoga College) **Parvathy Suresh** (Conestoga College) **Rachel Denzil** (Conestoga College)
Conestoga College / Conestoga College

13:00-13:30 **Joosung Min** (Simon Fraser University) **Daisy Yu** (Simon Fraser University) **Olga Vishnyakova** (Simon Fraser University) **Renny Doig** (Simon Fraser University)
Simon Fraser University / Simon Fraser University

13:00-13:30 **Thet Htet Chan Nyein** (University of Calgary) **Hamid Hamidi** (University of Calgary) **Yanzhao Qian** (University of Calgary) **Bahar Mousazadeh** (University of Calgary)
University of Calgary / University of Calgary

13:00-13:30 **Deniza Robinson** (York University) **Nga Nguyen** (York University) **S.M. Nehanul Hoque** (York University) **Andrew Fallone** (York University)
York University 2 / York University 2

13:00-13:30 **Jianan Wang** (McMaster University) **Chun Dong** (McMaster University) **Xiangyu Lyu** (McMaster University) **Xiaoyan Liu** (McMaster University)
McMaster University / McMaster University

13:30-15:00 **Invited / Sur invitation** (abstract/résumé 175)

Memorial Session for Donald A. S. Fraser
Session commémorative pour Donald A. S. Fraser

Chair/Président: Mary E. Thompson
Organizer/Responsable: Mary E. Thompson
Sponsor/Commanditaires: 50th Anniversary Committee / Le Comité du 50e anniversaire de la SSC

13:30-13:45 **Christian Genest** (McGill University)
Introductory Remarks / Remarques introductives

13:45-14:25 **Nancy Reid** (University of Toronto)
From Structural Inference to Asymptotic Theory / De l'inférence structurelle à la théorie asymptotique

14:25-15:00 **Mylène Bédard** (Université de Montréal)
Recent Advances in Statistical Inference / Avancées récentes en inférence statistique

---

13:30-15:00 **Invited / Sur invitation** (abstract/résumé 176)

Advances in Extreme Value Modelling
Avancées en modélisation des valeurs extrêmes

Chair/Président: Johanna G. Nešlehová
Organizer/Responsable: Léo Belzile

13:30-14:00 **Stanislav Volgushev** (University of Toronto) **Sebastian Engelke** (University of Geneva) **Michaël Lalancette** (University of Toronto)
Structure Learning for Extremal Graphical Models / Apprentissage de structure pour des modèles graphiques extrêmes

14:00-14:30 **Natalia Nolde** (The University of British Columbia)
Linking Representations for Multivariate Extremes via a Limit Set / Liaison de représentations d'extrêmes multivariés par l'entremise d'un ensemble limite

---

13:30-15:00 **Invited / Sur invitation** (abstract/résumé 178)

Statistical Learning and Inference on Streaming and Online Data
Apprentissage statistique et inférence sur les données "streaming" et en ligne

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

13:30-13:52 **Yingqi Zhao**
Constructing Stabilized Dynamic Surveillance Rules for Optimal Monitoring Schedules / Construction de règles de surveillance dynamique stabilisées pour programmes de surveillance optimaux

13:52-14:14 **Peter X. Song** (University of Michigan) **Emily Hector** (North Carolina State University) **Lan Luo** (University of Iowa)
Parallel-and-stream accelerator for computationally fast supervised learning with big data / Accélérateur parallèle et de diffusion pour l'apprentissage supervisé rapide sur le plan informatique avec des mégadonnées Ⓔ E

14:14-14:36 **Hengrui Cai** (North Carolina State University) **Ye Shen** (North Carolina State University) **Rui Song** (North Carolina State University)
Doubly Robust Interval Estimation for Optimal Policy Evaluation in Online Learning / Estimation doublement robuste d'intervalles pour une évaluation optimale des politiques en matière d'apprentissage en ligne Ⓔ E

14:36-14:58 **Linglong Kong** (University of Alberta)
Damped Anderson Mixing for Deep Reinforcement Learning: Acceleration, Convergence, and Stabilization / Mélange d'Anderson amorti pour l'apprentissage par renforcement profond : accélération, convergence et stabilisation Ⓔ E

13:30-15:00 **Invited / Sur invitation** (abstract/résumé 181)

Active Learning in Statistics: Where Are We Now?
Apprentissage actif en statistique : où en sommes-nous ?

Chair/Président: Chelsea Uggenti
Organizer/Responsable: Douglas G. Woolford, Chelsea Uggenti
Sponsor/Commanditaires: Statistical Education Section / Le Groupe d'éducation en statistique

13:30-15:00 **Alison L. Gibbs** (University of Toronto) **Wesley Burr** (Trent University) **Sohee Kang** (University of Toronto Scarborough)
Active Learning in Statistics: Where Are We Now? / Apprentissage actif en statistique : où en sommes-nous ? Ⓔ E

13:30-15:00 **Invited / Sur invitation** (abstract/résumé 182)

Collaborations and Consultations in an Academic World
Collaboration et consultations dans un monde académique

Chair/Président: Peijun Sang
Organizer/Responsable: Orla A Murphy
Sponsor/Commanditaires: Committee on New Investigators / Le Comité des nouveaux chercheurs

13:30-15:00 **Mireille Schnitzer** (Université de Montréal) **Thomas Loughin** (Simon Fraser University) **Dave Campbell** (Carleton University) **Gabriela Cohen Freue** (University of British Columbia)
Collaborations and Consultations in an Academic World / Collaborations et consultations dans le monde académique Ⓔ E

13:30-15:00 **Invited / Sur invitation** (abstract/résumé 183)

Recent Advancement and Application of Bayesian Causal Inference Methods
Progrès récents et application des méthodes d'inférence causale bayésienne

Chair/Président: Kuan Liu
Organizer/Responsable: Kuan Liu
Sponsor/Commanditaires: Biostatistics Section / Le Groupe de biostatistique

13:30-14:00 **Olli Saarela** (University of Toronto) **Thai-Son Tang** (University of Toronto) **Zhihui (Amy) Liu** (University Health Network)
Bayesian Non-Parametric Monotonic Regression for Radiotherapy Induced Normal Tissue Complications / Régression monotone non paramétrique bayésienne pour les complications aux tissus sains après radiothérapie Ⓔ E

14:00-14:30 **Arman Oganisian** (Brown University)
A Hierarchical Bayesian Bootstrap for Heterogenous Treatment Effect Estimation / Bootstrap bayésien hiérarchique pour l'estimation des effets de traitement hétérogènes Ⓔ E

14:30-15:00 **Paul Gustafson** (University of British Columbia) **Daniel Daly-Grafstein** (University of British Columbia) **Conor Morrison** (University of British Columbia)
Bayesian Approaches to Causal Inference: The Present Position and the Path Ahead / Situation actuelle et perspectives d'avenir des approches bayésiennes d'inférence causale Ⓔ E

13:30-15:00 **Contributed / Communications libres** (abstract/résumé 185)

New Statistical Models and Their Applications
Nouveaux modèles statistiques et leurs applications

Chair/Président: Devan G Becker

13:30-13:45 **Matthew R.P. Parker** (Simon Fraser University) **Jiguo Cao** (Simon Fraser University) **Laura L.E. Cowen** (University of Victoria) **Lloyd T. Elliott** (Simon Fraser University) **Junling Ma** (University of Victoria)
Estimating the Burden of COVID-19 in BC Using New Disease Analytic Multi-site Models / Estimer le fardeau de la COVID-19 en C.-B. à l'aide de nouveaux modèles multisites d'analyse de maladie Ⓔ E

13:45-14:00 **Pingbo Hu** (Western University)
Characterizing the COVID-19 Dynamics with a New Epidemic Model: Susceptible-Exposed-Symptomatic-Asymptomatic-Active-Removed / Caractériser les dynamiques de la COVID-19 à partir d'un nouveau modèle épidémique : susceptible, exposé, symptomatique, asymptomatique, actif et retiré Ⓔ E

14:00-14:15 **Leif Erik Lovblom** (University of Toronto) **Laurent Briollais** (University of Toronto) **Bruce A. Perkins** (University of Toronto) **George Tomlinson** (University of Toronto)
A Joint Model for a Longitudinal Outcome and a Multistate Process Under Intermittent Observation, with Applications for Diabetes Complications / Un modèle conjoint pour un résultat longitudinal et un processus multi-états sous observation intermittente, avec des applications pour les complications du diabète Ⓔ E

14:15-14:30 **Mai Ghannam** (University of Windsor) **Sévérien Nkurunziza** (University of Windsor)
Tensor Shrinkage Estimators in a Generalized Tensor Regression Model / Estimateurs à rétrécissement tensoriels dans un modèle de régression tensorielle généralisée Ⓔ E

14:30-14:45 **Katherine Burak** (University of Alberta) **Adam B. Kashlak** (University of Alberta)
Nonparametric confidence regions via the analytic wild bootstrap / Régions de confiance non paramétriques avec bootstrap sauvage analytique Ⓔ E

14:45-15:00 **Meixi Chen** (University of Waterloo) **Martin Lysy** (University of Waterloo) **Reza Ramezan** (University of Waterloo)
Decoding Multi-Neuronal Activities Through Latent Factor Models / Modèles de facteurs latents pour le décodage d'activités multineuronales Ⓔ E

13:30-15:00 **Contributed / Communications libres** (abstract/résumé 189)

Recent Developments in Survey methods and Capture-recapture Methods
Développements récents des méthodes d'enquête et des méthodes de capture-recapture

Chair/Président: Omidali Aghababaei Jazi

13:30-13:45 **Marie-Pier Lemieux** (Statistics Canada)
2021 Canadian Census: Using an Agile Non-Response Management Strategy to Obtain Quality Data during a Pandemic / Recensement Canadien de 2021 : Utilisation d'une stratégie agile pour la gestion de la non-réponse afin d'obtenir des résultats de qualité en temps de pandémie ⒺⒻ E F

13:45-14:00 **Audrey Béliveau** (University of Waterloo)
Design-Unbiased Trapezoid Area-Under-the-Curve Estimators for Estimating Salmon Escapement / Estimateurs de l'aire sous la courbe par la méthode des trapèzes qui soient sans biais par rapport au plan afin d'estimer l'échappée de saumons Ⓔ E

14:00-14:15 **Thomas Yoon** (Statistics Canada)
Modernization of the Canadian Census: An Administrative Data-Driven Approach to Defining Households / Modernisation du recensement canadien : Une approche axée sur les données administratives pour définir les ménages Ⓔ E

14:15-14:30 **Abel C. Dasylva** (Statistics Canada) **Arthur Goussanou** (Statistics Canada)
A new model for the automated identification of duplicate records / Nouveau modèle d'identification automatique des enregistrements en double Ⓔ E

14:30-14:45 **Yiran Wang** (University of Waterloo) **Martin Lysy** (University of Waterloo) **Audrey Béliveau** (University of Waterloo)
Genetic Mark-Recapture Methods for Estimating Seasonal River Run Size of Stock Populations / Méthodes de marquage et de recapture génétiques pour l'estimation de la taille saisonnière de l'effectif à la montaison en rivière de stocks Ⓔ E

14:45-15:00 **Inesh Prabuddha Munaweera Arachchilage** (University of Manitoba) **Saman Muthukumarana** (University of Manitoba) **Darren Gillis** (University of Manitoba) **Les N. Harris** (Fisheries and Oceans Canada)
Bayesian Multi-state Capture-recapture Modelling for Estimating Survival Probabilities of Arctic Char using Acoustic Telemetry Data / Modélisation bayésienne multi-états de capture-recapture pour l'estimation des probabilités de survie de l'omble chevalier à l'aide de données de télémétrie acoustique Ⓔ E

13:30-14:45 **Contributed / Communications libres** (abstract/résumé 193)

Missing Data, Causal Inference, and New Algorithms for Differential Equations
Données manquantes, inférence causale et nouveaux algorithmes pour les équations différentielles

Chair/Président: Nikola Surjanovic

13:30-13:45 **Abdoulaye Dioni** (Université Laval) **Alexandre Bureau** (Université Laval) **Lynne Moore** (Université Laval) **Aida Eslami** (Université Laval)
Development of a method for missing not at random / Développement d'une méthode pour les données manquantes non aléatoirement Ⓕ E F

13:45-14:00 **Renny Doig** (Simon Fraser University) **Liangliang Wang** (Simon Fraser University)
Probabilistic Numerical Solution of Differential Equations as a Remedy for Discretization-Induced Bias / Solution numérique probabiliste d'équations différentielles comme remède au biais induit par la discrétisation Ⓔ E

14:00-14:15 **Jonathan Ramkissoon** (University of Waterloo) **Martin Lysy** (University of Waterloo)
Smoothly Differentiable Particle Filters for Stochastic Differential Equations / Filtres à particules facilement différentiables pour des équations différentielles stochastiques (EDS) Ⓔ E

14:15-14:30 **Mohan Wu** (University of Waterloo) **Martin Lysy** (University of Waterloo)
Parameter Inference for Differential Equations using Bridge Proposal / Inférence de paramètres pour des équations différentielles à l'aide d'une proposition de pont Ⓔ E

14:30-14:45 **Pranav Subramani** (University of Waterloo) **Jonathan Ramkissoon** (University Of Waterloo) **Mohan Wu** (University Of Waterloo) **Martin Lysy** (University Of Waterloo)
A Method for Parameter Inference for Stochastic Differential Equations / Méthode d'inférence des paramètres pour équations différentielles stochastiques Ⓔ E

15:30-17:00 **Invited / Sur invitation** (abstract/résumé 196)

Recent Developments in Methodology and Applications of Mixture Models
Développements récents en méthodes et applications des modèles de mélange

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

15:30-16:00 **Nhat Ho** (University of Texas, Austin)
Bayesian Sieves and Excess Mass Behavior in Infinite Mixtures / Cribles bayésiens et comportement de masse excédante dans des mélanges infinis

16:00-16:30 **Tudor A. Manole** (Carnegie Mellon University) **Cody Mazza-Anthony** (Shopify) **Nhat Ho** (University of Texas, Austin) **Abbas Khalili** (McGill University)
Order Selection in Finite Mixture of Regression Models / Sélection de l'ordre dans un mélange fini de modèles de régression

16:30-17:00 **Alejandro Murua** (Université de Montréal)
A Bayesian Semi-parametric Mixture of Survival Regression Model for Survival Prediction / Un mélange bayésien semi-paramétrique de modèles de survie pour la prédiction de survie

15:30-17:00 **Invited / Sur invitation** (abstract/résumé 198)

Recent Advances on Model Assessment in Recurrent Event Analysis
Progrès récents en évaluation des modèles pour l'analyse des événements récurrents

Chair/Président: Hua Shen
Organizer/Responsable: Hua Shen

15:30-16:00 **Eleanor M. Pullenayegum** (Hospital for Sick Children)
The Role of Recurrent Event Models in Handling Longitudinal Data Subject to Irregular Observation: Determining the Assessment Mechanism and Undertaking Sensitivity Analysis. / Rôle des modèles d'événements récurrents dans le traitement des données longitudinales observées de façon irrégulière : détermination du mécanisme d'évaluation et analyse de sensibilité

16:00-16:30 **Candemir Cigsar** (Memorial University of Newfoundland)
Model Assessment for Dynamic Recurrent Event Processes with Dependent Gap Times / Évaluation de modèles pour des processus dynamiques d'événements récurrents avec des laps de temps dépendants

16:30-17:00 **Hua Shen** (University of Calgary)
Discussion / Discussion

15:30-17:00 **Invited / Sur invitation** (abstract/résumé 200)

Environmental Data Science: Growth and Opportunities
Science des données environnementales : Croissance et opportunités

Chair/Président: Wesley S. Burr
Organizer/Responsable: Wesley S. Burr
Sponsor/Commanditaires: Data Science and Analytics Section / Le Groupe de science des données et analytique

15:30-16:00 **Allison Horst** (University of California, Santa Barbara) **Samantha Csik** (National Center for Ecological Analysis & Synthesis) **Jamie Montgomery** (University of California, Santa Barbara)
Filling a Training Gap in Environmental Workplaces: The Emergence of Environmental Data Science Degree Programs, and Lessons Learned from Running One / Combler l'écart en formation dans les milieux de travail en environnement : l'émergence de programmes d'études en science des données environnementales et leçons tirées de la direction d'un tel programme

16:00-16:30 **Holly N. Steeves** (University of Western Ontario) **Sofiia Romanovska** (University of Victoria) **Laura L.E. Cowen** (University of Victoria)
Exploring the Robustness of Citizen Science Golden Eagle Data / Exploration de la robustesse des données sur l'aigle royal issues de la science citoyenne

16:30-17:00 **Susan Simmons** (North Carolina State University)
Best Practices for Virtual Research Groups / Les meilleures pratiques pour les groupes de recherche virtuels

---

15:30-17:00 **Invited / Sur invitation** (abstract/résumé 202)

Modeling Actuarial Risks
Modélisation des risques actuariels

Chair/Président: Mélina Mailhot
Organizer/Responsable: Mélina Mailhot
Sponsor/Commanditaires: Actuarial Science Section / Le Groupe de science actuarielle

15:30-16:00 **Silvana Manuela Pesenti** (University of Toronto) **Mélina Mailhot** (Concordia University) **Emily Wright** (Concordia University)
Renyi Divergence for Extreme Value Distributions / Divergence de Rényi pour les distributions des valeurs extrêmes

16:00-16:30 **Tobias Fissler** (Vienna University of Economics and Business) **Michael Merz** (University of Hamburg) **Mario V. Wüthrich** (ETH Zurich)
Deep Quantile and Deep Composite Model Regression / Régression quantile profonde et modèle de régression composite profond

16:30-17:00 **Klaus Herrmann** (Université de Sherbrooke) **Marius Hofert** (University of Waterloo) **Johanna G. Nešlehová** (McGill University)
Copula Diagonals, Distortions and the Asymptotic Distribution of Maxima / Diagonales des copules, distorsions et distribution asymptotique des maxima

---

15:30-17:00 **Contributed / Communications libres** (abstract/résumé 204)

New Developments for Analyzing Insurance and Finance Data
Nouveaux développements pour l'analyse des données d'assurance et de finance

Chair/Président: Yanbo Tang

15:30-15:45 **Di Meng** (Wilfrid Laurier University) **Mark Reesor** (Wilfrid Laurier University) **Adam Metzler** (Wilfrid Laurier University)
Calibration and Pricing of Contingent Convertible Securities / Calibrage et fixation du prix des titres convertibles conditionnels

15:45-16:00 **Félix Locas** (Université du Québec à Montréal)
De Finetti's Control Problem with Absolutely Continuous Strategies in a Diffusion Model / Problème de contrôle stochastique de De Finetti avec stratégies absolument continues dans un modèle de diffusion

16:00-16:15 **Yifan Li** (University of Western Ontario) **Reg Kulperger** (University of Western Ontario) **Hao Yu** (University of Western Ontario)
Semi-G-Structure: A Flexible Framework to Deal with Model Uncertainty / Semi-structure G : un cadre souple pour traiter l'incertitude du modèle

16:15-16:30 **Yunhong Lyu** (University of Windsor) **Sévérien Nkurunziza** (University of Windsor)
Estimation and Testing in Generalized Cox–Ingersoll–Ross Processes / Estimation et test dans les processus de Cox–Ingersoll–Ross Généralisés

16:30-16:45 **Elham Soufiani** (University of Regina)
Generalization of Hoeffding's inequality for Extended Acceptable Random Variables / Généralisation de l'inégalité d'Hoeffding pour les variables aléatoires acceptables étendues

16:45-17:00 **Sharandeep Singh Pandher** (University of Regina) **Shakhawat Hossain** (University of Winnipeg) **Andrei Volodin** (University of Regina)
Generalized Autoregressive Moving Average (GARMA) Models: An Efficient Estimation Approach / Modèles de moyenne mobile autorégressive généralisée (GARMA) : une approche d'estimation efficace

---

15:30-17:00 **Contributed / Communications libres** (abstract/résumé 207)

Statistical Methods for Handling Ordinal Data, Missing data, and Data with Measurement Error
Méthodes statistiques pour le traitement des données ordinales, des données manquantes et des données comportant des erreurs de mesure

Chair/Président: Joan X. Hu

15:30-15:45 **Lyubov Doroshenko** (Université Laval) **Brunero Liseo** (La Sapienza University of Rome)
Generalized Linear Mixed Model with Bayesian Rank Likelihood / Modèle linéaire généralisé mixte à l'aide de la probabilité de rang bayésienne

15:45-16:00 **Aya A. Mitani** (University of Toronto) **Osvaldo Espin-Garcia** (University Health Network, University of Toronto) **Victoria Landsman** (Institute of Work and Health, University of Toronto)
Using Stereotype Regression for Unbiased Inference from Ordinal Outcome-Dependent Samples / Emploi de régression de stéréotype pour une inférence non biaisée à partir d'échantillons dépendant des résultats ordinaux

16:00-16:15 **Gurbakhshash Singh** (Central Connecticut State University) **Gordon H. Fick** (University of Calgary)
Ordinal Outcomes: Considerations for the Generalized Linear Model with the Identity Link / Résultats ordinaux : Considérations au sujet du modèle linéaire généralisé avec lien d'identité

16:15-16:30 **Hon-Yiu So** (University of Waterloo) **Parminder Raina** (McMaster University) **Jinhui Ma** (McMaster University)
Application of Machine Learning in Imputing Heterogeneous Co-missing Data / Application de l'apprentissage automatique dans l'imputation de données co-manquantes hétérogènes

16:30-16:45 **Yifan Sun** (University of Western Ontario)
Estimation and Variable Selection for Function-on-scalar Linear Model with Covariate Measurement Error / Estimation et sélection de variables pour modèle linéaire à fonctions scalaires avec erreur de mesure de covariables

16:45-17:00 **Max Turgeon** (University of Manitoba)
Generalized Soft Impute for Matrix Completion / Imputation douce généralisée pour la complétion matricielle

---

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

Dynamic Treatment Regime Analysis and Dynamic Modelling
Analyse du régime de traitement dynamique et modélisation dynamique

Chair/Président: Andrea Benedetti

15:30-15:45 **Marzieh Mussavi Rizi** (University of Waterloo) **Joel A. Dubin** (University of Waterloo) **Michael Wallace** (University of Waterloo)
Dynamic Treatment Regimes in Dyadic Networks / Régimes de traitement dynamiques dans les réseaux dyadiques

15:45-16:00 **Cong Jiang** (University of Waterloo) **Michael Wallace** (University of Waterloo) **Mary E. Thompson** (University of Waterloo)
Doubly-Robust Dynamic Treatment Regimen Estimation for Binary Outcomes / Estimation dynamique du régime de traitement à double robustesse pour les résultats binaires

16:00-16:15 **Dan Liu** (Western University) **Wenqing He** (Western University)
Q-learning with Misclassified Response in Binary Regression / Apprentissage Q avec réponses mal classées dans une régression binaire

16:15-16:30 **Nathaniel David Osgood** (University of Saskatchewan) **Jeremy Eng** (Saskatchewan Polytechnic)
Effective Use of PMCMC for Daily Epidemiological Monitoring and Reporting: Methodological Lessons / Utilisation efficace du PMCMC pour la surveillance et le rapport épidémiologiques quotidiens : leçons méthodologiques

Thursday June 2 / jeudi 2 juin

11:00-12:30 **Invited / Sur invitation** (abstract/résumé 213)

2022 CRM-SSC Prize in Statistics Invited Address
Allocution du récipiendaire du prix CRM-SSC en statistique 2022

Chair/Président: David Haziza

11:00-12:00 **Pengfei Li** (University of Waterloo)
Density ratio model and its new applications / Modèle du rapport de densité et nouvelles applications

---

11:00-12:30 **Invited / Sur invitation** (abstract/résumé 214)

DataFest in Canada
DataFest au Canada

Chair/Président: Samantha-Jo Caetano
Organizer/Responsable: Samantha-Jo Caetano
Sponsor/Commanditaires: Statistical Education Section / Le Groupe d'éducation en statistique

11:00-11:30 **Nathan A. Taback** (University of Toronto)
ASA DataFest@UofT / ASA DataFest@UofT

11:30-12:00 **Karen Buro** (MacEwan University) **Jordan A. Slessor** (MacEwan University)
DataFests in Edmonton, 2019 and 2022, Two Perspectives / DataFests à Edmonton, 2019 et 2022, deux points de vue

12:00-12:30 **Shojaeddin Chenouri** (University of Waterloo)
ASA DataFest: The Waterloo Chapter / ASA DataFest : Le chapitre de Waterloo

---

11:00-12:30 **Invited / Sur invitation** (abstract/résumé 216)

Sports Analytics
Analyses sportives

Chair/Président: Tim B. Swartz
Organizer/Responsable: Shirley E. Mills
Sponsor/Commanditaires: Data Science and Analytics Section / Le Groupe de science des données et analytique

11:00-11:30 **Brian Macdonald** (Yale University)
Age, Experience, and Player Performance / Âge, expérience et performance des joueurs

11:30-12:00 **Alexander Hinton** (Vancouver Whitecaps Football Club)
Data Science at the Vancouver Whitecaps / La science des données chez les Whitecaps de Vancouver

12:00-12:30 **Lucas Friesen** (Canadian Tire Bank)
Owning The Podium: Supporting Team Canada Through Analytics / À nous le podium : soutenir l'équipe canadienne par l'analytique

---

11:00-12:30 **Invited / Sur invitation** (abstract/résumé 218)

Statistical Applications in P&C Insurance
Applications statistiques dans les assurances IARD

Chair/Président: Mathieu Pigeon
Organizer/Responsable: Mathieu Pigeon
Sponsor/Commanditaires: Actuarial Science Section / Le Groupe de science actuarielle

11:00-11:30 **Anas Abdallah** (McMaster University)
An Aggregate Trend Renewal Micro Model for Loss Reserving, with Trend, Inflation and Discount. / Un micro-modèle de renouvellement pour le provisionnement des pertes, avec tendance, inflation et escompte.

11:30-12:00 **Andrei L. Badescu** (University of Toronto) **Tsz Chai Fung** (Georgia State University) **X. Sheldon Lin** (University of Toronto)
Fitting censored and truncated regression data using the Mixture of Experts models / Ajustement de données de régression censurées et tronquées à l'aide de modèles de mélange d'experts

12:00-12:30 **Juan-Sebastian Yanez** (Université du Québec à Montréal)
Parametric Outstanding Claim Payment Count Modelling Through a Dynamic Claim Score / Modélisation paramétrique du nombre de paiements de sinistres en suspens grâce à un score de sinistres dynamique

---

11:00-12:30 **Contributed / Communications libres** (abstract/résumé 220)

Causal Inference and Causal Mediation Analysis
Inférence causale et analyse de médiation causale

Chair/Président: Mireille Schnitzer

11:00-11:15 **Mariia Samoilenko** (Université du Québec à Montréal) **Geneviève Lefebvre** (Université du Québec à Montréal)
On the Power to Detect a Natural Indirect Effect in Causal Mediation Analysis with a Categorical Mediator and a Binary Outcome / Sur la puissance à détecter un effet naturel indirect dans l'analyse de médiation causale avec un médiateur catégoriel et une réponse binaire

11:15-11:30 **Md Rashedul Hoque** (Simon Fraser University) **Yi Qian** (University of British Columbia) **Lawrence McCandless** (Simon Fraser University) **J. Antonio Aviña-Zubieta** (University of British Columbia) **Mary A. De Vera** (University of British Columbia) **Hui Xie** (Simon Fraser University)
An Index of Sensitivity to Non-Exchangeability / Indice de sensibilité à la non-échangeabilité

11:30-11:45 **Eric Morenz** (University of Washington)
Statistical Anatomy of Autopsy Studies / Anatomie statistique des études d'autopsie

11:45-12:00 **Blair Bilodeau** (University of Toronto) **Linbo Wang** (University of Toronto) **Daniel M. Roy** (University of Toronto)
Adaptively Exploiting d-Separators with Causal Bandits / Exploitation adaptative des séparateurs d avec des bandits causaux

12:00-12:15 **Lijia Wang** (University of Waterloo) **Yeying Zhu** (University of Waterloo) **Richard J. Cook** (University of Waterloo)
A Doubly Robust Joint Modelling Approach of Multiple Uncausally Correlated Mediators / Approche de modélisation conjointe doublement robuste de multiples médiateurs corrélés de manière non causale

---

11:00-12:15 **Contributed / Communications libres** (abstract/résumé 223)

New Developments in Survival Analysis
Nouveaux développements en analyse de survie

Chair/Président: Olli Saarela

11:00-11:15 **Changchang Xu** (University of Toronto) **Changchang Xu** (University of Toronto; Lunenfeld-Tanenbaum Research Institute, Sinai Health) **Shelley B. Bull** (University of Toronto; Lunenfeld-Tanenbaum Research Institute, Sinai Health)
Improving Mixture Cure Modelling of Molecular Genetic Biomarkers in Cancer Prognosis by Penalized Likelihood with Profile Likelihood Interval Estimation / Améliorer la modélisation du mélange avec taux de guérison des biomarqueurs génétiques moléculaires dans le pronostic du cancer par vraisemblance pénalisée avec une estimation des intervalles de vraisemblance du profil

11:15-11:30 **Shenita Pramij** (Memorial University of Newfoundland) **Candemir Cigsar** (Memorial University of Newfoundland)
A Dynamic Model for the Analysis of Recurrent Events with Application to Epidemic Data / Un modèle dynamique pour l'analyse d'événements récurrents avec application aux données épidémiques

11:30-11:45 **Shakhawat Hossain** (University of Winnipeg) **Jody Krahn** (Statistics Canada) **Shahedul Khan** (University of Saskatchewan)
An Efficient Estimation Approach to Joint Modelling of Longitudinal and Survival Data / Une approche d'estimation efficace pour la modélisation conjointe de données longitudinales et de survie

11:45-12:00 **Awa Diop** (Université Laval) **Denis Talbot** (Université Laval) **Caroline Sirois** (Université Laval)
History-Restricted Marginal Structural Model and Latent Class Growth Modeling of Treatment Trajectories for a Time-Dependent Outcome / Modèles structurels marginaux à historique restreint et modèles d'analyse de trajectoires groupées pour une issue qui varie dans le temps

12:00-12:15 **Denis Larocque** (HEC Montréal) **Weichi Yao** (New York University) **Halina Frydman** (New York University) **Jeffrey S. Simonoff** (New York University)
Ensemble Methods for Survival Function Estimation with Time-Varying Covariates / Méthodes d'ensemble pour l'estimation de la fonction de survie avec covariables qui varient dans le temps

---

12:30-13:30 **Invited / Sur invitation** (abstract/résumé 226)

Information on NSERC Competition Results and Discovery Grant Preparation
Information sur les résultats du concours du CRSNG et la préparation des subventions à la découverte

Chair/Président: Joanna Elizabeth Mills Flemming
Organizer/Responsable: Henrik Stryhn
Sponsor/Commanditaires: NSERC and the SSC Research Comittee / CRSNG et le comité de la recherche de la SSC

12:30-13:30 **Adele Ngi-Song** (NSERC) **Caroline Bicker** (NSERC) **Aurélie Labbe** (HEC Montreal)
Information on NSERC Competition Results and Discovery Grant Preparation / Information sur les résultats du concours du CRSNG et la préparation des subventions à la découverte

---

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

Recent Advances in Causal Inference: From Theory to Practice
Progrès récents en inférence causale : de la théorie à la pratique

Chair/Président: Linbo Wang
Organizer/Responsable: Linbo Wang
Sponsor/Commanditaires: Canadian Statistical Sciences Institute (CANSSI) / Institut canadien des sciences statistiques (INCASS)

13:30-14:00 **Yuexia Zhang** (University of Toronto) **Jianhua Hu** (Columbia University)
High dimensional mediation analysis for microbiome data / Analyse de médiation en haute dimension pour données du microbiome

14:00-14:30 **Geneviève Lefebvre** (Université du Québec à Montréal) **Miguel Caubet Fernandez** (Université du Québec à Montréal) **Mariia Samoilenko** (Université du Québec à Montréal)
Investigating the Performance of the Exact Estimator for Causal Mediation Analysis of Binary Outcomes and Binary Mediators in Case-control Designs / Étude de la performance de l'estimateur exact pour l'analyse de médiation causale pour les réponses et médiateurs binaires dans les devis cas-témoins

14:30-15:00 **Dehan Kong** (University of Toronto) **Zhenhua Lin** (National University of Singapore) **Linbo Wang** (University of Toronto)
Causal Inference on Distribution Functions / Inférence causale sur des fonctions de distribution

---

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

Frontier Statistical Research for Medical and Biological Data
Recherche statistique de pointe pour les données médicales et biologiques

Chair/Président: Xuekui Zhang

13:30-14:00 **Lynn Lin** (Duke University)
Multi-source Single-cell Data Integration by MAW Barycenter for Gaussian Mixture Models / Intégration de données à cellule unique et sources multiples par barycentre MAW pour les modèles de mélanges gaussiens

14:00-14:30 **Lihui Zhao** (Northwestern University)
Dynamic Risk Prediction for Cardiovascular Events / Prédiction de risque dynamique pour les événements cardiovasculaires

14:30-15:00 **Kailun Bai** (University of Victoria)
scSorterDL: a cell type annotation tool for single-cell RNA sequencing data / scSorterDL : outil d'annotation du type de cellule pour données de séquençage d'ARN unicellulaire

---

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

Statistical Challenges in Deep Learning
Défis statistiques en apprentissage profond

Chair/Président: Vahid Partovi Nia
Organizer/Responsable: Vahid Partovi Nia

13:30-14:00 **Masoud Asgharian** (McGill University)
Machine Learning and Neural Networks: Foundations and Some Fundamental Questions / Apprentissage automatique et réseaux neuronaux : fondements et questions fondamentales

14:00-14:30 **Ali Ghodsi** (University of Waterloo) **Mojtaba Valipour** (Cornell University) **Bowen You** (Cornell University) **Maysum Panju** (Cornell University)
SymbolicGPT: A Generative Transformer Model for Symbolic Regression / SymbolicGPT : Un modèle de transformateur génératif pour la régression symbolique

14:30-15:00 **Yaoliang Yu** (University of Waterloo) **Dockhorn Tim** (University of Waterloo) **Eyyüb Sari** (Huawei Noah's Ark Lab) **Mahdi Zolnouri** (Huawei Noah's Ark Lab) **Vahid Nia** (Huawei Noah's Ark Lab)
Demystifying and Generalizing BinaryConnect / Démystification et généralisation de la méthode BinaryConnect

---

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

Real-World Challenges and Recent Statistical Developments
Défis du monde réel et développements statistiques récents

Chair/Président: Yingwei (Paul) Peng
Organizer/Responsable: Joan X. Hu
Sponsor/Commanditaires: ICSA-Canada Chapter / Chapitre canadien de l'ICSA

13:30-14:00 **Trevor Thomson** (Simon Fraser University) **X. Joan Hu** (Simon Fraser University) **Bohdan Nosyk** (Simon Fraser University)
Recent Advances in Modelling Time-to-Event Data with Internal Covariates / Avancées récentes dans la modélisation de données de durée de vie avec covariables internes

14:00-14:30 **Leilei Zeng** (University of Waterloo)
Response Dependent Sampling in Observational Cohort Studies / Échantillonnage dépendant de la réponse dans les études observationnelles de cohortes

14:30-15:00 **Rong Chen** (Rutgers University)
Two Factor Models for High-Dimensional Tensor Time Series / Deux modèles factoriels pour séries chronologiques tensorielles en haute dimension

---

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

Stochastic Partial Differential Equations
Équations différentielles partielles stochastiques

Chair/Président: Yaozhong Hu
Organizer/Responsable: Yaozhong Hu
Sponsor/Commanditaires: Probability Section / Le Groupe de probabilité

13:30-14:00 **Xia Chen** (University of Tennessee)
Necessary and sufficient condition for the solvability of the hyperbolic Anderson models with Gaussian noise that is fractional in times / Condition nécessaire et suffisante pour la solvabilité des modèles hyperboliques d'Anderson avec bruit gaussien fractionné en temps

14:00-14:30 **Jian Song** (Shandong University) **Guanglin Rang** (Wuhan University)
The Scaling Limit of a Long-range Random Walk in Correlated Random Medium / La limite d'échelle d'une promenade aléatoire de longue portée dans un milieu aléatoire corrélé

14:30-15:00 **Samy Tindel** (Purdue University)
A coupling between Sinai's random walk and Brox diffusion / Couplage entre la marche aléatoire de Sinai et la diffusion de Brox

---

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

Statistical Analysis of Imperfect Data
Analyse statistique des données imparfaites

Chair/Président: Liqun Diao

13:30-13:45 **Dylan Z. Spicker** (University of Waterloo) **Michael Wallace** (University of Waterloo) **Grace Y. Yi** (University of Western Ontario)
Nonparametric Simulation Extrapolation for Measurement Error Models / Extrapolation par simulation non paramétrique pour des modèles d'erreur de mesure

13:45-14:00 **Jingyu Cui** (Western University) **Grace Y. Yi** (Western University)
Multivariate Regression Model with Measurement Error / Modèle de régression multivarié avec erreur de mesure

14:00-14:15 **Alexandra S. Bushby** (University of Toronto) **Eleanor M. Pullenayegum** (The Hospital for Sick Children)
Measurement Error in Longitudinal Data with Irregular Observation / Erreur de mesure des données longitudinales avec des observations irrégulières

14:15-14:30 **Melina Ribaud** (HEC Montréal) **Aurélie Labbe** (HEC Montreal) **Karim Oualkacha** (Université du Québec à Montréal)
Imputation in genetic methylation studies: A linear model of coregionalization (LMC) with informative covariates / Problèmes d'imputation dans les études génétiques de méthylation : un modèle de corrégionalisation linéaire (LMC) avec covariables.

14:30-14:45 **Mei Dong** (University of Toronto) **Aya A. Mitani** (University of Toronto)
Multiple imputation methods for missing multilevel ordinal outcomes / Méthodes d'imputation multiple de résultats manquants ordinaux à plusieurs niveaux

14:45-15:00 **Jinhui Ma** (McMaster University) **Parminder Raina** (McMaster University) **Lauren Griffith** (McMaster University) **Mylinh Duong** (McMaster University) **Alexandra Mayhew** (McMaster University) **Carol Bassim** (McMaster University) **Chris Verschoor** (Health Sciences North Research Institute) **Lehana Thabane** (McMaster University) **Hon-Yiu So** (Oakland University)
Imputation of Missing Spirometry Data in Population-based Studies / Imputation de données de spirométrie manquantes en études sur la population

---

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

New Statistical Methods in Genetic Studies
Nouvelles méthodes statistiques pour les études génétiques

Chair/Président: Qingrun Zhang

13:30-13:45 **Patrick Fournier** (Université du Québec à Montréal)
Accounting for Epistasis in PRSs Through the Coalescent / Prise en compte de l'épistasie dans les SRP grâce au coalescent

13:45-14:00 **Olga Vishnyakova** (Simon Fraser University) **Angela Brooks-Wilson** (Simon Fraser University, BC Cancer) **Lloyd T. Elliott** (Simon Fraser University)
Analysis of Homeostasis in Health / Analyse de l'homéostasie dans la santé

14:00-14:15 **Quan Long** (University of Calgary)
An Expression-directed Linear Mixed Model (edLMM) Discovering Low-effect Genetic Variant / Modèle mixte linéaire avec expression dirigée (edLMM) pour découvrir les variants génétiques à faible effet

14:15-14:30 **Guan Wang** (University of Toronto: Dalla Lana School of Public Health)
Two-Phase Design for Regional Genetic Sequencing Using Polygenic Risk Scores / Plan en deux phases pour un séquençage génétique régional utilisant des scores de risque polygénique

14:30-14:45 **Ting Zhang** (McGill University) **Jerome Dockes** (McGill University) **Nikhil Bhagwat** (McGill University) **Clara Moreau** (Pasteur Institute) **Celia Greenwood** (McGill University) **Jean-Baptiste Poline** (McGill University)
Kernel Selection for Linear Mixed Effect model on Estimating Variance Explained / Sélection de noyaux pour modèle linéaire à effets mixtes sur l'estimation de la variance expliquée

---

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

Statistical Analysis of Functional Data and Time Series Data
Analyse statistique des données fonctionnelles et des données de séries chronologiques

Chair/Président: Sharandeep Singh Pandher

13:30-13:45 **Thai-Son Tang** (University of Toronto) **Zhihui (Amy) Liu** (Princess Margaret Cancer Centre, University Health Network; Dalla Lana School of Public Health, University of Toronto) **Olli Saarela** (Dalla Lana School of Public Health, University of Toronto)
A marginal structural model for normal tissue complication probability / Modèle structurel marginal pour la probabilité de complication des tissus normaux

13:45-14:00 **Haixu Alex Wang** (Simon Fraser University) **Jiguo Cao** (Simon Fraser University)
Functional Nonlinear Learning / Apprentissage non linéaire fonctionnel

14:00-14:15 **Shivani Bhardwaj** (University of Manitoba) **Yuliya V. Martsynyuk** (University of Manitoba)
Finite-sample properties and applicability of functional CLT based confidence intervals for a population mean / Propriétés d'échantillon fini et applicabilité d'intervalles de confiance fonctionnels basés sur le théorème central limite d'une moyenne de population

14:15-14:30 **Boyi Hu** (Simon Fraser University) **Hua Liu** (Xi'an Jiaotong University) **Jinhong You** (Shanghai University of Finance and Economics) **Jiguo Cao** (Simon Fraser University)
Convolution Smoothed Functional Linear Quantile Regression with Locally Sparse Adaptation / Régression quantile linéaire fonctionnelle lissée par convolution avec adaptation localement éparse

14:30-14:45 **Chi-Kuang Yeh** (University of Waterloo) **Gregory Rice** (University of Waterloo) **Joel A. Dubin** (University of Waterloo)
Projection Based Model Validation and Identification Methods for Functional Time Series / Méthodes de validation et d'identification de modèles basés sur la projection pour séries chronologiques fonctionnelles

14:45-15:00 **Skye Paphora Griffith** (Queen's University)
Transfer Function Estimates and their Phase Distributions under the Multitaper Method / Estimations de fonctions de transfert et de leurs distributions de phase selon la méthode multitaper

---

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

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

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

15:30-15:52 **Peijun Sang** (University of Waterloo) **Zuofeng Shang** (New Jersey Institute of Technology) **Pang Du** (Virginia Polytechnic Institute and State University)
Statistical Inference for Functional Linear Quantile Regression / Inférence statistique pour la régression quantile linéaire fonctionnelle

15:52-16:14 **Yafei Wang** (University of Alberta)
M-estimation for varying coefficient model with functional response in reproducing kernel Hilbert space / Estimation M d'un modèle à coefficient variable avec réponse fonctionnelle dans un espace de Hilbert à noyau reproducteur

16:14-16:36 **Evan Sidrow** (The University of British Columbia) **Nancy Heckman** (University of British Columbia) **Sarah M.E. Fortune** (Dalhousie University) **Andrew W. Trites** (University of British Columbia) **Ian Murphy** (University of Florida) **Marie Auger-Mèthè** (University of British Columbia)
Modelling Functional Data with Hierarchical Hidden Markov Models: Applications to Animal Movement / Modélisation de données fonctionnelles avec modèles de Markov cachés hiérarchiques : applications au mouvement des animaux

16:36-16:58 **Haolun Shi** (Simon Fraser University) **Jiguo Cao** (Simon Fraser University)
Robust Regression-Based Functional Principal Component Analysis / Analyse en composantes principales fonctionnelle par régression robuste

---

15:30-17:00 **Invited / Sur invitation** (abstract/résumé 251)

Recent Advances in Methodologies and Applications of Innovative Survival Models
Progrès récents en méthodes et applications de modèles de survie innovants

Chair/Président: Longhai Li
Organizer/Responsable: Longhai Li

15:30-16:00 **Yingwei (Paul) Peng** (Queen's University) **Chyong-Mei Chen** (National Yang Ming Chiao Tung University) **Pao-sheng Shen** (Tunghai University) **Hsin-Jen Chen** (National Yang Ming Chiao Tung University)
Length-Biased and Interval-Censored Data with a Cure Fraction / Données biaisées en longueur et censurées par intervalle avec un taux de guérison

16:00-16:30 **Shahedul Khan** (University of Saskatchewan)
Accelerated Failure Time Models for Recurrent Event Data Analysis and Joint Modeling / Modèles à temps d'échec accélérés pour l'analyse de données d'événements récurrents et de modélisation conjointe

16:30-17:00 **Cindy Xin Feng** (Dalhousie University) **Tingxuan Wu** (University of Saskatchewan) **Longhai Li** (University of Saskatchewan)
A Comparative Study of R packages for Semiparametric Shared Frailty Models / Étude comparative de paquets R pour des modèles semi-paramétriques à fragilités partagées

---

15:30-17:00 **Invited / Sur invitation** (abstract/résumé 253)

Challenges and Examples in Data Science Consultation
Défis et exemples sur la consulation en sciences des données

Chair/Président: Jean-François Plante
Organizer/Responsable: Jean-François Plante
Sponsor/Commanditaires: Business and Industrial Statistics Section / Le Groupe de statistique industrielle et de gestion

15:30-16:00 **Steve Kanters** (RainCity Analytics)
Challenges and Highlights of Data Science Consulting / Défis et points forts de la consultation en science des données

16:00-16:30 **Ghislene Zerguini** (HEC Montréal)
Setting Your Data Workforce up for Success / Comment contribuer au succès des responsables de données en entreprise

16:30-17:00 **Sarah Legendre Bilodeau** (Videns Analytics) **Sébastien Duguay** (Videns Analytics)
Challenges and Examples in Data Science Consultation / La consultation en science des données - défis et exemples

---

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

Recent Advances on Approaches for Statistics in Biosciences
Progrès récents des approches de la statistique dans les biosciences

Chair/Président: Joan X. Hu
Organizer/Responsable: Joan X. Hu
Sponsor/Commanditaires: Biostatistics Section / Le Groupe de biostatistique

15:30-16:00 **Hongzhe Lee** (University of Pennsylvania)
Transfer Learning in High-dimensional Linear Regression and Graphical Models / Apprentissage par transfert dans une régression linéaire de haute dimension et des modèles graphiques

16:00-16:30 **Juxin Liu** (University of Saskatchewan)
Bias Analysis for Misclassification Errors in both the Response Variable and Covariate / Analyse de biais pour les erreurs de classification dans la variable de réponse et la covariable

16:30-17:00 **Richard J. Cook** (University of Waterloo) **Jerald F. Lawless** (University of Waterloo)
Analysis of Life History Data Obtained from Biased Sampling and Observation Schemes / Analyse de données de cycle de vie tirées d'échantillonnage et de schémas d'observation biaisés

---

15:30-17:00 **Invited / Sur invitation** (abstract/résumé 257)

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

Chair/Président: Félix Camirand Lemyre
Organizer/Responsable: Félix Camirand Lemyre
Sponsor/Commanditaires: Committee on New Investigators / Le Comité des nouveaux chercheurs

15:30-16:00 **Juliana Schulz** (HEC Montréal) **Erica E.M. Moodie** (McGill University)
Doubly Robust Estimation of Optimal Dosing Strategies / Estimation doublement robuste des stratégies de dosage optimales

16:00-16:30 **Kevin McGregor** (York University) **Nneka Okaeme** (York University)
Proportionality-Based Association Measures in Count-Based Compositional Data / Mesures d'association basées sur la proportionnalité pour des données de compatage compositionnelles

16:30-17:00 **Samantha-Jo Caetano** (University of Toronto) **Rohan Alexander** (University of Toronto)
Further Developments of a Toolkit for Learning R at All Levels. / Le développement supplémentaire d'une boîte à outils pour l'apprentissage du langage R à tous les niveaux

---

15:30-17:00 **Contributed / Communications libres** (abstract/résumé 259)

Copula-based Methods
Méthodes basées sur les copules

Chair/Président: Wesley S. Burr

15:30-15:45 **Robert Zimmerman** (University of Toronto) **Vianey Leos Barajas** (University of Toronto) **Radu V. Craiu** (University of Toronto)
Copula Modelling of Serially Correlated Multivariate Data with Hidden Structures / Modélisation par copules de données multivariées sériellement corrélées avec des structures cachées

15:45-16:00 **Guanjie Lyu** (University of Windsor) **Mohamed Belalia** (University of Windsor)
Testing Symmetry for Bivariate Copulas using Bernstein Polynomials / Test de symétrie pour copules bivariées à l'aide des polynômes de Bernstein

16:00-16:15 **H. Roland G. Dossa** (Université du Québec à Montréal)
Generalized Functional Linear Mixed Models for Binary Traits in Family-Based Designs via Copulas / Modèles mixtes linéaires fonctionnels généralisés pour les traits binaires dans les devis basés sur la famille via copules

16:15-16:30 **Xinyao Fan** (The University of British Columbia)
Proxies in High-dimensional Factor Copula Models / Proxys dans les modèles de copules factorielles à haute dimension

16:30-16:45 **Serge B. Provost** (The University of Western Ontario) **Yishan Zang** (Western University)
On Modeling Multivariate Data from Marginal Distributions / Modélisation de données multivariées à partir de distributions marginales

16:45-17:00 **Salah El Adlouni** (Université de Moncton) **A. Boukili-Makhoukhi** (Université de Moncton) **W. El Hannoun** (Université Mohamed) **A. Zoglat** (Université Mohamed)
Vine Copulas to Estimate Intensity-Duration-Frequency Curves / Copules en vignes et estimation des courbes Intensité-Durée-Fréquence

---

15:30-17:00 **Contributed / Communications libres** (abstract/résumé 262)

Longitudinal Data Analysis
Analyse des données longitudinales

Chair/Président: Zihang Lu

15:30-15:45 **Xiawen Zhang** (University of Toronto: Dalla Lana School of Public Health) **Eleanor M. Pullenayegum** (University of Toronto & SickKids)
The Bias of Parameters in Inverse-Intensity Weighted GEEs when People Without a Visit are Excluded / Biais des paramètres dans les EEG à pondération par intensité inverse lorsque les personnes sans visite sont exclues

15:45-16:00 **Rose Garrett** (University of Toronto)
Why Recommended Visit Intervals should be Extracted when Conducting Longitudinal Analyses using Electronic Health Record Data / Pourquoi extraire les intervalles de rendez-vous recommandés dans les analyses longitudinales à l'aide de données de dossiers médicaux électroniques

16:00-16:15 **Omidali Aghababaei Jazi** (University of Toronto Mississauga) **Eleanor M. Pullenayegum** (Hospital for Sick Children (Sickkids))
Dynamic Prediction for Longitudinal Data with Irregular and Outcome-dependent Follow-up / Prédiction dynamique pour données longitudinales avec suivi irrégulier et dépendant des résultats

16:15-16:30 **Menelaos Konstantinidis** (University of Toronto: Dalla Lana School of Public Health) **Lily S. H. Lim** (Children's Hospital Research Institute of Manitoba, University of Manitoba) **Eleanor M. Pullenayegum** (Dalla Lana School of Public Health, University of Toronto)
Designing an Accelerated Longitudinal Cohort for the Employment trajectories of Systemic Lupus Erythematosus Patients: A simulation Study / Conception de cohortes longitudinales accélérées pour des trajectoires d'emploi de patients atteints de lupus érythémateux systémique : une étude par simulations

16:30-16:45 **Lulu Guo** (Simon Fraser University - Burnaby, BC) **Hui Xie** (Faculty of Health Sciences, Simon Fraser University)
A Latent Variable Model for Longitudinal Trials with Multiple Endpoints and Treatment Noncompliance: an Application to a Study of Arthritis Health Journal / Modèle à variables latentes pour des essais longitudinaux en présence de multiples critères d'évaluation et de non-conformité au traitement : une application à une étude du Arthritis Health Journal

16:45-17:00 **Marzia Angela Cremona** (Université Laval) **Huy Dang** (The Pennsylvania State University) **Francesca Chiaromonte** (The Pennsylvania State University)
smoothEM: A New Approach for the Simultaneous Assessment of Smooth Curves and Spikes / smoothEM : une nouvelle approche pour l'évaluation simultanée des courbes lisses et des pics

---

15:30-17:00 **Contributed / Communications libres** (abstract/résumé 266)

Recent Advances and Applications of Machine-learning Methods
Progrès récents et applications des méthodes d'apprentissage automatique

Chair/Président: Ilia Sucholutsky

15:30-15:45 **Li Yi** (University of Western Ontario)
How Self-Supervised Contrastive Learning Helps Learning with Label Noise / Comment l'apprentissage auto-supervisé contrastif aide à apprendre lorsque les étiquettes sont bruitées

15:45-16:00 **Cansu Alakus** (HEC Montréal) **Denis Larocque** (HEC Montréal) **Aurélie Labbe** (HEC Montreal)
RFpredInterval: An R Package for Prediction Intervals with Random Forests and Boosted Forests / RFpredInterval : une bibliothèque R pour les intervalles de prévisions avec forêts aléatoires et forêts améliorées

16:00-16:15 **Leslie G. Fell** (University of Guelph) **Olaf Berke** (University of Guelph) **Lorna E. Deeth** (University of Guelph) **Lise A. Trotz-Williams** (Wellington-Dufferin-Guelph Public Health)
Predicting Bacterial Contamination of Private Well Water in Wellington-Dufferin-Guelph, Ontario / Prédiction de contamination bactérienne de l'eau de puits privé à Wellington-Dufferin-Guelph, en Ontario

16:15-16:30 **Henrik Stryhn** (University of Prince Edward Island)
A Random Effects Model for Sparse Cross-Classification Data / Modèle à effets aléatoires pour données éparses de classification croisée

16:30-16:45 **Alessandro Maria Maria Selvitella** (Purdue University Fort Wayne) **Kathleen Lois Foster** (Department of Biology - Ball State University)
Anolis Ecomorph Biomechanics across Arboreal Environments: What can Machine Learning tell us about Behavioral Plasticity in Lizards? / Biomécanique des écomorphes d'Anolis dans les environnements arboricoles : que peut nous apprendre l'apprentissage automatique sur la plasticité comportementale des lézards ?

16:45-17:00 **Yunfeng Yang** (University of Waterloo)
Multimodel Bayesian Analysis of Load Duration Effects in Lumber Reliability / Analyse bayésienne multimodèle des effets de la durée de chargement dans la fiabilité du bois d'œuvre

Friday June 3 vendredi 3 juin

11:00-12:30 **Invited / Sur invitation** (abstract/résumé 270)

Statistical Modelling and Computational Intelligence in Genomics
Modélisation statistique et intelligence informatique en génomique

Chair/Président: You Liang
Organizer/Responsable: You Liang
Sponsor/Commanditaires: Biostatistics Section / Le Groupe de biostatistique

11:00-11:30 **Wenqing He** (University of Western Ontario) **Juan Xiong** (Shengzhen University)
Identification of Survival Relevant Genes with Measurement Error in Gene Expression Incorporated / Identification de gène pertinent de survie avec erreur de mesure dans l'expression génique intégrée

11:30-12:00 **Xuekui Zhang** (University of Victoria)
Automated Cell-Type Annotation using scRNA-seq Data / Annotation automatique des types de cellules à l'aide de données de séquençage de l'ARN en cellule unique

12:00-12:30 **Liangliang Wang** (Simon Fraser University) **Shijia Wang** (Nankai University) **Alexandre Bouchard-Côté** (University of British Columbia)
Efficient Sequential Monte Carlo Methods for Bayesian Phylogenetic Inference / Méthodes de Monte Carlo séquentielles efficaces pour l'inférence phylogénétique bayésienne

---

11:00-12:30 **Invited / Sur invitation** (abstract/résumé 272)

50 Years of Statistical Community in Canada
50 ans de communauté statistique au Canada

Chair/Président: Rhonda J Rosychuk
Organizer/Responsable: Melody Ghahramani
Sponsor/Commanditaires: 50th Anniversary Committee / Le Comité du 50e anniversaire de la SSC

11:00-12:00 **David R. Bellhouse** (University of Western Ontario) **Christian Genest** (McGill University)
A Glimpse into SSC History / Un aperçu de l'histoire de la SSC

12:00-12:30 **Mary E. Thompson** (University of Waterloo)
Discussion / Discussion

11:00-12:30 **Invited / Sur invitation** (abstract/résumé 273)

Fisheries Statistics
Statistiques de pêche

Chair/Président: Joanna Elizabeth Mills Flemming
Organizer/Responsable: Joanna Elizabeth Mills Flemming

11:00-11:30 **Jonathan Babyn** (Dalhousie University)
Estimating both Population Effective and Census Size using Close-Kin Mark Recapture / Estimation de la taille effective de la population et de la taille de la population recensée à l'aide du marquage-recapture d'espèces qui ont un lien de parenté proche

11:30-12:00 **Ethan Lawler** (Dalhousie University) **Chris Field** (Dalhousie University) **Joanna Mills Flemming** (Dalhousie University)
Species Distribution Modelling using Spatio-temporal Nearest Neighbour Gaussian Processes / Modélisation de la distribution des espèces à l'aide de processus gaussiens spatio-temporels du plus proche voisin

12:00-12:30 **Andrea Perreault** (Fisheries and Oceans Canada) **Noel Cadigan** (Fisheries and Marine Institute of Memorial University)
Profile Likelihood Diagnostics for Integrated State-Space Models / Diagnostics du profile de vraisemblance pour les modèles d'espace d'états intégrés

11:00-12:30 **Invited / Sur invitation** (abstract/résumé 275)

2022 Pierre Robillard Award Address
Allocution du récipiendaire du prix Pierre-Robillard 2022

Chair/Président: Yingwei (Paul) Peng

11:00-12:00 **Janie Coulombe** (McGill University)
Causal inference on the marginal effect of an exposure: Addressing biases due to covariate-driven monitoring times and confounders / Inférence causale sur l'effet marginal d'une exposition : Comment tenir compte des biais dus aux temps de visite qui dépendent du patient et aux facteurs confondants E EF

11:00-12:30 **Invited / Sur invitation** (abstract/résumé 276)

Stochastic Population Models
Modèles stochastiques de population

Chair/Président: Xiaowen Zhou, Shui Feng
Organizer/Responsable: Xiaowen Zhou
Sponsor/Commanditaires: Probability Section / Le Groupe de probabilité

11:00-11:30 **Shui Feng** (McMaster University)
Kingman Coalescent and Bayesian Nonparametrics / Coalescent de Kingman et approche bayésienne non paramétrique E E

11:30-12:00 **Lam Ho** (Dalhousie University)
Theory of Ancestral State Reconstruction / Théorie de reconstruction d'état ancestral E E

12:00-12:30 **Xiaowen Zhou** (Concordia University)
Continuous-state Nonlinear Branching Processes / Processus de branchement non linéaires à l'état continu E E

11:00-12:30 **Contributed / Communications libres** (abstract/résumé 278)

Recent Developments in Clustering and Classification
Développements récents en matière de classification et de regroupement

Chair/Président: Utkarsh J. Dang

11:00-11:15 **Andrea Payne** (Carleton University) **Anjali Silva** (University of Toronto) **Steven Rothstein** (University of Guelph) **Paul D. McNicholas** (McMaster University) **Sanjeena Dang (Subedi)** (Carleton University)
Clustering High Dimensional Multivariate Count Data Using a Family of Mixtures of Multivariate Poisson Log-Normal Distributions / Regroupement de données de dénombrement multivariées à haute dimension à l'aide d'une famille de mélanges de distributions log-normales multivariées de Poisson E E

11:15-11:30 **Zahra Aghahosseinalishirazi** (Western University) **Dr Camila De Souza** (The University of Western Ontario)
Clustering Single-Cell RNA Sequencing Data via the Expectation-Maximization Algorithm / Regroupement des données de séquençage de l'ARN de cellules uniques avec l'algorithme espérance-maximisation E E

11:30-11:45 **Ashani N. Wickramasinghe** (University of Manitoba) **Saman Muthukumarana** (University of Manitoba) **Dan Loewen** (ioAirFlow) **Matt Schaubroeck** (ioAirFlow)
Temperature Clusters in Commercial Buildings Using K-means and Time Series Clustering / Regroupement de températures dans les bâtiments commerciaux à l'aide de K-moyennes et de séries chronologiques E E

11:45-12:00 **Michelle Wu** (University of Toronto) **Hyejung Jung** (University of Toronto)
Correlation Analysis and Machine Learning-Based Approaches to Assess Depression Severity / Analyse de corrélation et approches basées sur l'apprentissage machine pour évaluer la gravité de la dépression

12:00-12:15 **Michael John Ilagan** (McGill University) **Carl F. Falk** (McGill University)
Supervised Components, Unsupervised Mixing Proportions: Detection of Bots in Likert-type Surveys / Composantes supervisées, proportions du mélange non supervisé : la détection des bots dans les enquêtes de type Likert

12:15-12:30 **Wanhua Su** (MacEwan University)
Classification With Imbalanced Data / Classification avec données déséquilibrées

---

11:00-12:30 **Contributed / Communications libres** (abstract/résumé 282)

Recent Advances in Regression Methods
Progrès récents en méthodes de régression

Chair/Président: Max Turgeon

11:00-11:15 **Jason Hou-Liu** (University of Waterloo) **Ryan P. Browne** (University of Waterloo)
Fast Estimation of Generalized Linear Models via Sketching / Estimation rapide de modèles linéaires généralisés par esquisse

11:15-11:30 **Hui Guo** (Western University)
Variable Selection for Logistic Regression Models with Misclassified Response / Sélection de variables pour les modèles de régression logistique avec réponse classée incorrectement

11:30-11:45 **Yansan Han** (Brock University) **Mei Ling Huang** (Brock University) **William Marshall** (Brock University)
Quantile Regression Analysis on COVID-19 / Analyse de régression quantile sur la COVID-19

11:45-12:00 **Zeyu Chen** (University of Toronto: Dalla Lana School of Public Health) **Osvaldo Espin-Garcia** (University of Toronto: Dalla Lana School of Public Health)
The Perils of Ignoring the Study Design in High-Dimensional Settings: A Simulation-based Evaluation / Les dangers d'ignorer la conception de recherche dans des contextes à haute dimension : une évaluation en simulation

12:00-12:15 **Chong Gan** (University of Guelph) **Zeny Feng** (University of Guelph) **Jiahua Chen** (University of British Columbia)
Association Tests under Gaussian Mixture Regression Models / Tests d'association selon les modèles de régression de mélange gaussien

12:15-12:30 **Gunho Bae** (University of Manitoba) **Saumen Mandal** (University of Manitoba)
Optimal Experimental Designs for Estimating Parameters Independently of Each Other / Plans d'expérience optimaux pour l'estimation de paramètres indépendamment les uns des autres

---

12:30-13:30 **Invited / Sur invitation** (abstract/résumé 286)

CANSSI Programs and Plans
Programmes et plans de l'INCASS

Chair/Président: Donald Estep
Organizer/Responsable: Donald Estep
Sponsor/Commanditaires: Canadian Statistical Sciences Institute (CANSSI) / Institut canadien des sciences statistiques (INCASS)

12:30-13:30 **Andrea Benedetti** (McGill University) **Joanna Mills Flemming** (Dalhousie University) **Donald Estep** (CANSSI)
CANSSI Programs and Plans / Programmes et plans de l'INCASS

13:30-15:00 **Invited / Sur invitation** (abstract/résumé 287)

CANSSI Postdoctoral Showcase
Vitrine des boursiers postdoctoraux de l'INCASS

Chair/Président: Andrea Benedetti
Organizer/Responsable: Andrea Benedetti
Sponsor/Commanditaires: Canadian Statistical Sciences Institute (CANSSI) / Institut canadien des sciences statistiques (INCASS)

13:30-14:00 **Kaiqiong Zhao** (McGill University) **Linglong Kong** (University of Alberta) **Yanchun Bao** (University of Essex)
A Multi-Channel Fusion Framework for Statistical Learning and Inference with its Application in Multi-Omics Data Analysis / Cadre de fusion multicanal pour l'apprentissage et l'inférence statistiques et application à l'analyse de données multi-omiques Ⓔ E

14:00-14:30 **Caitlin Ward** (University of Calgary) **Rob Deardon** (University of Calgary) **Alexandra M. Schmidt** (McGill University)
Sound the alarm: modeling behavioral changes in response to epidemic intensity / Alarme ! Modélisation des changements de comportement en réponse à l'intensité d'une épidémie Ⓔ E

14:30-15:00 **Cédric Beaulac** (Simon Fraser University/University of Victoria)
Neural Network Classifiers for Features Extraction in Neuroimaging Genetics / Utiliser un réseau de neurones de classification pour extraire des variables d'imagerie cérébrale. Ⓔ E

13:30-15:00 **Invited / Sur invitation** (abstract/résumé 289)

Recent Advances in Mixture Models: Theory and Application
Avancées récentes en modèles de mélange : théorie et application

Chair/Président: Juxin Liu
Organizer/Responsable: Juxin Liu

13:30-14:00 **Zeny Feng** (University of Guelph) **Sanjeena Subedi** (Carleton University) **Stephen Bak** (University of Guelph) **Drew Neish** (University of Guelph)
Mixture of Dirichlet Multinomial (DM) Models in Microbiome Data Analysis / Mélange de modèles multinomiaux de Dirichlet (MD) dans l'analyse de données de microbiome Ⓔ E

14:00-14:30 **Abbas Khalili** (McGill University) **Tudor A. Manole** (Carnegie Mellon University)
A Group-Sort-Fuse Procedure for Estimating the Number of Components in Finite Mixture Models / Une procédure Group-Sort-Fuse pour l'estimation du nombre de composants dans des modèles de mélanges finis Ⓔ E

14:30-15:00 **Jiahua Chen** (The University of British Columbia) **Qiong Zhang** (University of British Columbia)
Gaussian Mixture Reduction based on Composite Transportation Divergence / Réduction de mélange gaussien en fonction de la divergence de transport composite Ⓔ E

13:30-15:00 **Invited / Sur invitation** (abstract/résumé 291)

2022 CJS Award Address
Allocution du récipiendaire du prix de la RCS 2022

Chair/Président: Andrei Volodin

13:30-14:30 **Li Xing** (University of Saskatchewan) **Xuekui Zhang** (University of Victoria) **Igor Burstyn** (Drexel University) **Paul Gustafson** (University of British Columbia)
The logistic Box-Cox regression helps investigate the exposure-disease relationship in epidemiological studies / La régression logistique de Box-Cox aide à étudier la relation exposition-maladie dans les études épidémiologiques Ⓔ E F

13:30-15:00 **Invited / Sur invitation** (abstract/résumé 292)

Perspectives on Open Education Resources
Perspectives sur les ressources éducatives libres

Chair/Président: Sotirios Damouras
Organizer/Responsable: Sotirios Damouras
Sponsor/Commanditaires: Statistical Education Section / Le Groupe d'éducation en statistique

13:30-14:00 **Surita Jhangiani** (The University of British Columbia)
Leveraging Open Educational Resources in Higher Education / Ressources éducatives libres dans l'enseignement supérieur

14:00-14:30 **Trevor Campbell** (The University of British Columbia) **Melissa Lee** (University of British Columbia) **Tiffany A. Timbers** (University of British Columbia)
Creating Open Resources for an Introductory Data Science Course / Création de ressources ouvertes pour un cours d'introduction à la science des données

14:30-15:00 **Toby Hodges** (The Carpentries)
Perspectives on Development and Use of Open Educational Resources at Scale / Perspectives sur le développement et l'utilisation de ressources éducatives libre à l'échelle

13:30-15:00 **Invited / Sur invitation** (abstract/résumé 294)

Data Science Applications in Computational Finance
Applications de la science des données en finance computationnelle

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

13:30-14:00 **Shelton Peiris** (The University of Sydney) **David Dowe** (Monash University) **Zheng Fang** (Monash University) **Dedi Rosadi** (University of Gadjah Mada) **Aerambamoorhty A. Thavaneswaran** (University of Manitoba)
A Novel ARFIMA-ANN Hybrid Model for Financial Time Series Forecasting / Nouveau modèle hybride de ARFIMA-RNA pour la prévision des séries chronologiques financières

14:00-14:30 **You Liang** (Toronto Metropolitan University)
Long Term Interval Forecasts of Demand using Data-Driven Dynamic Regression Models / Prévisions à intervalles à long terme de la demande à l'aide de modèles de régression dynamiques axés sur les données

14:30-15:00 **Ruppa K Thulasiram** (University of Manitoba) **Japjeet Singh** (University of Manitoba) **Sulalitha Bowala** (University of Manitoba) **Aerambamoorhty A. Thavaneswaran** (University of Manitoba) **Saumen Mandal** (University of Manitoba)
Hybrid Data-Driven Fuzzy Risk Forecasts for Cryptocurrencies / Prévisions de risque floues reposant sur des données hybrides pour les cryptomonnaies

13:30-15:00 **Contributed / Communications libres** (abstract/résumé 296)

New Stochastic Processes and Their Applications
Nouveaux processus stochastiques et leurs applications

Chair/Président: Yanbo Tang

13:30-13:45 **Mufan Li** (University of Toronto) **Sinho Chewi** (Massachusetts Institute of Technology) **Murat A. Erdogdu** (University of Toronto) **Ruoqi Shen** (University of Washington) **Matthew Zhang** (University of Toronto)
Analysis of Langevin Monte Carlo from Poincaré to Log-Sobolev / Analyse de l'algorithme Monte-Carlo de Langevin, de l'inégalité de Poincaré et de l'inégalité de Sobolev logarithmique

13:45-14:00 **Golara Zafari** (Simon Fraser University) **Jean-François Bégin** (Simon Fraser University)
Parametric Inference of Multifactor Stochastic Volatility Models with Variance-Dependent Pricing Kernel / Inférence paramétrique des modèles de volatilité stochastique multifactorielle avec noyau de tarification dépendant de la variance

14:00-14:15 **Roberto Casarin** (Ca' Foscari University of Venice) **Mauro Costantini** (University of Aquila, Italy) **Anthony Osuntuyi** (Ca' Foscari University of Venice)
Bayesian nonparametric panel Markov-switching GARCH models / Modèles GARCH bayésiens non paramétriques à changements de régimes markovien

14:15-14:30 **Mohsen Bahremani** (Wilfrid Laurier University) **Xu (Sunny) Wang** (Wilfrid Laurier University)
Modeling Multivariate Hopfield-Transformer Hawkes Process: Application to Sovereign Credit Default Swaps / Modélisation de processus transformateur-Hopfield multivarié de Hawkes : application au contrat d'échange sur défaillance de crédit souverain

14:30-14:45 **Adam B. Kashlak** (University of Alberta) **Giseon Heo** (University of Alberta) **Prachi Loliencar** (University of Alberta)
Topological Hidden Markov Models / Modèles de Markov cachés topologiques

14:45-15:00 **Giulia Carallo** (Università Ca' Foscari di Venezia) **Roberto Casarin** (Ca' Foscari University of Venice) **Christian P. Robert** (Université Paris-Dauphine)
Generalized Poisson Difference Autoregressive Processes / Processus autorégressifs de la différence du modèle généralisé de Poisson

---

13:30-15:00 **Contributed / Communications libres** (abstract/résumé 299)

Statistical Methods for Health Sciences, Extreme Risk, and Extremal Dependence
Méthodes statistiques pour les sciences de la santé, les risques extrêmes et la dépendance extrême

Chair/Président: Candemir Cigsar

13:30-13:45 **James A. Hanley** (McGill University) **Maryse Kochoedo** (McGill University) **Rajib Dey** (McGill University) **Wilber Deck** (Direction de Santé Publique, Gaspé)
Measuring the Numbers of Lung Cancer (LC) Deaths Averted by Screening / Mesure du nombre de décès dus au cancer du poumon évités par le dépistage

13:45-14:00 **Jennifer McNichol** (University of Victoria) **Connie Stewart** (University of New Brunswick Saint John)
Simultaneous Maximum Unified Fatty Acid Signature Analysis / Analyse simultanée de la signature maximale unifiée des acides gras

14:00-14:15 **Xiaoqing Zhang** (University of Regina) **Dianliang Deng** (University of Regina)
Lindley Binomial Model: A Flexible Approach for Modelling the Proportions with Sparseness and Excessive zeros / Modèle binomial de Lindley : une approche souple pour la modélisation des proportions lors de dispersion et de surreprésentation des zéros

14:15-14:30 **Philip J. Schmidt** (University of Waterloo) **Ellen Cameron** (University of Waterloo) **Kirsten Muller** (University of Waterloo) **Monica Emelko** (University of Waterloo)
Amplicon Sequencing Diversity Analysis: Multinomial Models and Variants You Don't Know You Didn't See / Analyse de la diversité du séquençage d'amplicons : modèles multinomiaux et variantes que vous ne savez pas que vous n'avez pas vues

14:30-14:45 **Jonathan Jalbert** (Polytechnique Montreal) **Gamet Philémon** (Polytechnique Montréal)
A flexible extended generalized Pareto distribution for tail estimation / Loi de Pareto généralisée étendue pour la modélisation des valeurs extrêmes

14:45-15:00 **Michaël Lalancette** (University of Toronto) **Sebastian Engelke** (Université de Genève) **Stanislav Volgushev** (University of Toronto)
Inference for Extremal Graphical Models / Inférence pour les modèles graphiques extrémaux

15:30-17:00 **Invited / Sur invitation** (abstract/résumé 303)

Methodological Advances in Classification Models for Complex Longitudinal Data
Avancées méthodologiques des modèles de classification pour données longitudinales complexes

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

15:30-15:52 **Anuradha Roy** (The University of Texas at San Antonio) **Ricardo Leiva** (Universidad Nacional de Cuyo, Argentina)
Linear discrimination for three-level multivariate data / Discrimination linéaire des données multivariées à trois niveaux Ⓔ E

15:52-16:14 **Anita Brobbey** (University of Calgary) **Lisa M. Lix** (University of Manitoba) **Alberto Nettel-Aguirre** (University of Wollongong) **Tyler Williamson** (University of Calgary) **Samuel Wiebe** (University of Calgary) **Tolu Sajobi** (University of Calgary)
Repeated Measures Discriminant Analysis using Generalized Estimating Equations / Analyse discriminante des mesures répétées utilisant des équations d'estimation généralisées Ⓔ E

16:14-16:36 **David Hughes** (University of Liverpool)
Dynamic Longitudinal Discriminant Analysis Using Multiple Longitudinal Markers of Different Types / Analyse discriminante longitudinale et dynamique au moyen de marqueurs longitudinaux multiples de différents types Ⓔ E

16:36-16:58 **Jeffrey L. Andrews** (University of British Columbia, Okanagan) **Ryan P. Browne** (University of Waterloo) **Liam Welsh** (University of Toronto)
Finite mixture models for longitudinal data with dynamic group membership / Modèles de mélanges finis pour les données longitudinales présentant une appartenance dynamique à un groupe Ⓔ E

15:30-17:00 **Invited / Sur invitation** (abstract/résumé 306)

Fairness in Data-driven Research
Recherche fondée sur les données et équité

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

15:30-16:00 **Veronique Tremblay** (Beneva / HEC Montréal)
Responsible use of algorithms in decision making: ethical principles and recommendations / Utilisation responsable des modèles dans la prise de décision : principes éthiques et recommandations ⒺⒻ E F

16:00-16:30 **David R. Hunter** (Pennsylvania State University)
Gratz v. Bollinger and Statistical Machine Learning / Gratz contre Bollinger et l'apprentissage automatique statistique Ⓔ E

16:30-17:00 **Warut Khern-am-nuai** (McGill University)
Addressing Fairness in Machine Learning Predictions: Strategic Best-Response Fair Discriminant Removed Algorithm / Aborder la justesse dans les prédictions d'apprentissage automatique : Algorithme stratégique de meilleure réponse juste discriminant éliminé Ⓔ E

15:30-17:00 **Invited / Sur invitation** (abstract/résumé 308)

Nonresponse Issues in Surveys
Problèmes de non-réponse dans les enquêtes

Chair/Président: Francois Brisebois
Organizer/Responsable: Francois Brisebois
Sponsor/Commanditaires: Survey Methods Section / Le Groupe des méthodes d'enquête

15:30-16:00 **Brady West** (University of Michigan)
New measures for assessing non-ignorable selection bias in non-probability samples and low response rate probability samples / Nouvelles mesures pour évaluer le biais de sélection important des échantillons non probabilistes et probabilistes à faible taux de réponse

16:00-16:30 **Yajuan Si** (University of Michigan)
A Case Study of Nonresponse Bias Analysis in Educational Assessment Surveys / Étude de cas de l'analyse du biais de non-réponse dans les enquêtes d'évaluation de l'éducation

16:30-17:00 **Peter G. Wright** (Statistics Canada) **Patrice Martineau** (Statistics Canada) **François Brisebois** (Statistics Canada)
Improving Response by Studying Citizen Participation in Social Surveys / Amélioration de la réponse en examinant la participation citoyenne aux enquêtes sociales

---

15:30-17:00 **Invited / Sur invitation** (abstract/résumé 310)

Maintaining Relevancy Through New Tools, Data Science and Data Visualizations
Maintien de la pertinence grâce à de nouveaux outils, à la science des données et aux visualisations de données.

Chair/Président: Beatrice D. Baribeau
Organizer/Responsable: Beatrice D. Baribeau
Sponsor/Commanditaires: Accreditation Committee / Comité d'accréditation

15:30-16:00 **Peter Solymos** (E Source) **Khalid Lemzouji** (Analythium Solutions)
Best Practices for Delivering Applied Statistics from Concept to Production / Les meilleures pratiques de livraison de statistiques appliquées, de la conception à la production

16:00-16:30 **Kenneth C.K. Chu** (Statistics Canada)
Spaceborne Radar Earth Observation (Big) Data, Emerging Opportunities for Statisticians and Data Scientists / (Méga)données d'observation terrestre par radar spatioporté, occasions émergentes pour les statisticiens et les scientifiques des données

16:30-17:00 **Martin Monkman** (BC Stats, Province of British Columbia)
Continuous Learning in Times of Continuous Change / L'apprentissage continu en période de changement permanent

---

15:30-17:00 **Contributed / Communications libres** (abstract/résumé 312)

Spatial Data Analysis
Analyse de données spatiales

Chair/Président: Kathryn Morrison

15:30-15:45 **Jeffrey W. Peitsch** (University of Winnipeg)
Classification-Based Inference for Spatially Stratified Infectious Disease Systems / Inférence basée sur la classification pour systèmes de maladies infectieuses spatialement stratifiées

15:45-16:00 **Rick E Danielson** (Fisheries and Oceans Canada) **Hui Shen** (Pêches et Océans Canada) **Jing Tao** (Pêches et Océans Canada) **Will Perrie** (Pêches et Océans Canada)
Towards a Characterization of North Atlantic Right Whale Habitat from Space: Dependence of Ocean Current Features on Wind / Vers une caractérisation de l'habitat des baleines noires de l'Atlantique Nord depuis l'espace : dépendance des caractéristiques des courants océaniques par rapport au vent

16:00-16:15 **Madeline Ward** (University of Calgary) **Lorna E. Deeth** (University of Guelph) **Rob Deardon** (University of Calgary)
Incorporating Behavioural Change into Spatial Individual-Level Models for Infectious Disease Transmission / Incorporer le changement de comportement dans les modèles spatiaux de transmission de maladies infectieuses au niveau individuel

16:15-16:30 **Selvakkadunko Selvaratnam** (University of Toronto)
Applications of Robust Methods in Modern Spatial Analysis / Applications de méthodes robustes en analyses spatiales modernes

16:30-16:45 **Kyran Cupido** (St Francis Xavier University) **Petar Jevtic** (Arizona State University) **Tim Boonen** (University of Amsterdam)
Space, Mortality, and Economic Growth / Espace, mortalité et croissance économique

16:45-17:00 **Sara Zapata-Marin** (McGill University) **Alexandra M. Schmidt** (McGill University) **Scott Weichenthal** (McGill University) **Eric Lavigne** (Health Canada)
Modelling Temporally Misaligned Data Across Space / Modélisation de données temporellement désalignées dans l'espace

---

15:30-17:00 **Contributed / Communications libres** (abstract/résumé 316)

Nonparametric and Semiparametric Methods
Méthodes non paramétriques et semi-paramétriques

Chair/Président: Meng Yuan

15:30-15:45 **Yanglei Song** (Queen's University) **Meng Zhou** (Queen's University)
Truncated LinUCB for Stochastic Linear Bandits / Algorithme LinUCB tronqué pour des bandits linéaires stochastiques

15:45-16:00 **Archer Gong Zhang** (University of British Columbia) **Jiahua Chen** (University of British Columbia)
Estimation Efficiency under a Two-Sample Density Ratio Model / Efficacité de l'estimation sous un modèle de rapport de densité à deux échantillons

16:00-16:15 **Jervis Gallanosa** (University of Manitoba) **Yuliya V. Martsynyuk** (University of Manitoba)
Nonparametric Asymptotic Tests for Change in the Mean with Better Balanced Power Functions / Tests asymptotiques non paramétriques de changements de la moyenne avec fonctions de puissance équilibrée supérieure

16:15-16:30 **Marc Angelo Parsons** (McGill University) **Jingjun (Victor) Chen** (McGill University) **Andrea Benedetti** (McGill University)
Modelling Non-linear Exposure-outcome Relationships in Quantitative Systematic Reviews: A Meta-epidemiological Review of Current Practice / Modélisation des liens exposition-effet non-linéaires dans les revues systématiques quantitatives : une revue méta-épidémiologique de la pratique courante

16:30-16:45 **Xiaoting Li** (The University of British Columbia) **Harry Joe** (University of British Columbia)
Nonparametric Estimation of Multivariate Tail Probabilities / Estimation non paramétrique des probabilités de queue multivariées

16:45-17:00 **Deli Li** (Lakehead University) **Yu Miao** (Henan Normal University, China) **George Stoica** (University of New Brunswick, Canada)
A General Large Deviation Result for Partial Sums of Super-Heavy Tailed Random Variables / Résultat général de grand écart pour les sommes partielles de variables aléatoires à queue super lourde

---

15:30-17:00 **Contributed / Communications libres** (abstract/résumé 320)

New Sampling Techniques and High-dimensional Data Analysis
Nouvelles techniques d'échantillonnage et analyse des données à haute dimension

Chair/Président: Yixiu Liu

15:30-15:45 **Johanna de Haan-Ward** (University of Western Ontario) **Simon Bonner** (University of Western Ontario) **Douglas G. Woolford** (University of Western Ontario)
Comparison of Subsampling Methods for Prediction of Rare Events, with Application to Human-Caused Wildland Fire Prediction / Comparaison de méthodes de sous-échantillonnage pour la prédiction d'événements rares, avec application à la prédiction des incendies de forêt d'origine humaine

15:45-16:00 **Lorenzo Frattarolo** (European Commission Joint Research Centre) **Roberto Casarin** (University Ca' Foscari of Venice) **Radu V. Craiu** (University of Toronto) **Christian P. Robert** (CEREMADE, University Paris-Dauphine PSL and University of Warwick)
Living on the Edge: An Unified Approach to Antithetic Sampling / Vivre à la limite : une approche unifiée de l'échantillonnage antithétique

16:00-16:15 **Xiaotong Liu Zihang Lu** (Queen's University) **Myrtha Reyna** (The Hospital for Sick Children)
Defining Lifestyle Patterns Using High Dimensional Questionnaire Data / Définition des modes de vie à l'aide de données de questionnaire à haute dimension

16:15-16:30 **Derek Latremouille** (University of Toronto) **Dehan Kong** (University of Toronto) **Linglong Kong** (University of Alberta)
High-Dimensional, Low-Sample Tests of Normality Based on Concentration / Tests de normalité en haute dimension avec petite taille d'échantillon basés sur la concentration

16:30-16:45 **Richard Le Blanc** (CHUS)
Noncentral Distributions' Bayesian Inference in terms of Orthogonal Polynomials / Inférence bayésienne concernant les distributions noncentrales en termes de polynômes orthogonaux

16:45-17:00 **Jiarui Zhang** (Simon Fraser University)
An Annealed Sequential Monte Carlo Method for Generalized Bayesian Multidimensional Scaling / Méthode de Monte-Carlo séquentielle recuite pour l'échelonnement multidimensionnel bayésien généralisé

Chair/Président: Fernando Camacho

Organizer/Responsable: Fernando Camacho

Date: Sunday May 29 / dimanche 29 mai

Time/Heure: 11:00-18:00

Abstract/Résumé

[11:00-18:00]

Peter Solymos (E Source) **Khalid Lemzouji** (Worley)

Delivering applied statistics from concept to production

Fournir des statistiques appliquées, du concept à la production

Modern applied statistics involve communicating the results to various audiences. This communication increasingly takes place in interactive media rather than status reports. Traditional education for statisticians does not adequately prepare applied scientists for effectively handling such requirements. However, healthy exposure to software engineering skills and practices can greatly facilitate the timely delivery of results. This is due to the shorter time to working prototypes, shorter feedback loops involving stakeholders, and easier communication with IT/engineering when it comes to scale and performance.

Our 1-day course will introduce the thought process of making modular and reusable software code. Such code lays the foundation for quickly building prototypes and interfaces. We will introduce cloud-native technologies, such as Docker, and will use the R statistical programming language. The R language supports building full-featured web applications using the Shiny framework and developing web interfaces. We will use free and open-source software with a focus on R. The workshop organizers will pre-configure cloud instances for the participants to use thus cutting down on preparation and installation time, and also removing setup-related issues.

Participants will be expected to be familiar with R but extensive knowledge is not required. We will use the RStudio integrated development environment for programming and for accessing servers. Participants will need their own laptop with a modern internet browser and access to the internet. AV equipment depends on the in-person/remote/hybrid nature of the workshop (projector, conferencing software for remote participants).

Les statistiques appliquées modernes impliquent de communiquer les résultats à divers publics. Cette communication se fait de plus en plus par le biais de médias interactifs plutôt que par des rapports périodiques. L'éducation traditionnelle des statisticiens ne prépare pas adéquatement les scientifiques appliqués à gérer efficacement de telles exigences. Toutefois, une saine exposition aux compétences et aux pratiques de l'ingénierie logicielle peut grandement faciliter la livraison des résultats en temps voulu. Cela s'explique par le délai plus court pour obtenir des prototypes fonctionnels, des boucles de rétroaction avec les parties prenantes plus courtes et une communication plus facile avec l'informatique/l'ingénierie en ce qui concerne l'échelle et les performances.

Notre cours d'une journée présente le processus de réflexion permettant de créer un code logiciel modulaire et réutilisable. Un tel code jette les bases d'une construction rapide de prototypes et d'interfaces. Nous introduirons des technologies natives en nuage, telles que Docker, et utiliserons le langage de programmation statistique R. Le langage R permet de créer des applications Web complètes à l'aide du cadre Shiny et de développer des interfaces Web. Nous utiliserons des logiciels libres et gratuits qui s'appuient sur R. Les organisateurs de l'atelier configureront à l'avance des instances en nuage que les participants pourront utiliser, ce qui réduira le temps de préparation et d'installation et supprimera les problèmes liés à la configuration.

Les participants devront être familiarisés avec R mais une connaissance approfondie n'est pas requise. Nous utiliserons l'environnement de développement intégré RStudio pour la programmation et pour l'accès aux serveurs. Les participants auront besoin de leur propre ordinateur portable avec un navigateur Internet moderne et un accès à Internet. L'équipement audiovisuel dépend de la nature en personne/à distance/hybride de l'atelier (projecteur, logiciel de conférence pour les participants à distance).

Accreditation Workshop
Atelier du Comité d'accréditation

Outline:
The 6-hour workshop will be structured into 4 blocks, each approximately 1.5 hours long.

Grandes lignes :
L'atelier de 6 heures sera structuré en 4 blocs, chacun d'une durée d'environ 1,5 heure.

- Présentations
- Transformation d'un concept en modèle de modélisation des données
- Coffee break
- Développement d'interfaces de l'application et prototypes d'application
- Lunch break
- Partage des résultats avec les parties prenantes
- Déploiement d'applications activées dans le nuage
- Coffee break
- Performance et échelle : découplage de la logique commerciale de la présentation

Chair/Président: Jean-Francois Plante

Organizer/Responsable: Jean-Francois Plante

Date: Sunday May 29 / dimanche 29 mai

Time/Heure: 13:00-16:00

Abstract/Résumé

[13:00-17:00]

Sarah Legendre Bilodeau (Videns Analytics) **Sébastien Duguay** (Videns Analytics)

Kubernetes, containers and the cloud: an overview of the tools and challenges to put models in production

Kubernetes, conteneurs et cloud : tour d'horizon des outils et défis pour mettre des modèles en production

This workshop aims to familiarize participants with the reality of putting machine learning models into production. Participants will have the chance to test the effectiveness of these tools and use them in their work as needed.

Outline:

- Introduction to the complexity of computer systems environements
- R&D vs. Production: Can you keep up?
- Containers and orchestration: why should you care?
- Cloud computing: Write, Plan, Apply, Destroy (Rinse & Repeat)

Cet atelier vise à familiariser les participants avec les défis de la mise en production de modèles d'apprentissage machine. Les participants auront la chance de tester l'efficacité de ces outils et de les utiliser dans leur travail au besoin.

Structure :

- Introduction à la complexité des environements informatiques
- R&D vs production : Pouvez-vous suivre la parade ?
- Conteneurs et orchestration : pourquoi devriez-vous y porter attention ?
- Infonuagique : écrire, planifier, appliquer, détruire (rincer et répéter)

Statistical Education Workshop
Atelier du Groupe d'éducation en statistique

Chair/Président: Bruce Dunham

Organizer/Responsable: Bruce Dunham

Date: Sunday May 29 / dimanche 29 mai

Time/Heure: 13:30-17:00

Abstract/Résumé

[13:30-17:00]

Tiffany A. Timbers (The University of British Columbia) **Wesley Burr** (Trent University)

Reproducibility Workshop

Atelier sur la reproductibilité

This workshop will equip participants with tools (Git/GitHub, R/Rstudio and 'renv'- an R dependency management tool) and best practices for implementing data analysis workflows that promote collaboration and reproducibility. These tools and workflows are integral to creating reproducible and transparent research - research where the same result can be reached given the same input, computational methods, and conditions, as well as one that has a history which records how and why decisions were made that shaped the analysis. These tools and principles have also transformed the teaching of statistics, facilitating the development of active and experiential learning.

Cet atelier dotera les participants d'outils (Git/GitHub, R/Rstudio et 'renv'- outil de gestion des dépendances de R) et de bonnes pratiques pour mettre en œuvre des flux de travail d'analyse de données qui favorisent la collaboration et la reproductibilité. Ces outils et ces flux de travail sont essentiels à la création d'une recherche reproductible et transparente - une recherche où le même résultat peut être obtenu avec les mêmes données, les mêmes méthodes de calcul et les mêmes conditions, ainsi qu'une recherche qui dispose d'un historique qui enregistre comment et pourquoi les décisions qui ont façonné l'analyse ont été prises. Ces outils et principes ont également transformé l'enseignement de la statistique, en facilitant le développement d'un apprentissage actif et expérientiel.

Chair/Président: Jean-François Beaumont

Organizer/Responsable: Jean-François Beaumont

Date: Saturday June 4 / samedi 4 juin

Time/Heure: 12:30-17:00

Abstract/Résumé

[12:30-17:00]

Changbao Wu (University of Waterloo)

From Sample Surveys to Missing Data and Causal Inference

Des enquêtes par sondage aux données manquantes et à l'inférence causale

We provide an introduction to analysis of complex probability survey samples with a major focus on weighting and calibration methods. We show that the techniques, combined with the empirical likelihood methods, can be used as general inferential tools for missing data analysis and causal inference with observational data. Some recent developments on doubly robust and multiply robust inference and methods for analyzing non-probability survey samples are discussed to illustrate the use of weighting and calibration for these seemingly different but intrinsically connected topics. The workshop is designed for senior undergraduate or graduate students in statistics, biostatistics or related fields and young researchers interested in the topics.

Nous proposons une introduction à l'analyse d'échantillons d'enquêtes probabilistes complexes, en mettant l'accent sur les méthodes de pondération et de calibration. Nous montrons que ces techniques, combinées aux méthodes de vraisemblance empirique, peuvent être utilisées comme outils inférentiels généraux pour l'analyse des données manquantes et l'inférence causale avec des données d'observation. Nous discutons de certains développements récents en matière d'inférence doublement robuste et multiplement robuste et de méthodes d'analyse des échantillons d'enquête non probabilistes pour illustrer l'utilisation de la pondération et de la calibration dans ces domaines apparemment différents mais intrinsèquement liés. L'atelier est conçu pour les étudiants de premier cycle ou diplômés en statistique, biostatistique ou domaines connexes et les jeunes chercheurs intéressés par ces sujets.

Chair/Président: Nathaniel Tyler Stevens

Organizer/Responsable: Nathaniel Tyler Stevens

Date: Saturday June 4 / samedi 4 juin

Time/Heure: 13:00-16:30

Abstract/Résumé

[13:00-16:30]

Rodolfo Lourenzutti (University of British Columbia) **Arman Seyed-Ahmadi** (University of British Columbia) **Diego Ardila** (Shopify)

Intro to Databases in Industry: Data Cleaning, Querying, and Modeling at Scale

Introduction aux bases de données en industrie : nettoyage des données, interrogation et modélisation à grande échelle

This workshop is intended to walk the participants through the journey of data from the "raw" to an "analysis-ready" state. Using R, we will explore the basic flow of data cleaning and organizing raw data such that the outcome is error-free, consistent and accurate. The participants will then be introduced to relational databases—the most widely used option for storing clean, well-structured data. We will explore how to interact with and efficiently retrieve data from relational databases using their well-known, powerful query language of SQL. Finally, we will show how to connect R to SQL databases for reading and writing purposes.

Cet atelier a pour but de guider les participants dans le parcours des données, de l'état « brut » à un état « prêt à l'analyse ». À l'aide de R, nous explorerons le flux de base du nettoyage des données et de l'organisation des données brutes de sorte que le résultat soit exempt d'erreurs, cohérent et précis. Les participants seront ensuite initiés aux bases de données relationnelles - option la plus largement utilisée pour stocker des données propres et bien structurées. Nous explorerons comment interagir avec les bases de données relationnelles et en extraire efficacement des données à l'aide du langage d'interrogation puissant et bien connu SQL. Enfin, nous montrerons comment connecter R aux bases de données SQL à des fins de lecture et d'écriture.

Chair/Président: Ting Kam Leonard Wong

Organizer/Responsable: Ting Kam Leonard Wong

Date: Sunday June 5 / dimanche 5 juin

Time/Heure: 11:00-17:00

Abstract/Résumé

[11:00-17:00]

Ting Kam Leonard Wong (University of Toronto) **Jun Zhang** (University of Michigan) **Paul Marriott** (University of Waterloo) **Guido Montufar** (University of California, Los Angeles) **Gabriel Khan** (Iowa State University) **Melvin Loek** (University of California, San Diego) **Tian Han** (Stevens Institute of Technology) **Wuchen Li** (University of South Carolina)

Information geometry and applications

Géométrie de l'information et applications

Information geometry studies the geometry of spaces of probability distribution, which are also known as statistical manifolds. It provides a unified mathematical framework to study objects such as entropy, KL-divergence, the Fisher-Rao metric, and exponential families, as well as their generalizations. It has been applied to statistics and machine learning among other fields, particularly due to its close connections with optimal transport and statistical physics.

La géométrie de l'information est un domaine d'étude des espaces de distributions de probabilité exploitant des objets que sont les variétés statistiques. Cette géométrie fournit un cadre mathématique unifié pour étudier des objets tels que l'entropie, la divergence Kullback-Leibler, la métrique Fisher-Rao et la famille exponentielle, ainsi que leurs généralisations. Parmi d'autres champs d'application, il a été appliqué à la statistique et l'apprentissage automatique notamment en raison de liens étroits avec le transport optimal et la physique statistique.

Schedule

11am-12pm: Tutorial on information geometry (Jun Zhang, University of Michigan)

12pm-1pm: Lunch break

1pm-1:30pm: Paul Marriott (University of Waterloo)

1:30pm-2pm: Guido Montufar (University of California, Los Angeles)

2pm-2:30pm: Gabriel Khan (Iowa State University)

2:30pm-2:45pm: Break

2:45pm-3:15pm: Melvin Leok (University of California, San Diego)

3:15pm-3:45pm: Tian Han (Stevens Institute of Technology)

3:45pm-4:15pm: Wuchen Li (University of South Carolina)

4:15pm-4:45pm: Leonard Wong (University of Toronto)

4:45pm-5pm: Discussion

Programme

11h-12h: Tutoriel sur la géométrie de l'information (Jun Zhang, University of Michigan)

12h-13h: Lunch-déjeuner

13h-13h30: Paul Marriott (University of Waterloo)

13h30-14h: Guido Montufar (University of California, Los Angeles)

14h-14h30: Gabriel Khan (Iowa State University)

14h30-14h45: Break-café

14h45-15h15: Melvin Leok (University of California, San Diego)

15h15-15h45: Tian Han (Stevens Institute of Technology)

15h45-16h15: Wuchen Li (University of South Carolina)

16h15-16h45: Leonard Wong (University of Toronto)

16h45-17h: Discussion

Chair/Président: Rob Deardon

Organizer/Responsable: Rob Deardon

Date: Sunday June 5 / dimanche 5 juin

Time/Heure: 12:00-15:30

Abstract/Résumé

[12:00-15:30]

Jessica Gronsbell (University of Toronto)

Electronic Health Records Phenotyping

Phénotypage des dossiers médicaux électroniques

The widespread adoption of electronic health records (EHRs) has resulted in an unprecedented opportunity to leverage routinely collected medical data for purposes beyond patient care and billing. Vast amounts of longitudinal, patient-level information that were once locked away in paper format are being tapped for epidemiological research, clinical decision making, disease surveillance, and real-world predictive modeling of disease risk factors. The first step in nearly every EHR-based application is phenotyping, the process of identifying the subset of patients among the hundreds of thousands in the database who have the disease, condition, or characteristic that qualify them for analysis. Although a ubiquitous aspect of EHR research, phenotyping is a time consuming and financially demanding task due to the amount of expert knowledge required to precisely translate a clinical condition into criteria that describe its manifestation in the EHR. In this workshop, I will introduce statistical learning methods designed to expedite the phenotyping process in order to improve the scalability of EHR research.

Topics Covered & Timetable

This will be a half-day workshop covering the following topics:

- EHR Data: What DSE: De quoi s'agit-il ? (20 min)
- Brief history of EHR phenotyping Bref historique du phénotypage des DSE (30 min)
- Break – 10 min
- Methods for supervised phenotyping Méthodes d'apprentissage supervisé pour le phénotypage (90 min)
  - Background on statistical learning Contexte de l'apprentissage statistique
  - Approach overview Approche générale
  - Processing unstructured data with clinical natural language processing Traitement des données non structurées avec le traitement du langage naturel clinique
  - Model selection methods Méthodes de sélection des caractéristiques
- Recent research in phenotyping Recherche actuelle en phénotypage (15 min)

L'adoption généralisée des dossiers de santé électroniques (DSE) a donné lieu à une opportunité sans précédent d'exploiter les données médicales collectées de manière routinière à des fins autres que les soins aux patients et la facturation. De vastes quantités d'informations longitudinales au niveau du patient, autrefois enfermées dans un format papier, sont exploitées pour la recherche épidémiologique, la prise de décisions cliniques, la surveillance des maladies et la modélisation prédictive des facteurs de risque de maladie dans le monde réel. La première étape de presque toutes les applications basées sur les DSE est le phénotypage, ce processus d'identification du sous-ensemble de patients parmi les centaines de milliers de patients de la base de données qui présentent la maladie, l'état ou la caractéristique qui les qualifie pour l'analyse. Bien qu'il s'agisse d'un aspect omniprésent de la recherche sur les DSE, le phénotypage est une tâche qui demande beaucoup de temps et d'argent, en raison de la quantité de connaissances spécialisées requises pour traduire précisément une condition clinique en critères qui décrivent sa manifestation dans le DSE. Dans cet atelier, je présenterai des méthodes d'apprentissage statistique conçues pour accélérer le processus de phénotypage et améliorer l'évolutivité de la recherche sur les DSE.

Sujets couverts et calendrier

Il s'agira d'un atelier d'une demi-journée couvrant les sujets suivants :

- Les DSE : De quoi s'agit-il ? (20 min)
- Bref historique du phénotypage des DSE (30 min)
- Pause – 10 min
- Méthodes d'apprentissage supervisé pour le phénotypage (90 min)
  - Contexte de l'apprentissage statistique
  - Approche générale
  - Traitement des données non structurées avec le traitement du langage naturel clinique
  - Méthodes de sélection des caractéristiques
- Recherche actuelle en phénotypage (15 min)

- Break – 10 mins
- Hands-on practice (40 mins)

Learning Objectives

- Understand the benefits and challenges of using EHR data for research
- Understand the challenges of EHR-type types DSE and basic approaches to address them
- Capacité à mettre en œuvre un algorithme de phénotypage en utilisant des méthodes d'apprentissage statistique

Objectifs d'apprentissage

- Comprendre les avantages et les défis de l'utilisation des données du DSE pour la recherch
- Comprendre les défis du EHR-type types DSE et des approches de base pour les relever
- Capacité à mettre en œuvre un algorithme de phénotypage en utilisant des méthodes d'apprentissage statistique

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

Chair/Président: Grace Y. Yi

Organizer/Responsable: Grace Y. Yi

Date: Monday May 30 / lundi 30 mai

Time/Heure: 11:00-12:30

Abstract/Résumé

[11:00-12:00]

Anthony Davison (École polytechnique fédérale de Lausanne)

How Long Could a Human Live?

Y a-t-il une durée maximale de longévité humaine ?

There is long-standing and widespread interest in understanding if there is any limit to the human lifespan. Apart from its intrinsic interest, changes in survival in old age have implications for the sustainability of social security systems. Recent analyses of data on the oldest human lifespans have led to competing claims about survival and to some controversy, due in part to inappropriate use of statistical methods. One central question is whether the endpoint of the underlying lifetime distribution is finite. This talk will discuss the particularities associated with such data, outline correct ways of handling them and present suitable models and methods for their analysis. We illustrate the ideas through analysis of data on semi-supercentenarian lifetimes, which suggests that any upper limit to human lifetimes lies well beyond the highest lifetime yet reliably recorded, with lower limits to 95% confidence intervals around 130 years, and maximum likelihood estimates well above 130 years. The work is joint with Léo Belzile, Jutta Gampe, Holger Rootzén and Dmitrii Zholud.

La question de la longévité humaine exerce une fascination et trouve de nombreux échos dans la presse. Outre son intérêt intrinsèque, l'évolution de la survie des aînés a des implications pour la pérennité des systèmes de sécurité sociale. Des analyses récentes ont conduit à des affirmations contradictoires sur l'existence ou non d'une durée de vie maximale. Cette controverse est due en partie à une utilisation inappropriée des méthodes statistiques. En termes mathématiques, la question centrale est de savoir s'il y une limite finie à la distribution de la durée des vies. Cette présentation abordera les particularités associées à ces données, décrira les bonnes manières de les traiter et présentera des modèles et des méthodes appropriés pour leur analyse. Nous illustrons ces idées en étudiant les durées de vie des semi-supercentenaires ; ces dernières suggèrent que toute limite supérieure de la durée de vie humaine se situe bien au-delà de la durée de vie la plus élevée enregistrée, avec des limites inférieures des intervalles de confiance de 95% autour de 130 ans et des estimations de maximum de vraisemblance bien au-dessus de 130 ans. Le travail est conjoint avec Léo Belzile, Jutta Gampe, Holger Rootzén et Dmitrii Zholud.

Date: Monday May 30 / lundi 30 mai

Time/Heure: 12:30-13:30

Abstract/Résumé

[12:30-13:00]

Jizhou Tian (Lady Davis Institute for Medical Research, Jewish General Hospital) **Yi Liu** (Lady Davis Institute for Medical Research, Jewish General Hospital, Montreal, Quebec, Canada) **Andrea Benedetti** (McGill University)

An Empirical Comparison of the Two-Stage and One-Stage Bootstrap Approaches in the Context of an Individual Participant Data Meta-Analysis

Comparaison empirique d'approches bootstrap à deux étapes et à une étape dans un contexte de méta-analyse de données individuelles de participants

The bootstrap approach is generally adopted in constructing confidence intervals (CIs) to evaluate the accuracy of screening tools among subgroups in individual participant data meta-analyses (IPDMAs). Several approaches are commonly used including simple case bootstrap and two-stage bootstrap. However, there is hardly any direct comparison of the two-stage and one-stage bootstrap approaches in the context of IPDMA. In this study, we aimed to empirically compare the two bootstrap approaches using two different individual participant data. We generated 1000 bootstrap samples via the two approaches separately and fitted bivariate random-effects models. In IPDMA, the one-stage bootstrap was less time consuming and was able to obtain similar results to the two-stage bootstrap with regards to the mean difference in sensitivity and specificity estimates, although it had narrower CI width. A simulation study is needed to determine the best bootstrap approach in the context of IPDMA.

L'approche bootstrap est généralement adoptée pour la construction d'intervalles de confiance (CI) aux fins d'évaluer la précision des outils de filtrage parmi des sous-groupes dans les méta-analyses de données individuelles de participants (IPDMA). Bon nombre d'approches sont couramment utilisées, y compris le bootstrap à cas individuel ou à deux étapes. Cependant, il n'existe pratiquement aucune comparaison directe des approches bootstrap à deux étapes et à une étape dans un contexte d'IPDMA. Cette étude a pour but de comparer de façon empirique les deux approches bootstrap à l'aide de deux ensembles de données individuelles de participants. Nous avons généré séparément 1 000 échantillons bootstrap selon les deux approches et ajusté des modèles à effets aléatoires bivariés. Dans l'IPDMA, le bootstrap à une étape a pris moins de temps et a donné des résultats similaires au bootstrap à deux étapes quant à l'estimation de la différence de moyennes de sensibilité et de spécificité, mais la largeur des intervalles de confiance était plus étroite. Une étude de simulation est nécessaire pour déterminer la meilleure approche bootstrap dans un contexte d'IPDMA.

[12:30-13:00]

Xi Zhang (McMaster University) **Orla Murphy** (Dalhousie University) **Paul D. McNicholas** (McMaster University)

Longitudinal Data Clustering with a Copula Kernel Mixture Model

Regroupement de données longitudinales avec modèle de mélange à noyau de copules

Multivariate longitudinal data is composed of multiple highly autocorrelated time, therefore many commonly used clustering methods cannot be used for this data type. In this work, a copula kernel mixture model is proposed for clustering this type of data. By using copulas, each multivariate distribution component in the finite mixture model is decomposed into its marginal and dependence structure. For the marginal distributions, the Gaussian and gamma kernel functions are used. The multivariate Gaussian copula function is used as a de-

Les données longitudinales multivariées sont composées de plusieurs temps fortement autocorrélés, si bien que de nombreuses méthodes de regroupement couramment utilisées ne peuvent pas leur être appliquées. Dans ce travail, nous proposons un modèle de mélange à noyau de copules pour le regroupement de ce type de données. En utilisant les copules, on décompose chaque composante de distribution multivariée du modèle de mélange fini en structures marginales et de dépendance. Pour les distributions marginales, on utilise les fonctions de noyau gaussiennes et gamma. La fonction de copule gaussienne multivariée est utilisée comme

pendence structure due to its mathematical tractability. The expectation-maximization algorithm is used to estimate bandwidths and correlation matrices. A simulation study and real data are used to show the good performance of this method. Keywords: Multivariate longitudinal data clustering; copula kernel mixture model; expectation-maximization algorithm.

structure de dépendance en raison de sa maniabilité mathématique. On utilise l'algorithme d'espérance-maximisation pour estimer les largeurs de fenêtres et les matrices de corrélation. Nous proposons une étude de simulation et des données réelles pour montrer les bonnes performances de cette méthode. Mots-clés : Regroupement de données longitudinales multivariées ; modèle de mélange à noyau de copules ; algorithme d'espérance-maximisation.

[12:30-13:00]

Sidi Wu (Simon Fraser University) **Cédric Beaulac** (Simon Fraser University) **Jiguo Cao** (Simon Fraser University)

Neural Networks with Functional Response

Réseaux neuronaux avec réponse fonctionnelle

In recent years, there is a trend in applying machine learning techniques to known statistical fields. We are inspired to adapt the neural network (NN) architecture to functional data analysis. Most existing works concentrate on developing a NN taking functional data as inputs and outputting a scalar response, while our work looks at the other side of the coin. We design a feed-forward NN meant to predict a functional response using scalar inputs. The proposed method firstly transforms the functional response to a finite-dimensional vector of coefficients which act as the outputs of the NN, and then modifies the objective function of the NN to directly use the functional response for network training. We also apply a roughness penalty to control the smoothness over the predicted curves. In application, it is shown that our model outperforms the classic function-on-scalar regression model in both linear and nonlinear scenarios and owns superiority on computational cost with big data.

Ces dernières années, la tendance est d'appliquer les techniques d'apprentissage automatique à des domaines statistiques connus. Nous sommes inspirés d'adapter l'architecture des réseaux de neurones (RN) à l'analyse de données fonctionnelles. La plupart des travaux existants se concentrent sur le développement d'un RN avec des données fonctionnelles pour entrée et une réponse scalaire, alors que notre travail examine l'autre côté de la médaille. Nous concevons un RN feed-forward destiné à prédire une réponse fonctionnelle en utilisant des entrées scalaires. La méthode proposée transforme d'abord la réponse fonctionnelle en un vecteur à dimensions finies de coefficients qui agissent comme les sorties du RN, puis modifie la fonction objective du RN pour utiliser directement la réponse fonctionnelle pour l'entraînement du réseau. Nous appliquons également une pénalité de rugosité pour contrôler le lissage des courbes prédites. Dans l'application, il est démontré que notre modèle surpasse le modèle classique de régression de type fonction sur scalaire dans les scénarios linéaires et non linéaires et est moins couteux en calcul pour les données volumineuses.

[12:30-13:00]

Ruwan C. Karunanayaka (University of the Fraser Valley) **Boxin Tang** (Simon Fraser University)

On the Existence and Constructions of Orthogonal Designs

De l'existence et des constructions de plans orthogonaux

This project presents some new results on orthogonal designs, which are a useful class of designs for computer experiments. We first establish a non-existence result on orthogonal designs, generalizing an early result on orthogonal Latin hypercubes, and then present some construction results. We obtain a collection of orthogonal designs with small run sizes by computer search. Using these results and existing methods in the literature, we create a comprehensive catalogue of orthogonal designs for up to 100 runs.

Ce projet présente de nouveaux résultats sur les plans orthogonaux, classe de plans utile pour les expériences informatiques. Nous établissons d'abord un résultat de non-existence sur les plans orthogonaux, en généralisant un premier résultat sur les hypercubes latins orthogonaux, puis nous présentons quelques résultats de construction. Nous obtenons une collection de plans orthogonaux avec de petites tailles d'exécution par recherche informatique. En utilisant ces résultats et les méthodes existantes dans la littérature, nous créons un catalogue complet de plans orthogonaux pour un maximum de 100 exécutions.

[13:00-13:30]

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 d'importance séquentiel

A genealogy traces the ancestry of DNA sequences 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. However, a full genealogy can be so large that it greatly decreases the efficiency of existing sampling techniques. Partial genealogies trace the ancestry to a specified number of lineages, and dramatically improve the efficiency of some commonly-used sampling methods. 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 (SIS) and accommodates coalescence and mutation events in the ancestral history of the sequences. SIS methods are computationally intensive and have become popular as an alternative to MCMC methods for inference in population genetics.

Une généalogie retrace l'origine des séquences d'ADN dans le temps jusqu'à l'origine commune de l'ADN. Nous ne pouvons pas examiner les généalogies, mais les données des séquences d'ADN nous fournissent des informations qui permettent d'échantillonner la distribution a posteriori des généalogies sous-jacentes. Cependant, une généalogie complète peut être si vaste qu'elle réduit considérablement l'efficacité des techniques d'échantillonnage actuelles. Les généalogies partielles retracent les origines jusqu'à un nombre précis de lignées, et améliorent considérablement l'efficacité de certaines méthodes d'échantillonnage couramment utilisées. Nous présentons un algorithme d'échantillonnage des généalogies partielles d'un ensemble de séquences d'ADN à partir de leur distribution a posteriori. Notre algorithme fonctionne par échantillonnage d'importance séquentiel et tient compte des événements de coalescence et de mutation dans les origines des séquences. Les méthodes d'échantillonnage d'importance séquentiel sont intensives en matière de calcul et sont devenues très prisées pour remplacer les méthodes de Monte-Carlo par chaînes de Markov concernant l'inférence en génétique des populations.

[13:00-13:30]

Jingjun (Victor) Chen (McGill University) **Andrea Benedetti** (McGill University) **Zelalem F. Negeri** (McGill University; Lady Davis Institute for Medical Research, Jewish General Hospital) **Brett D. Thombs** (McGill University; Lady Davis Institute for Medical Research, Jewish General Hospital)

Individual Participant Data Meta-Analyses Using Bivariate Random Effect Models

Méta-analyses des données des participants individuels à l'aide de modèles à effets aléatoires bivariés

Bivariate random effects models (BREM) are commonly used to synthesize diagnostic test sensitivity and specificity from several primary studies either in the context of aggregate data meta-analysis (ADMA) or individual participant data meta-analysis (IPDMA). In the context of IPDMA, the examination of the BREM and its simplified versions are less than that of ADMA. For ADMA, most issues arise as a result of convergence issues introduced by sparse data, model complexity, default optimization, etc. The objective of this research is to empirically evaluate the BREM and its simplified versions in the context of IPDMA using the Patient Health Questionnaire-9 (PHQ-9) data accrued by the DEPRESsion Screening Data (DEPRESSD) Project. We compare the models based on their (a) pooled sensitivity and specificity estimates, (b) confidence interval widths, (c) shape and area of the prediction region, and (d) convergence properties. Results show that estimated specificity and sensitivity at each cut-off are compara-

Les modèles à effets aléatoires bivariés (BREM) sont couramment utilisés pour synthétiser la sensibilité et la spécificité des tests de diagnostic à partir de plusieurs études primaires, soit dans le contexte d'une méta-analyse de données agrégées (ADMA) ou d'une méta-analyse de données individuelles sur les participants (IPDMA). Dans le cadre de l'IPDMA, l'examen du BREM et de ses versions simplifiées est inférieur à celui de l'ADMA. Pour l'ADMA, la plupart des problèmes surviennent à la suite de problèmes de convergence introduits par des données rares, la complexité du modèle, l'optimisation par défaut, etc. L'objectif de cette recherche est d'évaluer empiriquement le BREM et ses versions simplifiées dans le contexte de l'IPDMA à l'aide du Patient Health Questionnaire -9 (PHQ-9) données accumulées par le projet DEPRESsion Screening Data (DEPRESSD). Nous comparons les modèles en fonction de leurs (a) estimations de sensibilité et de spécificité regroupées, (b) largeurs d'intervalle de confiance, (c) forme et surface de la région de prédiction et (d) propriétés de convergence. Nos résultats empiriques suggèrent que lorsqu'un nombre modéré d'études sont incluses, le BREM et ses simplifi-

ble across models. However, confidence interval width and the shape and area of the prediction region vary markedly across models. Our empirical findings suggest that when a moderate number of studies are included, the BREM and its simplifications do not have a drastic distinction in terms of diagnostic accuracy estimates but do exhibit some differences in other statistical aspects.

cations n'ont pas de distinction drastique en termes d'estimations de la précision du diagnostic, mais présentent certaines différences dans d'autres aspects statistiques.

[13:00-13:30]

Timofei Biziaev (University of Calgary)

Comparison of Frequentist and Bayesian Approaches to Ordinal Regression Model Validation

Comparaison des approches fréquentiste et bayésienne pour la validation de modèle de régression ordinal

Validation of probabilistic models is essential if the models are to be used clinically or be generalized to other populations. From a frequentist perspective this is generally comprised of assessing the model's ability to separate outcomes and provide accurate predictions of risk for new data. Validation through a Bayesian lens was considered as: (1) parameter estimates from fitting one data set are used as an informative prior for fitting a similar model with the other data set to assess model validity and (2) prior-data conflict evaluations that will be extended to ordinal response regression models. Data from the Alberta Tomorrow Project (N=2112) were fit in two sex-specific cancer-stage prediction models (stages I, II, III, or IV) and the models were validated using British Columbia Generations Project data (N=855). Predictors included lifestyle and family history information. Insights from applying both approaches are discussed as well as future research directions.

La validation de modèles probabilistes est essentielle si les modèles doivent être employés cliniquement ou généralisés à d'autres populations. Selon la perspective fréquentiste, il faut évaluer la capacité du modèle à séparer les résultats et à prédire le risque de façon fiable pour de nouvelles données. L'approche bayésienne fonctionne comme suit : (1) on se sert des estimations de paramètre tiré de l'ajustement d'un jeu de données à titre d'a priori informatif pour ajuster un modèle similaire avec l'autre jeu de données afin d'évaluer la validité du modèle, puis (2) évaluer le conflit des données a priori qui sera étendu aux modèles de régression de réponse ordinale. Les données provenant du Alberta Tomorrow Project (N=2112) ont été ajustées en deux modèles de prédiction du stade du cancer spécifique au sexe (stades 1, 2, 3 ou 4) et les modèles furent validés à l'aide des données du British Columbia Generations Project (N=855). Les prédicteurs comprennent de l'information relative au mode de vie et aux antécédents familiaux. Nous aborderons la rétroaction reçue en référence à l'application des deux approches et discuterons de nos orientations de recherche à venir.

[13:00-13:30]

Mengjie Bian (McMaster University) **Angelo J. Canty** (McMaster University)

Identification of Invalid Genetic Variants in Mendelian Randomization

Identification de variants génétiques invalides dans une randomisation mendélienne

Mendelian randomization (MR) uses genetic variants to investigate the causal relationship between exposure and outcome. The method relies on the assumption that the genetic variants are independent of outcome conditional on the exposure. If this assumption is not satisfied, the causal estimates may be biased. We propose a simple method to detect invalid variants using hypothesis testing in linear regression conditional on the exposure. Our method uses part of the data to detect invalid variants and rest for MR estimation. The first step needs individual-level data. The selection step could also use Bayesian variable selection or Lasso. Simulations show

La randomisation mendélienne (MR) utilise des variants génétiques pour étudier la relation causale entre exposition et issue. La méthode se fonde sur l'hypothèse que les variants génétiques sont indépendants de l'issue conditionnellement à l'exposition. Si la méthode ne satisfait pas cette hypothèse, les estimations de la causalité peuvent être biaisées. Nous proposons une méthode simple pour détecter les variants invalides par un test d'hypothèse dans une régression linéaire conditionnelle à l'exposition. Notre méthode utilise une partie des données pour la détection des variants invalides et l'autre pour l'estimation de la randomisation mendélienne. La première étape requiert des données individuelles. L'étape de sélection peut aussi utiliser une sélection de

our method has greater power to detect invalid variants than existing methods such as MR-Lasso and MR-PRESSO. All methods give unbiased estimators of the causal effect, but Wald intervals have low coverage due to selection effects. We propose using the bootstrap to deal with this issue.

variables bayésienne ou le Lasso. Des simulations montrent que la puissance de notre méthode pour détecter des variants invalides est plus élevée que celle des méthodes existantes comme la MR-Lasso et MR-PRESSO. Toutes les méthodes donnent des estimateurs non biaisés de l'effet causal, mais les intervalles de Wald ont une faible couverture en raison des effets de la sélection. Nous proposons l'utilisation d'un bootstrap pour traiter ce problème.

New Methods for Structured Variable Selection
Nouvelles méthodes de sélection structurée de variables

Chair/Président: Mireille Schnitzer

Organizer/Responsable: Guanbo Wang

Date: Monday May 30 / lundi 30 mai

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-13:52]

Marie Denis (Centre de coopération internationale en recherche agronomique pour le développement) **Mahlet G. Tadesse** (Georgetown University)

Graph-Structured Variable Selection with Gaussian Markov Random Field Horseshoe Prior

Sélection de variables structurées en graphe avec un a priori en fer à cheval de champs aléatoires gaussiens de Markov

A graph structure is commonly used to characterize the dependence between variables. The Bayesian approach provides a natural framework to integrate dependence structure through the priors. In this talk we present an approach that combines Gaussian Markov random field (MRF) prior with global–local (GL) shrinkage prior for selection of graph-structured variables. The local shrinkage parameters capture the dependence between connected covariates and take into account the sign of their empirical correlations. This encourages a similar amount of shrinkage for the regression coefficients while allowing them to have opposite signs. For non-connected variables, a standard horseshoe prior is specified. We illustrate the performance of the model with simulated data and real data applications, one in quantitative trait loci mapping with dependence between adjacent genetic markers and the other in gene expression data with a general estimated dependence structure between genes.

Une structure de graphe est couramment utilisée pour caractériser la dépendance entre des variables. L'approche bayésienne fournit un cadre naturel pour intégrer la structure de dépendance par le biais des a priori. Nous présentons une approche qui combine l'a priori des champs aléatoires gaussiens de Markov (MRF) et un a priori de rétrécissement global-local (GL) pour la sélection de variables structurées en graphe. Les paramètres de rétrécissement local saisissent la dépendance entre les covariables connectées et prennent en compte le signe de leurs corrélations empiriques. Cela favorise un degré similaire de rétrécissement pour les coefficients de régression, tout en leur permettant d'avoir des signes contraires. Pour les variables non connectées, un a priori en fer à cheval standard est spécifié. Nous illustrons la performance du modèle à l'aide de données simulées et d'applications avec données réelles, dont l'une avec mappage de locus de caractères quantitatifs avec dépendance entre des marqueurs génétiques adjacents et l'autre avec données sur l'expression génique avec une dépendance estimée générale entre les gènes.

[13:52-14:14]

Guanbo Wang (McGill University) **Mireille Schnitzer** (Université de Montréal) **Tom Chen** (Harvard Pilgrim Health Care Institute and Harvard Medical School) **Rui Wang** (Harvard Pilgrim Health Care Institute and Harvard Medical School) **Robert Platt** (McGill University)

A general framework for identification of permissible variable subsets in structured model selection

Cadre général d'identification des sous-ensembles de variables admissibles dans la sélection structurée de modèles

Oftentimes in variable selection, a selection rule that prescribes the permissible variable combinations in the final model is desirable due to the inherent structural constraints among the variables. Penalized regression methods can integrate these restrictions ("selection rules") by assigning the covariates to different groups.

Souvent, dans la sélection des variables, une règle de sélection, qui prescrit les combinaisons de variables admissibles dans le modèle final, est souhaitable en raison des contraintes structurelles inhérentes entre les variables. Les méthodes de régression pénalisée peuvent intégrer ces restrictions (appelées « règles de sélection ») par l'attribution des covariables à différents

However, no general framework has yet been proposed to formalize selection rules and their application. In this work, we develop a mathematical language for constructing selection rules in variable selection, where the resulting combination of permissible sets of selected covariates ("selection dictionary"), is formally defined. We show that all selection rules can be represented as a combination of operations on constructs, and these can be used to identify the related selection dictionary. We also present a condition for a grouping structure used with penalized regressions to carry out variable selection under an arbitrary selection rule.

groupes. Cependant, aucun cadre général n'a encore été proposé pour définir concrètement les règles de sélection et leur application. Dans le cadre de ces travaux, nous développons un langage mathématique pour créer des règles de sélection dans la sélection de variables, dans lequel la combinaison obtenue des ensembles admissibles de covariables sélectionnées (≪ dictionnaire de sélection ≫) est explicitement définie. Nous montrons que toutes les règles de sélection peuvent être représentées comme une combinaison d'opérations sur des constructions, et que celles-ci peuvent être utilisées pour identifier le dictionnaire de sélection correspondant. Nous présentons également une condition pour qu'une structure de regroupement utilisée avec des régressions pénalisées puisse effectuer une sélection de variables selon une règle de sélection arbitraire.

[14:14-14:36]

Yi Yang (McGill University) **Yuwen Gu** (University of Connecticut) **Yue Zhao** (University of York) **Jun Fan** (McGill University)

Flexible Regularized Estimating Equations: Some New Perspectives

Nouvelles perspectives d'équations d'estimation régularisées souples

We make some observations about the equivalences between regularized estimating equations, fixed-point problems and variational inequalities: (a) A regularized estimating equation is equivalent to a fixed-point problem, specified via the proximal operator of the corresponding penalty. (b) A regularized estimating equation is equivalent to a (generalized) variational inequality. Both equivalences extend to any estimating equations with convex penalty functions. To solve large-scale regularized estimating equations, it is worth pursuing computation by exploiting these connections. While fast computational algorithms are less developed for regularized estimating equation, there are many efficient solvers for fixed-point problems and variational inequalities. In this regard, we apply some efficient and scalable solvers which can deliver hundred-fold speed improvement. These connections can lead to further research in both computational and theoretical aspects of the regularized estimating equations.

Nous présentons quelques observations sur les équivalences entre les équations d'estimation régularisées, les problèmes de point fixe et les inégalités variationnelles : a) une équation d'estimation régularisée est équivalente à un problème en virgule fixe, spécifié par l'opérateur proximal de la pénalité correspondante ; b) une équation d'estimation régularisée est équivalente à une inégalité variationnelle (généralisée). Les deux équivalences sont applicables à toutes les équations d'estimation avec des fonctions de pénalité convexes. Pour résoudre les équations d'estimation régularisées à grande échelle, il est judicieux de poursuivre le calcul en exploitant ces connexions. Bien que les algorithmes de calcul rapide soient moins développés pour les équations d'estimation régularisées, il existe de nombreux résolveurs efficaces pour les problèmes de calcul en virgule fixe et les inégalités variationnelles. À cet égard, nous appliquons certains résolveurs efficaces et évolutifs qui peuvent améliorer la vitesse de cent fois. Ces liens peuvent engendrer des recherches plus poussées sur les aspects informatiques et théoriques des équations d'estimation régularisées.

[14:36-14:58]

Tingting Yu (Harvard Medical School and Harvard Pilgrim Health Care Institute)

Variable selection in high dimensional linear regression accounting for heterogeneity in covariate effects across multiple data sources

Sélection de variables dans la régression linéaire de haute dimension tenant compte de l'hétérogénéité des effets des covariables dans les sources de données multiples

When analyzing data combined from multiple sources, the heterogeneity across different sources must be ac-

Lors de l'analyse de données issues de sources multiples, il faut tenir compte de l'hétérogénéité entre les différentes sources. Nous

counted for. We consider high-dimensional linear regression models for integrative data analysis with heterogeneity across units modeled as unit-specific covariate effects. We propose a new adaptive clustering penalty (ACP) method, which penalizes distances from parameters to multiple cluster centers, to simultaneously select variables and cluster unit-specific covariate effects with sub-homogeneity. We show that the estimator based on the ACP method enjoys a strong oracle property under certain regularity conditions, and develop an efficient alternating direction method of multipliers (ADMM) algorithm for parameter estimation. We conduct simulation studies to compare the performance of the proposed method to existing methods and apply the method to real datasets.

examinons des modèles de régression linéaire de haute dimension pour l'analyse intégrative de données avec une hétérogénéité modélisée entre les unités, comme des effets de covariables spécifiques à l'unité. Nous proposons une nouvelle méthode de regroupement avec pénalité adaptative, qui pénalise les distances entre les paramètres et les centres de regroupement multiples, pour sélectionner simultanément les variables et regrouper les effets des covariables spécifiques aux unités présentant une sous-homogénéité. Nous montrons que l'estimateur reposant sur la méthode de regroupement avec pénalité adaptative possède une forte propriété d'oracle dans certaines conditions de régularité, puis nous développons un algorithme des directions alternées efficace pour l'estimation des paramètres. Nous réalisons des études de simulation pour comparer les résultats de la méthode proposée aux méthodes actuelles. Enfin, nous appliquons notre méthode à des ensembles de données réelles.

Statistical Disclosure Control Methods for Privacy
Méthodes de contrôle de la divulgation statistique et vie privée

Chair/Président: Linglong Kong

Organizer/Responsable: Bei Jiang

Date: Monday May 30 / lundi 30 mai

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-13:55]

Fang Liu (University of Notre Dame)

Utility Analysis of Differentially Private Gradient-based Optimization Algorithms

Analyse d'utilité d'algorithmes d'optimisation différentiellement confidentiels à base de gradients

Differentially private machine learning procedures output results with privacy guarantees. Noisy stochastic gradient descent (SGD) is an approach to achieving privacy guarantees by perturbing the gradients of a gradient-based optimizer. We provide theoretical analysis on the usefulness of noisy SGD with respect to the dimensionality of the optimization problem, the training data size, the batch size, and the privacy loss. We also derive a lower bound on the sample complexity for useful noisy SGD procedures. Our results shed light and provide general guidance on whether useful privacy-preserving results may be obtained via noisy SGD in a ML procedure.

Les résultats tirés de procédures d'apprentissage automatique différentiellement confidentielles sont accompagnées de garanties quant à la confidentialité. La descente de gradient stochastique (SGD) bruitée est une approche qui offre des garanties de confidentialité en perturbant les gradients d'un optimiseur à base de gradient. Nous présentons une analyse théorique sur l'utilité d'une SGD avec bruit en respectant la dimensionnalité du problème d'optimisation, la taille des données d'apprentissage, la taille du lot et la perte de confidentialité. Nous dérivons aussi une limite inférieure dans la complexité d'échantillon pour les procédures pratiques de la SGD avec bruit. Nos résultats soulignent et offrent une référence générale pour savoir si les résultats utiles préservant la confidentialité peuvent être obtenus au moyen d'une SGD dans une procédure ML.

[13:55-14:20]

Hui Xie (Simon Fraser University) **Yi Qian** (University of British Columbia)

Fast Distribution-free Statistical Control Methods to Construct Large-scale Privacy-preserving Databases

Méthodes statistique rapides et libre du contrôle de la distribution pour construire des bases de données à grande échelle préservant la confidentialité

We develop a class of fast distribution-free data perturbation, shuffling and multiple imputation synthetic data methods for building large-scale secure databases protecting confidential data while maintaining high utility. These distribution-free procedures are applicable to mask mixed continuous and discrete confidential variables with arbitrary unknown distributions. Furthermore, the methods nest generalized linear models commonly used for data analysis as special cases. Thus, a wide range of statistical relationships can be recovered from the secure databases with results similar to those achieved using original data. To overcome the computational bottle neck for continuous sensitive at-

Nous développons une classe de méthodes de données synthétiques de perturbation rapide, de brassage et d'imputation de données et sans distribution pour la construction de bases de données sécurisées à grande échelle protégeant les données confidentielles tout en maintenant une grande utilité. Ces procédures sans distribution peuvent servir à masquer les variables confidentielles mixtes discrètes et continues à l'aide de distributions d'inconnu arbitraire. De plus, les méthodes comprennent des modèles linéaires généralisés fréquemment employés pour les analyses de données dans des cas particuliers. Conséquemment, un large éventail de liens statistiques peut être récupéré à partir de bases de données protégées tout en obtenant des résultats similaires à ceux obtenus à partir de données d'origine. Afin de surmonter le goulot

tributes with many unique values, these procedures employ a profile-likelihood approach to eliminating the need to estimate the large number of nuisance parameters for baseline distributions, resulting in a large reduction (¿90%) of computational time. Thus the proposed methods are scalable for building robust large-scale secure databases. Simulation studies and empirical applications demonstrate that they outperform existing data-masking methods.

d'étranglement informatique pour les attributs sensibles continus ayant de nombreuses valeurs uniques, ces procédures adoptent une approche de vraisemblance profilée pour éviter d'avoir à estimer le grand nombre de paramètres de nuisance des distributions de base, ce qui réduit énormément le temps de calcul (plus de 90 %). Les méthodes proposées sont ainsi extensibles pour la construction de bases de données protégées, robustes et à grande échelle. Des études en simulation et des applications empiriques démontrent que nos méthodes de masquage de données surpassent les autres méthodes existantes.

[14:20-14:45]

Liu Yi (University of Alberta) **Ke Sun** (University of Alberta) **Bei Jiang** (University of Alberta) **Linglong Kong** (University of Alberta)

A Bridge to Gaussian Differential Privacy

Un pont vers la confidentialité différentielle gaussienne

Gaussian differential privacy (GDP) is a single-parameter family of privacy notions that provides coherent guarantees to avoid the exposure of sensitive individual information. Relative to DP, GDP provides more interpretability and tighter bounds under composition. Many widely used mechanisms (e.g., the Laplace mechanism) inherently provide GDP guarantees but often fail to take advantage of this new framework because their privacy guarantees were derived under a different background. We develop an easy-to-verify criterion to identify such algorithms and give an efficient method to narrow down possible values of an optimal privacy measurement, μ with an arbitrarily small and quantifiable margin of error.

La confidentialité différentielle gaussienne (GDP) est une famille à paramètre unique de notions de confidentialité qui fournit des garanties cohérentes pour éviter l'exposition de renseignements personnels sensibles. Comparativement à la confidentialité différentielle (DP), la GDP offre une plus grande interprétabilité et des bornes plus étroites sous composition. Bon nombre de mécanismes largement utilisés (par ex. : le mécanisme de Laplace) fournissent de façon inhérente des garanties GDP, mais échouent souvent à tirer parti de ce nouveau cadre parce que leurs garanties de confidentialité sont dérivées sous un arrière-plan différent. Nous développons un critère facilement vérifiable pour identifier de tels algorithmes et offrir une méthode efficace pour réduire le nombre de valeurs possibles d'une mesure de confidentialité optimale, μ avec une marge d'erreur arbitrairement petite et quantifiable.

Use of Machine Learning Methods for Handling Missing Survey Data
Utilisation de méthodes d'apprentissage automatique pour le traitement des données d'enquête manquantes

Chair/Président: Keven Bosa, Jean-François Beaumont

Organizer/Responsable: Keven Bosa

Date: Monday May 30 / lundi 30 mai

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-14:00]

Mehdi Dagdoug (Université de Bourgogne Franche-Comté) **Camelia Goga** (Université de Bourgogne Franche-Comté) **David Haziza** (University of Ottawa)

Model-Assisted Estimation with Machine Learning Methods in High-Dimensional Settings for Survey Data

Estimation assistée par modèle avec des méthodes d'apprentissage automatique dans des contextes de grande dimension pour les données d'enquête

For decades, methodologies attempting to use efficiently the auxiliary information available, such as model-assisted estimators, have attracted a lot of attention in the literature. Most of the asymptotic properties of these estimators have been established in a framework in which both the population and sample sizes diverge, with a fixed number of auxiliary variables. Yet, nowadays, survey practitioners face the emergence of data sets with a large number of covariates. Therefore, a more realistic framework would be one in which the number of auxiliary variables is allowed to diverge as well. In this work, we adopt this high-dimensional asymptotic framework and establish the convergence rates of model-assisted estimators built upon various statistical learning procedures. We also conducted a large simulation study that included many estimators based on popular machine learning algorithms, in various high-dimensional settings and with several sampling designs.

Depuis plusieurs décennies, les méthodes permettant l'utilisation d'informations auxiliaires, telles que les estimateurs par modélisation assistée, ont attiré beaucoup d'attention. La majorité des propriétés asymptotiques de ces estimateurs ont été obtenues dans un cadre où la taille de l'échantillon et celle de la population divergent toutes les deux, tout en considérant un nombre de covariables fixe. Cependant, de nos jours, les statisticiens ont fréquemment en leur possession des jeux de données incluant un très grand nombre de covariables. Il est par conséquent nécessaire d'étudier aussi le scénario dans lequel le nombre de covariables est autorisé à diverger. Dans ce travail, la convergence de plusieurs estimateurs est étudiée dans le cadre asymptotique en grande dimension précédemment mentionné. De plus, une étude par simulations incluant plusieurs estimateurs par modélisation assistée construits à partir de modèles provenant de l'apprentissage statistique est réalisée.

[14:00-14:30]

Sixia Chen (University of Oklahoma Health Sciences Center) **Chao Xu** (University of Oklahoma Health Sciences Center)

Handling High Dimensional Data with Missing Values by Modern Machine Learning Techniques

Traitement de données de haute dimension avec valeurs manquantes par des techniques modernes d'apprentissage automatique

High dimensional data has been regarded as one of the most important types of big data in practice. It happens frequently in practice including genetic study, financial study, and geographical study. Missing data in high dimensional data analysis should be handled properly to reduce nonresponse bias. We discuss some mod-

Les données de haute dimension sont reconnues comme étant les types de mégadonnées les plus importantes concrètement. On les rencontre fréquemment en pratique, par exemple dans les études génétiques, financières et géographiques. Les données manquantes dans les analyses de données de haute dimension doivent être traitées adéquatement afin de réduire le biais de non-réponse. Nous

ern machine learning techniques including penalized regression approaches, tree based approaches, and deep learning for handling missing data with high dimensionality. Specifically, our proposed methods can be used for estimating general parameters of interest including population means and percentiles with imputation based estimators, propensity score estimators, and doubly robust estimators. We compare those methods through some limited simulation studies and one real application. Both simulation studies and real application show the benefits of deep learning and XGboost approaches compared with other methods in terms of balancing bias and variance.

aborderons certaines techniques modernes d'apprentissage automatique comprenant les approches de régression pénalisées, les approches à base d'arbre et l'apprentissage profond dans le traitement de données manquantes dans un jeu de haute dimension. Plus précisément, la méthode que nous proposons peut servir à estimer les paramètres généraux avantageux comme les moyennes et pourcentages de population avec des estimateurs à base d'imputation, avec score de propension, et doublement robuste. Nous comparons ces méthodes par l'entremise d'études en simulation et d'une application réelle. Les deux études en simulation et l'application réelle démontrent les avantages de l'apprentissage profond et de l'approche XGboost par rapport aux autres méthodes en termes d'équilibre du biais et de la variance.

[14:30-15:00]

Olanrewaju Michael Akande (Duke University) **Zhenhua Wang** (University of Missouri) **Jason Poulos** (Harvard Medical School) **Fan Li** (Duke University)

Are Deep Learning Models Superior for Missing Data Imputation in Surveys? Evidence from an Empirical Comparison

Les modèles d'apprentissage profond permettent-ils une meilleure imputation des données manquantes dans les enquêtes ? Résultats d'une comparaison empirique

Multiple imputation (MI) is a popular approach for dealing with missing data arising from non-response in sample surveys. Multiple imputation by chained equations (MICE) is one of the most widely used MI algorithms for multivariate data, but it lacks theoretical foundation and is computationally intensive. Recently, missing data imputation methods based on deep learning models have been developed with encouraging results in small studies. However, there has been limited research on evaluating their performance in realistic settings compared to MICE, particularly in large-scale surveys. We conduct extensive simulation studies based on American Community Survey data to compare the repeated sampling properties of four machine learning based MI methods: MICE with classification trees, MICE with random forests, generative adversarial imputation networks, and multiple imputation using denoising autoencoders. We find the deep learning based imputation methods is superior to MICE in terms of computational time. However, with the default choice of hyperparameters in the common software packages, MICE with classification trees consistently outperforms, often by a large margin, the deep learning imputation methods in terms of bias, mean squared error, and coverage under a range of realistic settings.

L'imputation multiple (IM) est une approche populaire pour traiter les données manquantes résultant de la non-réponse dans les enquêtes par sondage. L'imputation multiple par équations chaînées (MICE) est l'un des algorithmes d'IM les plus utilisés pour les données multivariées, mais elle manque de fondement théorique et est lourde à calculer. Récemment, des méthodes d'imputation des données manquantes basées sur des modèles d'apprentissage profond ont été développées avec des résultats encourageants dans de petites études. Cependant, peu de recherches ont été menées sur l'évaluation de leurs performances dans des contextes réalistes par rapport à MICE, notamment sur des enquêtes à grande échelle. Nous menons des études de simulation approfondies basées sur les données de l'American Community Survey pour comparer les propriétés d'échantillonnage répété de quatre méthodes d'IM basées sur l'apprentissage machine : MICE avec arbres de classification, MICE avec forêts aléatoires, réseaux d'imputation adversariale générative et imputation multiple à l'aide d'autoencodeurs de débruitage. Nous constatons que les méthodes d'imputation basées sur l'apprentissage profond sont supérieures à MICE en termes de temps de calcul. Cependant, avec le choix par défaut des hyperparamètres des progiciels courants, MICE avec arbres de classification surpasse systématiquement, et souvent de loin, les méthodes d'imputation par apprentissage profond en termes de biais, d'erreur quadratique moyenne et de couverture pour une gamme de paramètres réalistes.

Isobel Loutit Invited Address
Allocution Isobel-Loutit

Chair/Président: Hugh Chipman

Organizer/Responsable: Jean-François Plante

Date: Monday May 30 / lundi 30 mai

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-14:30]

Jeff Wu (Georgia Tech)

My Five Years in Canada: How it Impacted my Work and the Field of Experimental Design

Mes cinq années au Canada : leur impact sur mon travail et le domaine de la conception expérimentale

In 1988-93 I worked at the U of Waterloo as the GM/NSERC Chair in Quality and Productivity. In this talk I will explain how thisacademic/industrial interactions had helped me to develop new ideas and methods, primarily in design and analysis of experiments. I have done the following work: analysis of experiments with complex aliasing, effect heredity principle, various aspects of robust parameter design, theory of minimum aberration, orthogonal arrays of small size. The new method of Conditional Main Effects, which was fully developed in recent years, had its germination then. From these work, I started to envision a new system of experimental design, which serves as a blue print for the Wu-Hamada book "Experiments". Many people had contributed to the work in this period, some of which are still working in Canada. My own experience was transformational. Prior to 1988, I was working primarily in theory and methodology and had no idea about the real world problems and how they can inspire and impact innovative research. After 1993 I have become a more complete statistician.

De 1988 à 1993, j'ai travaillé à la University of Waterloo en tant que titulaire de la chaire GM/CRSNG en qualité et productivité. Dans cet exposé, j'expliquerai comment ces interactions entre le monde universitaire et l'industrie m'ont aidé à développer de nouvelles idées et méthodes, principalement dans le domaine de la conception et de l'analyse d'expériences. J'ai effectué les travaux suivants : analyse d'expériences avec aliasing complexe, principe d'hérédité des effets, divers aspects de la conception de paramètres robustes, théorie de l'aberration minimale, tableaux orthogonaux de petite taille. La nouvelle méthode des effets principaux conditionnels, qui a été pleinement développée ces dernières années, a eu sa germination à cette époque. À partir de ces travaux, j'ai commencé à envisager un nouveau système de plans d'expérience, qui a servi de schéma directeur pour le livre de Wu-Hamada « Experiments ». De nombreuses personnes ont contribué à mes travaux pendant cette période, dont certaines travaillent encore au Canada. Ma propre expérience a été transformatrice. Avant 1988, je travaillais principalement en théorie et en méthodologie et je n'avais aucune idée des problèmes du monde réel et de la façon dont ils peuvent inspirer et influencer la recherche innovante. Après 1993, je suis devenu un statisticien plus complet.

Chair/Président: Maciej Augustyniak

Organizer/Responsable: Maciej Augustyniak

Date: Monday May 30 / lundi 30 mai

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-14:00]

Alexandru Badescu (University of Calgary) **Maciej Augustyniak** (Université de Montréal) **Jean-François Bégin** (Simon Fraser University) **Sarath Kumar Jayaraman** (University of Calgary)

Long Memory in Option Pricing: A Fractional Discrete-Time Framework

Mémoire longue pour la tarification des options : un cadre fractionnaire en temps discret

This article studies the impact of long memory on modelling asset returns and pricing options in discrete-time. We propose a general pricing framework based on affine multi-component volatility models that admit ARCH(∞) representations, which not only nests a plethora of option pricing models from the literature, but also allows for the introduction of novel fractionally integrated processes for valuation purposes. We carry out an extensive empirical analysis which includes single and joint calibrations of a variety of short and long memory models to historical returns and S&P 500 options. Our results indicate that the inclusion of long memory into modelling the returns substantially improves the option pricing performance. Moreover, using an expanding window out-of-sample exercise, we show that a single-component long-memory model outperforms a richer-parametrized two-component model with short-memory dynamics, the difference becoming even larger when combining the two features.

Cet article met en lumière l'impact de la mémoire longue sur la modélisation en temps discret du rendement des actifs et de la tarification des options. Nous proposons un cadre général de tarification basé sur des modèles affines à volatilité multicomposantes qui admettent des représentations ARCH(∞). En plus d'inclure une foule de modèles de tarification des options disponibles dans la littérature, ce cadre permet aussi de présenter des nouveaux processus fractionnaires intégrés à des fins d'évaluation. Nous procédons à une vaste analyse empirique qui comprend des calibrages simples et conjoints de divers modèles à mémoire courte et longue de rendements historiques et d'options S&P 500. Nos résultats indiquent que l'inclusion de mémoire longue dans la modélisation des rendements améliore sensiblement la performance de la tarification des options. De plus, à l'aide d'un exercice avec fenêtre croissante hors échantillon, nous montrons qu'un modèle à mémoire longue avec composante unique surpasse un modèle à deux composantes à paramétrisation plus riche avec une dynamique à mémoire courte, la différence étant encore plus marquée lorsque les deux caractéristiques sont combinées.

[14:00-14:30]

Jean-François Bégin (Simon Fraser University)

On Complex Economic Scenario Generators: Is Less More?

Sur les générateurs de scénarios économiques complexes : Moins, c'est mieux ?

This article proposes a complex economic scenario generator that nests versions of well-known actuarial frameworks. The generator estimation relies on the Bayesian paradigm and accounts for both model and parameter uncertainty via Markov chain Monte Carlo methods. So, to the question is less more?, we answer maybe, but it depends on your criteria. From an in-sample fit perspective, on the one hand, a complex economic scenario gen-

Cet article propose un générateur de scénarios économiques complexe qui imbrique des versions de modèles actuariels bien connus. L'estimation du générateur repose sur le paradigme bayésien et tient compte de l'incertitude du modèle et des paramètres, le tout en utilisant des méthodes de Monte Carlo à chaîne de Markov. Ainsi, à la question est-ce que moins, c'est mieux ?, nous répondons peut-être, mais cela dépend de vos critères. Du point de vue de l'adéquation à l'échantillon, d'une part, un générateur de

erator seems better. From the conservatism, forecasting, and coverage perspectives, on the other hand, the situation is less clear: having more complex models for the short rate, term structure, and stock index returns is clearly beneficial. However, that is not the case for inflation and the dividend yield.

scénarios économiques complexe semble préférable. En revanche, du point de vue du conservatisme, de la prévision et de la couverture, la situation est moins claire : disposer de modèles plus complexes pour le taux court, la structure des taux et les rendements des indices boursiers est clairement bénéfique. En revanche, ce n'est pas le cas pour l'inflation et le rendement des dividendes.

[14:30-15:00]

Anne Mackay (Université de Sherbrooke) **Michael A. Kouritzin** (University of Alberta)

On stochatic approximation and option pricing

Tarification d'options via un algorithme d'approximation stochastique

We consider almost sure convergence rates of averaged linear stochastic approximation algorithms, when applied to data with triangular dependence structure and heavy tails. We find that when the data is replaced by its running average in the algorithm, convergence may be faster. We then obtain rates of convergence of price estimates in the context of American option pricing via a dynamic programming algorithm with stochastic approximation. From a methodological point of view, our results show that using averaged data in the pricing algorithm leads to speeds of convergence that are more robust to the choice of parameters.

On s'intéresse au taux de convergence presque sûre d'un algorithme d'approximation stochastique linéaire appliqué à des données présentant une structure de dépendance triangulaire et des queues lourdes. On démontre que lorsque les données entrantes dans l'algorithme sont remplacées par une moyenne mobile, la convergence peut être plus rapide. On présente également le taux de convergence des prix obtenus lorsque cet algorithme est utilisé pour tarifer des options américaines par programmation dynamique. D'un point de vue méthodologique, nos résultats démontrent que l'utilisation d'une moyenne mobile dans l'algorithme de tarification rend la vitesse de convergence plus robuste au choix des paramètres.

Stochastic Processes, Monte Carlo Integration, and AFT Model
Processus stochastiques, intégration Monte Carlo et modèle de temps de défaillance accéléré

Chair/Président: Adam B. Kashlak

Date: Monday May 30 / lundi 30 mai

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-13:45]

Mamadou Yamar Thioub (HEC Montréal) **Bouchra Nasri** (Université de Montréal) **Bruno Rémillard** (HEC Montréal)

Goodness-of-fit Tests and Robust Regime Selection Procedure for General Hidden Markov Models

Tests d'adéquation et procédure robuste de sélection de régimes pour les modèles de Markov cachés

In this work, we present powerful goodness-of-fit procedures for general Markov regime-switching models with covariates when the outcomes have continuous, discrete, or zero-inflated distributions. The EM algorithm is used for the estimation and a randomized Rosenblatt's transform is applied to obtain formal goodness-of-fit tests. The latter then served for selecting the number of regimes. Numerical experiments are used to assess the finite sample performance of the proposed methodologies and to compare with other criteria for selection of models, including Bayesian methods.

Dans ce travail, nous présentons de puissants tests d'adéquation pour les modèles généraux markovien à changement de régimes avec covariables, lorsque les réponses ont des distributions continues, discrètes ou à inflation de zéros. L'algorithme EM est utilisé pour l'estimation et une transformation de Rosenblatt randomisée est appliquée pour obtenir des tests formels d'adéquation. Ces derniers servent ensuite à sélectionner le nombre de régimes. Des expériences numériques sont réalisées pour évaluer la performance des méthodologies proposées et à les comparer avec d'autres critères de sélection de modèles, incluant les méthodes bayésiennes.

[13:45-14:00]

Sabrina Sixta (University of Toronto)

Convergence Rate Bounds for Iterative Random Functions Using One-Shot Coupling

Limites de taux de convergence pour les fonctions aléatoires itératives utilisant le couplage à un coup

One-shot coupling is a method of bounding the convergence rate between two copies of a Markov chain in total variation distance. The method is divided into two parts: the contraction phase, when the chains converge in expected distance and the coalescing phase, which occurs at the last iteration, when there is an attempt to couple. In this paper, we present a general theorem for finding the upper bound on the Markov chain convergence rate that uses the one-shot coupling method. Our theorem does not require the use of any exogenous variables like a drift function or minorization constant. We then apply the general theorem to two families of Markov chains: the random functional autoregressive process and the randomly scaled iterated random function. The one-shot coupling method appears to generate tight geometric convergence rate bounds.

Le couplage à un coup est une méthode permettant de limiter le taux de convergence entre deux copies d'une chaîne de Markov en distance de variation totale. La méthode est divisée en deux parties : la phase de contraction, lorsque les chaînes convergent en distance attendue et la phase de coalescence, qui se produit à la dernière itération, lorsqu'il y a une tentative de couplage. Dans cet article, nous présentons un théorème général pour trouver la limite supérieure du taux de convergence des chaînes de Markov qui utilise la méthode de couplage à un coup. Notre théorème ne nécessite pas l'utilisation de variables exogènes comme une fonction de dérive ou une constante de minorisation. Nous appliquons ensuite le théorème général à deux familles de chaînes de Markov : le processus autorégressif fonctionnel aléatoire et la fonction aléatoire itérée à échelle aléatoire. La méthode de couplage à un coup semble générer des limites de taux de convergence géométriques serrées.

[14:00-14:15]

Dinh-Toan Nguyen (Université du Québec à Montréal)

Scaling Limit of the Collision Measures of Multiple Random Walks

Limite d'échelle des mesures de collision de plusieurs marches aléatoires

For an integer $k \geq 2$, let $S^{(1)}, S^{(2)}, \ldots, S^{(k)}$ be k independent simple random walks in $\mathbb{Z}$. A pair (n,z) is called a collision event if there are at least two distinct random walks, namely, $S^{(i)}, S^{(j)}$ satisfying $S_n^{(i)} = S_n^{(j)} = z$. We show that under the same scaling as in Donsker's theorem, the sequence of random measures representing these collision events converges to a non-trivial random measure on $[0,1] \times \mathbb{R}$. Moreover, the limiting random measure can be characterized using Wiener chaos. The proof is inspired by methods from statistical mechanics, especially, by a partition function that has been developed for the study of directed polymers in random environment.

Étant donné $k \geq 2$ marches aléatoires simples indépendantes sur $\mathbb{Z}$, on appelle (n,z) un évènement de collision s'il existe deux marches aléatoires distinctes $S^{(i)}, S^{(j)} (1 \leq i < j \leq k)$ telles que $S_n^{(i)} = S_n^{(j)} = z$. Dans cet article, on montre que sous le même changement d'échelle que dans le théorème de Donsker, les mesures aléatoires représentant ces évènement de collision convergent vers une mesure aléatoire non-triviale sur $[0,1] \times \mathbb{R}$. De plus, cette mesure aléatoire limite peut être caractérisée en utilisant les chaos de Wiener. La preuve s'inspire de méthodes issues de la mécanique statistique.

[14:15-14:30]

Yanbo Tang (University of Toronto)

Monte Carlo Integration in High Dimensions

L'intégration de Monte-Carlo en hautes dimensions

Monte Carlo integration is a commonly used technic to integrate low dimensional integrals. However, it is typically assumed to perform poorly for high-dimensional integrals. We examine the naïve Monte Carlo integration scheme through the lens of high dimensional statistics where the dimension of the integral is allowed to increase. In doing so, we derive non-asymptotic bounds for the relative error of the integral approximation in some general scenarios through some concentration inequalities. We demonstrate that the scaling in the number of points sampled needed to guarantee a consistent estimate can vary between polynomial and exponential, depending on the function being integrated, therefore showing that Monte Carlo integration in high dimensions is not an uniformly bad idea.

L'intégration de Monte-Carlo est une technique couramment utilisée pour intégrer des intégrales de faible dimension. Toutefois, sa performance est généralement présumée être faible pour des intégrales de haute dimension. Nous examinons le schème de l'intégration de Monte-Carlo naïve sous l'angle de statistiques de haute dimension dans lesquelles la dimension de l'intégrale peut s'accroître. Ce faisant, nous dérivons des limites non asymptotiques pour l'erreur relative de l'approximation des intégrales dans quelques scénarios généraux par le biais de certaines inégalités de concentration. Nous faisons valoir que la mise à l'échelle du nombre de points échantillonnés nécessaire pour garantir une estimation convergente peut varier entre polynomiale et exponentielle, dépendamment de la fonction qui est intégrée, ce qui montre conséquemment que l'intégration de Monte-Carlo en hautes dimensions n'est pas uniformément un mauvais choix.

[14:30-14:45]

Weinan Qi (University of Waterloo) **Paul Marriott** (University of Waterloo) **Yi Shen** (University of Waterloo)

Excursion sets and critical points of Gaussian random fields over high thresholds

Ensembles d'excursions et points critiques des champs aléatoires gaussiens au-delà de seuils élevés

Modeling the critical points of a Gaussian random field is an important challenge in stochastic geometry. In this talk, we focus on isotropic Gaussian random fields and study the location and type of critical points over high thresholds. Under certain conditions, we show that when the threshold tends to infinity and the searching area expands with a matching speed, both the location of the local maxima and the location of all critical points

La modélisation des points critiques d'un champ aléatoire gaussien présente un enjeu de taille en géométrie stochastique. Dans cette présentation, nous nous penchons particulièrement sur les champs aléatoires gaussiens isotropes et nous étudions l'emplacement et le type des points critiques au-delà de seuils élevés. Nous montrons que, dans certaines conditions, quand le seuil tend vers l'infini et que la zone de recherche s'étend à une vitesse correspondante, l'emplacement des maxima locaux et l'emplacement de

above the threshold converge weakly to a Poisson point process. We will further discuss the possibility to approximate these locations when the threshold is high but not extremely high. In particular, we explore the local behavior of critical points by looking at the type of a critical point given there is another critical point close to it. This is a joint work with Paul Marriott and Yi Shen.

tous les points critiques au-dessus du seuil convergent faiblement vers un processus ponctuel de Poisson. Nous traitons ensuite de la possibilité d'approximer ces emplacements lorsque le seuil est élevé, mais pas extrêmement élevé. Plus particulièrement, nous explorons le comportement local des points critiques en examinant le type d'un point critique lorsqu'il existe un autre point critique proche de celui-ci. Il s'agit de travaux conjoints avec Paul Marriott et Yi Shen.

[14:45-15:00]

Quinn Forzley (University of Winnipeg) **Shakhawat Hossain** (University of Winnipeg)

Pretest and Shrinkage Estimation Strategies in Accelerated Failure Time Model

Stratégies d'estimation de rétrécissement et de prétests dans un modèle à temps d'échec accéléré

This paper focuses on the accelerated failure time (AFT) model which is often suggested as an alternative to the Cox proportional hazards model. We propose pretest and shrinkage estimation methods for estimating regression parameters in the AFT model for right-censored data, when some parameters shrink to a restricted subspace. Shrinkage estimators take information from the unrestricted model and provide an efficient estimate of regression parameters by shrinking the unrestricted estimate toward the restricted model estimate. This technique increases bias in the estimation process but reduces overall mean-squared error, offsetting the bias. We show that if the shrinkage dimension exceeds two, the risk of the shrinkage estimator is strictly less than that of the maximum-likelihood estimator. Extensive numerical studies are conducted to evaluate the performance of the proposed estimators. An empirical application is provided to illustrate the practical usefulness of these estimators.

Cet article se concentre sur le modèle à temps d'échec accéléré (TÉA) que l'on suggère fréquemment en guise d'option pour remplacer le modèle de risque proportionnel de Cox. Nous proposons des méthodes d'estimation de rétrécissement et de prétests pour estimer les paramètres de régression dans le modèle à TÉA pour les données censurées à droite, lorsque certains paramètres rétrécissent en un sous espace restreint. Les estimateurs tirent l'information du modèle non restreint et procurent une estimation efficace des paramètres de régression en rétrécissant l'estimation non restreinte vers l'estimation du modèle restreint. Cette technique augmente le biais dans le processus d'estimation, mais réduit l'erreur quadratique moyenne globale, ce qui compense le biais. Nous démontrons que, lorsque la dimension de rétrécissement dépasse deux, le risque de l'estimateur de rétrécissement est sévèrement moindre par rapport à l'estimateur du maximum de vraisemblance. Nous menons des études numériques approfondies afin d'évaluer la performance des estimateurs proposés. Nous illustrons l'utilité pratique de ces estimateurs à partir d'une application empirique.

Identifying and Utilizing Group Structures in Heterogeneous Populations
Identification et utilisation des structures de groupe dans les populations hétérogènes

Chair/Président: Kevin McGregor

Date: Monday May 30 / lundi 30 mai

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-13:45]

Gyanendra Pokharel (The University of Winnipeg)

Classification-Based Inference for Spatial Infectious Disease Models Incorporating Infection Time Uncertainty

Inférence basée sur une classification pour des modèles spatiaux de maladies infectieuses incorporant une incertitude du temps d'infection

Mechanistic models are key to provide reliable information that can be used in developing strategies to control the spread of infectious diseases. These models are generally fitted in Bayesian Markov chain Monte Carlo (MCMC) frameworks. The parameter estimates from these frameworks are accurate but computationally expensive. This problem is more severe for the case when the epidemic history is incomplete, such as unknown infection times. As an alternative, we propose to use supervised learning methods to analyze the incomplete infectious disease data where the epidemic generating models are classified based on the epidemic summary statistics generated over the pre-defined design matrix of the parameters. We use both simulated and experimental data incorporating infection time uncertainty into the models to investigate the validity. We show that the proposed methods can be used to infer the transmission dynamics of infectious disease in the presence of infection events uncertainty.

Les modèles mécanistes sont importants, car ils fournissent de l'information fiable qui peut être utilisée pour contrôler la propagation de maladies infectieuses. Ces modèles sont généralement ajustés dans des cadres bayésiens de Monte-Carlo par chaînes de Markov (MCMC). L'estimation des paramètres tirée de ces cadres est précise mais coûteuse sur le plan computationnel. Ce problème s'aggrave lorsque l'historique de l'épidémie est incomplet, comme quand les temps d'infection sont inconnus. Comme solution de rechange, nous proposons l'utilisation de méthodes d'apprentissage supervisé pour l'analyse des données incomplètes relatives aux maladies infectieuses dans lesquelles les modèles générateurs d'épidémie sont classifiés selon les statistiques épidémiologiques résumées provenant de la matrice de schéma prédéfinie des paramètres. Nous utilisons des données simulées et expérimentales incorporant une incertitude du temps d'infection dans les modèles pour en étudier la validité. Nous montrons que les méthodes proposées peuvent servir à inférer la dynamique de transmission d'une maladie infectieuse en présence d'une incertitude concernant les événements infectieux.

[13:45-14:00]

Wangshu Tu (Carleton University) **Sanjeena Subedi** (Carleton University) **Ryan P. Browne** (University of Waterloo)

Mixtures of Logistic Skew-normal Multinomial Models

Mélanges de modèles logistiques multinomiaux asymétriques

The logistic normal multinomial distribution is gaining interest for modeling microbiome data. It utilizes a hierarchical structure such that the observed counts conditional on the compositions are assumed to be multinomial random variables and the log-ratio transformed compositions are assumed to be from a Gaussian distribution. While multinomial distribution accounts for the compositional nature of the data, and a Gaussian prior offers flexibility in structure of covariance matrices, the log-ratio transformed compositions of the microbiome

Il y a un intérêt grandissant pour la distribution normale logistique multinomiale pour la modélisation des données sur le microbiome. Cette distribution utilise une structure hiérarchique qui fait en sorte que les comptes observés conditionnels aux compositions sont supposés être des variables aléatoires multinomiales et les compositions transformées par log-ratio sont supposées découler d'une distribution gaussienne. Même si la distribution multinomiale prend en compte la nature compositionnelle des données et que l'a priori gaussien offre une souplesse dans la structure des matrices de covariables, les compositions transformées par

data can be highly skewed, especially at a lower taxonomic level. Fitting a Gaussian distribution can be detrimental to the model fit. Here, we propose a novel mixture of logistic skew-normal multinomial distributions in which a multivariate skew-normal distribution is utilized as a prior for the log-ratio transformed compositions. A variational Gaussian approximation in conjunction with the EM algorithm is utilized for parameter estimation.

log-ratio des données sur le microbiome peuvent être grandement asymétriques, en particulier à un niveau taxonomique plus bas. Ajuster une distribution gaussienne peut nuire à l'adéquation du modèle. Nous proposons un nouveau mélange de distributions logistiques multinomiales normales asymétriques dans lequel une distribution de normale multivariée asymétrique est utilisée comme a priori pour les compositions transformées par log-ratio. Une approximation gaussienne variationnelle conjuguée à l'algorithme espérance-maximisation est utilisée pour l'estimation des paramètres.

[14:00-14:15]

Dexen D.Z. Xi (Western University) **Masoud Adelzadeh** (National Research Council Canada)

Finite Mixture Models and Shared Frailty Models for Fire Department Response Time in Building Fires

Modèles de mélanges finis et modèles à fragilités partagées du temps de réponse des services d'incendie en cas de feu dans un bâtiment

The National Building Code of Canada limits/increases the spatial separation between buildings if the time from receipt of notification of a fire by the fire department until the arrival of the first fire department vehicle at the building exceeds 10 min in 10% or more of all fire department calls to the building. We develop a stepwise approach to understand the effect of fire preventive features (e.g., smoke detectors) and geo-spatial information to response time using data from the National Fire Information Database. A regression mixture suggests that among 55% of the incidents in Alberta whose response time are identified as long, fire preventive features are related to decreases in response time. A shared frailty model suggests that the agglomerations have unobserved effect on the average levels of the response time across the province, while within the agglomerations of Calgary, the response time is related to the distance to the nearest fire station.

Le Code national du bâtiment du Canada limite ou accroît l'espace entre les bâtiments si le délai depuis la réception d'une notification d'incendie par un service d'incendie jusqu'à l'arrivée de son premier véhicule au bâtiment dépasse 10 minutes dans au moins 10 % de tous les appels au service à se rendre au bâtiment. Nous développons une approche par étapes pour comprendre l'effet des mesures préventives contre l'incendie (par ex. : le détecteur de fumée) et de l'information géospatiale du temps de réponse, à l'aide de données tirées de la base de données de la National Fire Information Database. Un mélange de régression semble indique que dans 55 % des incidents en Alberta au cours desquels on a déterminé un long temps de réponse, les mesures de prévention contre l'incendie sont liées à une réduction du temps de réponse. Un modèle à fragilités partagées suggère que les agglomérations ont un effet non observé sur les délais moyens du temps de réponse dans la province, tandis dans les agglomérations de Calgary le temps de réponse est lié à la distance de la caserne de pompiers la plus proche.

[14:15-14:30]

Xiaoke Qin (Carleton University) **Sanjeena Dang** (Carleton University)

Mixtures of Generalized Dirichlet-Multinomial Models for Microbiome Data

Mélanges de modèles multinomiaux généralisés de Dirichlet pour des données sur le microbiome

Variations in microbiome compositions have been associated with several diseases. Mixtures of Dirichlet-multinomial distribution have been previously used to cluster individuals into groups based on their microbiome composition. A Dirichlet-multinomial distribution is a hierarchical distribution such that the observed counts conditional on the compositions are assumed to be multinomial and the compositions are assumed to be a random variable from a Dirichlet dis-

Les variations dans les compositions du microbiome ont été associées à plusieurs maladies. Des mélanges de lois multinomiales de Dirichlet ont été utilisés antérieurement pour séparer des individus dans des groupes selon la composition de leur microbiome. Une loi multinomiale de Dirichlet est une loi hiérarchique dans laquelle on suppose que conditionnellement aux compositions, les comptages observés sont multinomiaux, et que les compositions sont une variable aléatoire suivant une loi de Dirichlet. Cependant, la loi de Dirichlet a une structure de covariance

tribution. However, Dirichlet distribution has a restrictive covariance structure. An alternate to mixtures of Dirichlet-multinomial distributions are mixtures of generalized Dirichlet-multinomial (GDM) distributions, previously proposed in the literature. Here, we extend the mixtures of GDM distributions to cluster microbiome data and develop a parameter estimation framework using the minorization-maximization algorithm. Furthermore, generalizing these models for accounting for inter-individual heterogeneity will be discussed.

restrictive. Une solution alternative aux mélanges de lois multinomiales de Dirichlet sont les mélanges de lois multinomiales généralisées de Dirichlet (GDM), proposées antérieurement dans la littérature. Nous étendons les mélanges GDM pour regrouper des données sur le microbiome et développer un cadre d'estimation des paramètres en utilisant l'algorithme minimisation-maximisation. De plus, nous abordons une généralisation de ces modèles pour prendre en compte l'hétérogénéité entre les individus.

[14:30-14:45]

Zihang Lu (Queen's University) **Wendy Lou** (University of Toronto)

A Model-Based Approach for Clustering Developmental Trajectories with Complex Longitudinal Data

Approche basée sur un modèle pour le regroupement de trajectoires de développement avec des données longitudinales complexes

Longitudinal data are commonly collected in health studies and identifying subgroups of patients based on longitudinal profiles is a common task to understand the developmental patterns of diseases. Motivated by a Canadian birth cohort study, we propose a statistical method for clustering subjects into homogenous subgroups based on longitudinal markers. The proposed method allows modeling multiple discrete or continuous longitudinal markers measured at marker-specific time points. Comparison between the proposed and existing methods using real and simulated data will be presented and discussed.

Les données longitudinales sont couramment recueillies dans les études de santé et il est d'usage courant d'identifier des sous-groupes de patients sur la base de profils longitudinaux pour mieux comprendre les schémas de développement des maladies. Motivés par une étude canadienne de cohorte de naissance, nous proposons une méthode statistique qui permet de regrouper les sujets en sous-groupes homogènes sur la base de marqueurs longitudinaux. La méthode proposée permet de modéliser plusieurs marqueurs longitudinaux discrets ou continus mesurés à des points de temps spécifiques au marqueur. Nous présenterons et discuterons d'une comparaison entre la méthode proposée et les méthodes existantes à l'aide de données réelles et simulées.

[14:45-15:00]

Aida Eslami (Université Laval) **Hervé Abdi** (School of Behavioral and Brain Sciences, The University of Texas at Dallas)

Integrating Group Structure in the Multiple Correspondence Analysis

Intégration de la structure de groupe dans l'analyse des correspondances multiples

Nowadays, we have access to large amounts of heterogeneous and multidimensional data. For categorical/categorized variables multiple correspondence analysis (MCA) is the appropriate statistical method for exploratory research. MCA reduces the dimensionality of the data space and provides graphical representations that describe the relationships between the original variables and between the observations. Multivariate analysis methods -such as MCA-assume that observations are independent and originate from a homogeneous population. However, the observations often comprise several groups known a priori (e.g., sex, ethnicity); but, so far, there is no extension, for qualitative data, of discriminant analysis that can optimally consider group structure to identify a common structure to the different groups in

De nos jours, nous avons accès à de grandes quantités de données hétérogènes et multidimensionnelles. Pour les variables catégorielles/catégorisées, l'analyse des correspondances multiples (ACM) est la méthode statistique appropriée pour la recherche exploratoire. L'ACM réduit la dimensionnalité de l'espace de données et fournit des représentations graphiques qui décrivent les relations entre les variables d'origine et entre les observations. Les méthodes d'analyse multivarié -telles que l'ACM- supposent que les observations sont indépendantes et proviennent d'une population homogène. Cependant, les observations comprennent souvent plusieurs groupes connus a priori (e.g., sexe, ethnie) ; mais, jusqu'à présent, il n'existe pas d'extension de l'analyse discriminante qui puisse prendre en compte de manière optimale une structure de groupe pour identifier une structure commune aux différents groupes en ACM 1) en étudiant les groupes simul-

MCA, by 1) studying the groups simultaneously while 2) imposing appropriate constraints integrating within-group variations. To address these constraints, we developed a new method by extending approaches that we previously developed within the framework of factorial analysis in the case of quantitative variables.

tanément et 2) en imposant des contraintes appropriées intégrant la variation intra-groupes. Pour répondre à ces contraintes, nous avons développé une nouvelle méthode en utilisant les approches que nous avions précédemment développées dans le cadre de l'analyse factorielle pour le cas des variables quantitatives.

Chair/Président: Suborna Shekhor Ahmed

Date: Monday May 30 / lundi 30 mai

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-13:45]

Douglas Whitaker (Mount Saint Vincent University)

Investigation of Bivariate Grid-Type Items for Measuring Attitudes in Statistics Education: Preliminary Results

Items de type grille bivariée pour mesurer les attitudes en enseignement de la statistique : résultats préliminaires

Likert-type items are ubiquitous in attitude research in statistics education but imply a reciprocal relationship between positivity and negativity in the construct being measured. Based on historical challenges measuring some constructs in a widely used framework in statistics education (Eccles's Expectancy-Value Theory [EVT]), we speculate that the reciprocal relationship implied by the Likert-type items may not be appropriate. Evaluative Space Grid (ESG) items have been proposed as an alternative: respondents indicate their positivity and negativity on a grid that does not impose a reciprocal relationship. However, there have been relatively few studies that focus on ESG items. This presentation reports on a set of preliminary studies that seek to describe the psychometric properties of ESG items and document evidence of their appropriateness (or lack thereof) for measuring EVT constructs. Data have been collected from introductory statistics students and a general participant pool.

Les items de type Likert sont souvent utilisés dans la recherche sur les attitudes en enseignement de la statistique, mais ils impliquent une relation réciproque entre la positivité et la négativité dans le concept mesuré. Puisqu'il semble difficile de mesurer certains concepts dans un cadre largement utilisé en enseignement de la statistique (théorie de l'espérance-valeur d'Eccles, ou EVT), nous supposons que la relation réciproque impliquée par les items de type Likert n'est peut-être pas appropriée. Les items de type Evaluative Space Grid (ESG) ont été proposés comme alternative : les répondants indiquent leur positivité et leur négativité sur une grille qui n'impose pas de relation réciproque. Cependant, relativement peu d'études se sont concentrées sur les items ESG. Cette présentation rend compte d'un ensemble d'études préliminaires qui cherchent à décrire les propriétés psychométriques des items ESG et à documenter les preuves de leur adéquation (ou non) à la mesure des concepts EVT. Les données ont été recueillies auprès d'étudiants en introduction aux statistiques et d'un groupe de participants général.

[13:45-14:00]

Tharshanna Nadarajah (University of Toronto)

Teaching Through Collaboration

Enseigner par la collaboration

Collaboration-based teaching is another innovative teaching method where students collaborate on various projects/assignments. The modern world is globalized, and collaboration is a crucial skill necessary for all careers. Students can develop this skill in the classroom by taking part in lessons, studying and working in groups. Promoting collaboration is a challenging endeavour-and it doesn't just happen. Our learning activities need to be intentionally designed to stimulate real collaboration. We developed some strategies to encourage effective collaboration, such as constructing complex learning

L'enseignement basé sur la collaboration est une autre méthode d'enseignement innovante où les étudiants collaborent sur divers projets/travaux. Le monde moderne est globalisé, et la collaboration est une compétence cruciale nécessaire pour toutes les carrières. Les élèves peuvent développer cette compétence en classe en participant aux cours, en étudiant et en travaillant en groupe. Promouvoir la collaboration est une entreprise difficile - et cela ne se fait pas tout seul. Nos activités d'apprentissage doivent être conçues intentionnellement pour stimuler une véritable collaboration. Nous avons développé quelques stratégies pour encourager une collaboration efficace, comme construire des activités

activities; students are prepared to work as part of a team; avoiding free-riding opportunities; develop many opportunities for discussion and consensus. We surveyed to gather information on students' attitudes and behaviors regarding collaboration, and these data were correlated with course outcomes. Results indicate that collaboration improves student performance and that students believe collaboration is an effective learning method.

d'apprentissage complexes, préparer les étudiants à travailler en équipe, éviter les occasions de parasitisme et développer de nombreuses occasions de discussion et de consensus. Nous avons mené une enquête pour recueillir des informations sur les attitudes et les comportements des étudiants concernant la collaboration, et mis ces données en corrélation avec les résultats du cours. Les résultats indiquent que la collaboration améliore les performances des étudiants et que ces derniers pensent que la collaboration est une méthode d'apprentissage efficace.

[14:00-14:15]

Diana Katherine Skrzydlo (University of Waterloo)
Designing Authentic Assessments for Learning
Concevoir des évaluations authentiques pour l'apprentissage

Assessment is both to give students feedback on their learning as well as to assign grades. But unfortunately most students only see them as a means to obtain grades. How can we change this perception, and get students to see assessment as valuable learning experiences? As part of the Enhance Assessment Practices project in the Faculty of Math at UWaterloo, I have been conducting an extensive literature review of what assessment types STEM faculty are using, as well as outcomes and best practices. In this talk, I will discuss how statistics courses can use non-traditional assessment types such as interactive tutorials, communication-based activities, peer learning, case studies, and oral exams. Participants will come away with tangible and research-supported suggestions for incorporating more authentic assessment types into their courses.

L'évaluation sert à donner aux étudiants une rétroaction sur leur apprentissage et à attribuer des notes. Mais malheureusement, la plupart des étudiants ne les voient que comme un moyen d'obtenir des notes. Comment pouvons-nous amener les étudiants à considérer l'évaluation comme une expérience d'apprentissage profitable ? Dans le cadre du projet Améliorer les pratiques d'évaluation à la faculté de mathématiques de l'Université de Waterloo, j'ai mené une analyse documentaire des types d'évaluation utilisés par les professeurs de STEM, ainsi que des résultats et des meilleures pratiques. Je discuterai comment les cours de statistiques peuvent utiliser des types d'évaluation non traditionnels tels que des tutoriels interactifs, des activités de communication, l'apprentissage par les pairs, des études de cas et des examens oraux. Les participants repartiront avec des suggestions tangibles et fondées sur la recherche pour incorporer des types d'évaluation plus authentiques dans leurs cours.

[14:15-14:30]

Nathalie Moon (University of Toronto) **Liza Bolton** (University of Toronto) **Rebecca Christensen** (University of Toronto)
Reflective Writing in Statistics Courses
L'écriture réflexive dans les cours de statistiques

Students come to statistics courses for many different reasons: for some it is a program requirement towards a non-statistics degree, others intend on specializing in statistics, and others are still figuring out their path. Catering to a diverse audience poses many challenges, and one strategy I have found useful to support my students and increase student satisfaction is to include opportunities for reflection in my courses. For students, reflective writing tasks offer a framework to formulate their objectives, identify what they've already accomplished towards these, and plan concrete next steps to continue progressing. As an instructor, I have found that reading student reflections has helped me better under-

Les étudiants dans nos cours de statistique y sont pour plusieurs raisons : pour certains, il s'agit d'une exigence de leur programme d'études, d'autres ont l'intention de se spécialiser en statistique, et d'autres n'ont pas encore de trajectoire établie. Répondre aux besoins d'un groupe si diversifié pose de nombreux défis. Une stratégie que j'ai trouvé utile pour soutenir mes étudiants est d'inclure des tâches de réflexion dans mes cours. Pour les étudiants, les tâches d'écriture réflexive sont une chance de formuler leurs objectifs, d'identifier ce qu'ils ont déjà accompli pour les atteindre et de définir des prochaines étapes concrètes pour continuer à progresser. De mon côté, lire leurs réflexions m'aide à mieux comprendre mes étudiants et à adapter le cours à leurs besoins. Dans cette présentation, je présenterai les tâches de réflexion que j'ai

stand my students and adapt the course to their needs. In this session, I will share reflective writing prompts which I have successfully used in undergraduate statistics courses at various levels and comment on feedback received and lessons learned.

utilisées avec succès dans deux cours de statistiques de premier cycle de différents niveaux et je partagerai mon expérience et mes recommandations.

[14:30-14:45]

Nooshin Khobzi Rotondi (Ontario Tech University) **David Rudoler** (Ontario Tech University) **William Hunter** (Ontario Tech University) **Olayinka Sanusi** (Ontario Tech University) **Chris Collier** (Ontario Tech University) **Michael Rotondi** (York University)

Using a "midterm warning system" to improve student performance and engagement in an introductory statistics course: A randomized controlled trial

L'utilisation d'un « système d'avertissement à mi-parcours » pour améliorer les performances et l'engagement des étudiants dans un cours d'introduction aux statistiques : un essai comparatif aléatoire

We evaluated the effectiveness of e-mailed grade nudges on students' performance and engagement in an introductory statistics course for undergraduate health science students. In 2020-21, 358 students were randomized to an e-mail (n=178) or no e-mail (n=180) group. The intervention e-mail contained their predicted final grades. Our statistical analysis shows a higher compatibility with a model of no mean difference in final grades for students in the e-mail vs. no e-mail group. Comparison of the distributions of final grades between the two groups suggests the e-mailed nudges may be related to slight improvements in grades. Students also completed the Scale of Student Engagement in Statistics (SSE-S). Total engagement, affective and cognitive subscale scores were higher in the e-mail group, indicating low compatibility with a model of no difference in engagement scores. Our results show there is potential for our simple and cost-effective intervention to improve student outcomes.

Nous avons évalué l'efficacité des encouragements aux notes, envoyés par courriel, sur la performance et l'engagement des étudiants dans un cours d'introduction aux statistiques pour les étudiants de premier cycle en sciences de la santé. En 2020-2021, 358 étudiants ont été choisi aléatoirement dans un groupe avec e-mail (n = 178) ou sans e-mail (n = 180). Le courriel d'intervention contenait leurs notes finales prévues. Notre analyse statistique montre une plus grande compatibilité avec un modèle d'absence de différence moyenne dans les notes finales pour les étudiants du groupe e-mail par rapport au groupe sans e-mail. La comparaison des distributions des notes finales entre les deux groupes suggère que les encouragements envoyés par courriel pourraient être liés à de faibles améliorations des notes. Les étudiants ont également rempli l'Échelle d'engagement des étudiants en statistiques [Scale of Student Engagement in Statistics] (SSE-S). L'engagement total, les scores des sous-échelles affectives et cognitives étaient plus élevés dans le groupe des e-mails, indiquant une faible compatibilité avec un modèle sans différence dans les scores d'engagement. Nos résultats montrent que notre intervention simple et rentable offre le potentiel d'améliorer les résultats des élèves.

[14:45-15:00]

Melanie C. H. Gibbons (University of Saskatchewan) **Marc T. Avey** (Canadian Council on Animal Care) **Phyllis G. Paterson** (University of Saskatchewan)

Experimental Design and Statistics Training in Select Canadian Graduate Programs at U15 Universities

Formation en planification d'expérience et statistique dans certains programmes canadiens d'études supérieures des universités U15

Irreproducibility of animal studies is a documented problem, attributed in part to suboptimal experimental designs and flawed statistical analyses. Evidence of how investigators are trained in experimental design and statistics (EDS) is limited. We aimed to determine if courses in EDS are required for completion of thesis-based graduate programs at U15 universities

Le manque de reproductibilité des études sur les animaux est un problème documenté, attribué en partie à des plans d'expériences sous-optimaux et à des analyses statistiques défaillantes. On ne sait que trop peu comment les investigateurs sont formés en planification d'expérience et statistique (PES). Nous avons cherché à déterminer si des cours de PES sont requis dans les universités U15 pour compléter des programmes d'études supérieures

in a selection of disciplines in which animal studies are commonly performed. Through electronic searches of official sources of program information (e.g. online academic calendars), we found that EDS course requirements, recommendations, and pre-requisites are uncommon (¡ 30% of programs) for U15 graduate programs in animal science, neuroscience, and pharmacology, but more common (up to 93% of programs) for graduate programs in psychology. While our methods cannot determine what EDS training students receive, it is clear that this training often falls outside of formal program requirements. Funding: NSERC USRA

avec thèse dans une sélection de disciplines où des études sur les animaux sont couramment effectuées. Par des recherches électroniques dans les sources officielles d'information sur les programmes (par exemple, calendriers universitaires en ligne), nous avons constaté que les exigences, les recommandations et les prérequis des cours de PES sont peu fréquents (¡ 30 % des programmes) pour les programmes d'études supérieures en science animale, neuroscience et pharmacologie, mais plus courants (jusqu'à 93 % des programmes) pour les programmes d'études supérieures en psychologie. Bien que nos méthodes ne permettent pas de déterminer quelle formation en PES les étudiants reçoivent, il est clair que cette formation échappe souvent aux exigences officielles des programmes. Financement : CRSNG, USRA

Statistical Analysis of Complex Large-Scale Health Data
Analyse statistique des données complexes à grande échelle sur la santé

Chair/Président: Peter X. Song

Organizer/Responsable: Peter X. Song

Date: Monday May 30 / lundi 30 mai

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-16:00]

Ji Zhu (University of Michigan)

Fast Network Community Detection with Profile-Pseudo Likelihood Methods

Détection communautaire à réseau rapide avec des méthodes de pseudo vraisemblance profilée

The stochastic block model is one of the most studied network models for community detection. It is known that most algorithms proposed for fitting the stochastic block model likelihood function cannot scale to large-scale networks. One prominent work that overcomes this computational challenge is Amini et al. (2013), which proposed a fast pseudo-likelihood approach for fitting stochastic block models to large sparse networks. However, this approach does not have a convergence guarantee. In this talk, we present a novel likelihood based approach that decouples row and column labels in the likelihood function, which enables a fast alternating maximization; the new method is computationally efficient and has provable convergence guarantee. We also show that the proposed method provides strongly consistent estimates of the communities in a stochastic block model. As demonstrated in simulation studies, the proposed method outperforms the pseudo-likelihood approach in terms of both estimation accuracy and computation efficiency, especially for large sparse networks. We further consider extensions of the proposed method to handle networks with degree heterogeneity and bipartite properties. This is joint work with Jiangzhou Wang, Jingfei Zhang, Binghui Liu, and Jianhua Guo.

Le modèle de bloc stochastique est l'un des modèles de réseaux les plus étudiés pour la détection communautaire. Il est bien connu que la plupart des algorithmes proposés pour ajuster la fonction de vraisemblance du modèle de bloc stochastique ne peuvent pas échelonner les réseaux à grande échelle. Le travail d'Amini et autres (2013) est un exemple important pour avoir surmonté ce défi informatique en proposant une approche de pseudo-vraisemblance rapide dans le but d'ajuster les modèles de bloc stochastiques aux grands réseaux épars. Cependant, cette approche n'assure pas une convergence. Lors de cet exposé, nous présentons une nouvelle approche basée sur la vraisemblance qui dissocie des étiquettes de rangées et de colonnes dans la fonction de vraisemblance, qui permet une maximisation en alternance rapide. La nouvelle méthode est efficace sur le plan calculatoire et assure une convergence prouvable. Nous montrons aussi que la méthode proposée procure des estimations fortement cohérentes des communautés dans un modèle de bloc stochastique. Comme démontré dans les études en simulation, la méthode proposée surpasse l'approche de pseudo-vraisemblance en termes de précision de l'estimation et de l'efficacité sur le plan calculatoire, tout particulièrement pour les grands réseaux épars. De plus, nous abordons les extensions de la méthode proposée pour gérer les réseaux ayant un degré d'hétérogénéité et de propriétés bipartites. Il s'agit d'un travail conjoint avec Jiangzhou Wang, Jingfei Zhang, Binghui Liu, et Jianhua Guo.

[16:00-16:30]

Bin Nan (University of California, Irvine) **Yue Wang** (University of California) **Jack Kalbfleisch** (University of Michigan)

Kernel Estimation of Bivariate Time-varying Coefficient Model for Longitudinal Data with Terminal Event

Estimation par noyau d'un modèle bivarié à coefficients variant dans le temps pour données longitudinales avec événement terminal

We propose a nonparametric bivariate time-varying coefficient model for longitudinal measurements with the

Nous proposons un modèle non paramétrique bivarié à coefficients variant dans le temps pour les mesures longitudinales avec occur-

occurrence of a terminal event that is subject to right censoring. The time-varying coefficients capture the longitudinal trajectories of covariate effects along with both the follow up time and the residual lifetime. The proposed model extends the parametric conditional approach given terminal event time in recent literature, and thus avoids potential model misspecification. We consider a kernel smoothing method for estimating regression coefficients in our model and use cross-validation for bandwidth selection, applying undersmoothing in the final analysis to eliminate the asymptotic bias of the kernel estimator. We show that the kernel estimates follow a finite-dimensional normal distribution asymptotically under mild regularity conditions, and provide an easily computable sandwich covariance matrix estimator. We conduct extensive simulations that show desirable performance of the proposed approach, and apply the method to analyzing the medical cost data for patients with end-stage renal disease.

rence d'un événement terminal soumis à une censure à droite. Les coefficients variant dans le temps capturent les trajectoires longitudinales des effets des covariables ainsi que le temps de suivi et la durée de vie résiduelle. Le modèle proposé étend l'approche conditionnelle paramétrique compte tenu du temps à événement terminal dans la littérature récente, et évite ainsi la mauvaise spécification potentielle du modèle. Nous considérons une méthode de lissage à noyau pour estimer les coefficients de régression dans notre modèle et utilisons la validation croisée pour sélectionner la largeur de bande, en appliquant le sous-lissage dans l'analyse finale pour éliminer le biais asymptotique de l'estimateur à noyau. Nous montrons que les estimateurs à noyau suivent asymptotiquement une distribution normale de dimension finie sous de légères conditions de régularité, et fournissons un estimateur de matrice de covariance sandwich facilement calculable. Nous effectuons des simulations approfondies qui montrent les performances souhaitables de l'approche proposée et nous appliquons la méthode à l'analyse des données sur les coûts médicaux des patients atteints d'insuffisance rénale terminale.

[16:30-17:00]

Annie Qu (University of California, Irvine)

Optimal Individualized Omni-channel Treatment Decision Rule Under Budget Constraints

Règle décisionnelle de traitement omnicanal individualisé optimal selon des contraintes budgétaires

Individualized treatment rule (ITR), which recommends the optimal treatment based on individual characteristics, has drawn considerable interests from many areas such as precision medicine, personalized education, and personalized marketing. Existing ITR estimation methods mainly adopt one of two or more treatments. However, a combination of multiple treatments, or omni-channel treatments, could be more powerful in various areas. In this talk, we propose a novel double-encoder framework to estimate the individualized omni-channel treatment rule (IOTR). The proposed method incorporates the interaction effects among different channels and utilizes correlations among different omni-channel treatments. In addition, we propose a novel IOTR estimation under budget constraints to facilitate optimal decisions with limited resources. In theory, we show that the proposed method achieves a faster convergence rate of the value reduction bound compared with existing methods for multi-arm treatments. Our simulation studies show that the proposed method outperforms the existing ITR estimation in various settings. We also demonstrate the superior performance of the proposed method in a real data application that recommends optimal omni-channel treatments for Type-2 diabetes pa-

La règle de traitement individualisée (ITR), qui recommande un traitement optimal fondé sur les caractéristiques d'un individu, a beaucoup attiré l'attention dans plusieurs domaines comme la médecine de précision, l'enseignement personnalisé et le marketing personnalisé. Les méthodes d'estimation de l'ITR adoptent principalement un traitement parmi plusieurs. Toutefois, une combinaison de plusieurs traitements (ou traitement omnicanal) pourrait s'avérer efficace dans certains domaines. Dans cet exposé, nous proposons un nouveau double encodeur pour estimer la règle de traitement omnicanal individualisé (IOTR). La méthode proposée intègre les effets d'interaction entre différents canaux et se sert des corrélations parmi les différents traitements omnicanals. De plus, nous proposons une nouvelle estimation de IOTR selon des contraintes budgétaires pour mieux arriver à des décisions optimales avec des ressources limitées. En théorie, nous montrons que la méthode proposée parvient à un taux de convergence plus rapide de la limite de réduction de valeur par rapport aux méthodes existantes de traitements multibranches. Nos études par simulations démontrent que la méthode proposée surpasse l'estimation d'ITR actuelle dans plusieurs contextes. De plus, sa performance est aussi supérieure dans son application sur des données réelles qui recommandent des traitements omnicanal optimaux pour les patients souffrant de diabète de type 2.

tients.

New Statistical Methods for Adaptive Clinical Trial Design
Nouvelles méthodes statistiques pour la conception d'essais cliniques adaptatifs

Chair/Président: Wei Xu

Organizer/Responsable: Depeng Jiang

Date: Monday May 30 / lundi 30 mai

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-16:00]

Ying Yuan (University of Texas MD Anderson Cancer Center)

Elastic Priors to Dynamically Borrow Information from Historical Data in Clinical Trials

Distributions a priori élastiques pour l'emprunt dynamique d'information tirée de données historiques d'essais cliniques

Use of historical data and real-world evidence holds great potential to improve the efficiency of clinical trials. One major challenge is to effectively borrow information from historical data while maintaining a reasonable type I error and minimal bias. We propose the elastic prior approach to address this challenge. Unlike existing approaches, this approach proactively controls the behavior of information borrowing and type I errors by incorporating a well-known concept of clinically significant difference through an elastic function, defined as a monotonic function of a congruence measure between historical data and trial data. The elastic function is constructed to satisfy a set of prespecified criteria such that the resulting prior will strongly borrow information when historical and trial data are congruent, but refrain from information borrowing when historical and trial data are incongruent. The elastic prior approach has a desirable property of being information borrowing consistent, i.e. asymptotically controls type I error at the nominal value, no matter that historical data are congruent or not to the trial data. Our simulation study that evaluates the finite sample characteristic confirms that, compared to existing methods, the elastic prior has better type I error control and yields competitive or higher power. The proposed approach is applicable to binary, continuous and survival endpoints.

L'emploi de données historiques et de preuves réelles a un potentiel élevé pour améliorer l'efficacité des essais cliniques. L'un des défis principaux est d'emprunter efficacement l'information tirée des données historiques tout en conservant une erreur de type I raisonnable et un minimum de biais. Pour relever ce défi, nous proposons une distribution a priori élastique. Contrairement aux approches existantes, la nôtre contrôle de façon proactive le comportement d'emprunt d'information et des erreurs de type I en intégrant un concept bien connu de différence significative clinique par l'entremise d'une fonction élastique, définie comme une fonction monotone d'une mesure de congruence entre les données historiques et les données d'un essai. La fonction élastique est construite pour satisfaire un ensemble de critères préalablement spécifiés pour faire en sorte que la distribution a priori générée empruntera fortement l'information lorsque les données historiques et d'essais sont congruentes, mais pas si elles ne sont pas congruentes. La distribution a priori élastique comporte l'avantage d'être cohérente dans son emprunt d'information, p. ex. elle peut contrôler asymptotiquement l'erreur de type I à la valeur nominale, peu importe si les données sont congruentes ou non par rapport aux données d'un essai. Notre étude en simulation qui évalue la caractéristique de l'échantillon fini confirme que la distribution a priori élastique a un meilleur contrôle de l'erreur de type I et génère une puissance supérieure par rapport aux méthodes existantes. L'approche proposée est applicable aux indicateurs de résultats binaires, continus et de survie.

[16:00-16:30]

Suyu Liu (MD Anderson Cancer Center) **Beibei Guo** (Louisiana State University) **Elizabeth Garrett-Mayer** (American Society of Clinical Oncology)

A Bayesian Phase I/II Design for Cancer Clinical Trials Combining Immunotherapy and Chemotherapy

Plan bayésien de phase I/II pour les essais cliniques de traitement du cancer combinant l'immunothérapie et la chimiothérapie

Immunotherapy is an innovative treatment approach that harnesses a patient's immune system to treat cancer. It has provided an alternative and complementary treatment modality to conventional chemotherapy. Combining immunotherapy with cytotoxic chemotherapy agent has become the leading trend and the most active research field in oncology. To accommodate this growing trend, we propose a Bayesian phase I/II dose-finding design to identify the optimal biological dose combination (OBDC), defined as the dose combination with the highest desirability in the risk-benefit tradeoff. We propose new statistical models to describe the relationship between the doses and treatment outcomes, including immune response, toxicity, and progression-free survival (PFS). During the trial, based on accrued data, we continuously update model estimates and adaptively assign patients to dose combinations with high desirability. The simulation study shows that our design has desirable operating characteristics.

L'immunothérapie est une approche innovante de traitement qui mise sur le système immunitaire du patient pour traiter le cancer. Cette approche a fourni une solution de rechange et un mode de traitement complémentaire à la chimiothérapie conventionnelle. La combinaison de l'immunothérapie avec un agent de chimiothérapie cytotoxique est devenue la tendance prédominante et le champ de recherche le plus actif en oncologie. Pour répondre à cette tendance croissante, nous proposons un plan bayésien de dosage de phase I/II afin d'identifier la combinaison optimale de doses biologiques (OBDC), définie comme la combinaison de doses la plus désirable sur le plan du rapport risques-bienfaits. Nous proposons de nouveaux modèles statistiques pour décrire la relation entre les doses et les résultats du traitement, y compris la réponse immunitaire, la toxicité et la survie sans progression (PFS). Pendant l'essai basé sur des données accumulées, nous avons constamment mis à jour les estimations des modèles et apparié de façon adaptative les patients à des combinaisons de doses à forte désirabilité. L'étude en simulation indique que notre concept a des caractéristiques opérationnelles désirables.

[16:30-17:00]

Depeng Jiang (University of Manitoba) **Bosheng Li** (University of Manitoba) **Fangrong Yan** (China Pharmaceutical University)

A Bayesian Adaptive Design for an Immunotherapy with Heterogeneous Delayed Treatment Effect

Plan adaptatif bayésien pour une immunothérapie à effet tardif du traitement hétérogène

The delayed treatment effect is common for the immunotherapies and the delay times are heterogeneous. There are however few statistical methods available to specify the delay time distribution in immunotherapy trial design. The misspecification of the delay time distribution may lead to a substantial loss of desired power. In this paper, we proposed a Bayesian adaptive design with the promising zone defined purely based on the prediction of a clinical efficacy in terms of the value of the log-rank test statistics. The final sample size will be changed at the interim analysis if the log-rank test statistics falls into the promising zone. Specifically speaking, we will re-estimate the parameters of the distribution of the delay time using the MCMC approach and then modify the final critical value and calculate the conditional power of the log-rank test. The simulation studies illustrate the merits of our proposed Bayesian adaptive design over traditional designs.

L'effet tardif du traitement est fréquent dans le cas des immunothérapies, et les temps de retard sont hétérogènes. Il existe cependant peu de méthodes statistiques permettant de définir la distribution des temps de retard dans les plans d'essais d'immunothérapie. Par ailleurs, une mauvaise définition de la distribution des temps de retard peut entraîner une perte substantielle de la puissance souhaitée. Dans cette présentation, nous proposons un plan adaptatif bayésien dans lequel la zone favorable est définie uniquement sur la base de la prédiction d'une efficacité clinique en termes de valeur des statistiques du test logarithmique par rangs. La taille finale de l'échantillon est modifiée lors de l'analyse intermédiaire si la statistique du test logarithmique par rangs tombe dans la zone favorable. Plus précisément, nous réestimons les paramètres de la distribution des temps de retard à l'aide de la méthode de Monte-Carlo par chaînes de Markov, puis nous modifions la valeur critique finale et calculons la puissance conditionnelle du test logarithmique par rangs. Nous illustrons, par des études de simulation, les avantages de notre plan adaptatif bayésien par rapport aux plans traditionnels.

Leadership and Women in Statistics
Leadership et femmes en statistique

Chair/Président: Thérèse A. Stukel

Organizer/Responsable: Thérèse A. Stukel

Date: Monday May 30 / lundi 30 mai

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-17:00]

Charmaine B. Dean (University of Waterloo) **Nadia Ghazzali** (Université du Québec à Trois-Rivières) **Amanda Golbeck** (University of Arkansas for Medical Sciences) **Lisa Strug** (University of Toronto)

Leadership and Women in Statistics

Leadership et femmes en statistique

Women have made major contributions to statistical theory and methods although they have generally been in the background with regards to leadership in government, academia and professional associations. This invited panel session will bring together a diverse panel from leadership positions in statistics and related STEM fields, and explore barriers that women encounter in their path to leadership; skills that enable the pathways to leadership; and strategies and behaviours that successful leaders in STEM fields embody and practice.

Les femmes ont largement contribué à la théorie et aux méthodes statistiques, mais elles sont généralement restées en retrait du leadership au sein du gouvernement, du monde universitaire et des associations professionnelles. Cette table ronde invitée réunira des personnes occupant divers postes de direction en statistique et domaines STEM connexes, et explorera les obstacles que les femmes rencontrent dans leur parcours vers le leadership, les compétences qui permettent d'y accéder et les stratégies et comportements qu'incarnent et pratiquent les leaders qui réussissent dans les domaines STEM.

Fifty Years of Statistics Teaching
Cinquante ans d'enseignement de la statistique

Chair/Président: Becky Wei Lin

Organizer/Responsable: Becky Wei Lin

Date: Monday May 30 / lundi 30 mai

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-16:00]

Kenneth Laurence Weldon (Simon Fraser University)

Teaching the Big Ideas of Statistics

L'enseignement des grands concepts statistiques

In this talk I present my personal perspective, informed by teaching, research and consulting experiences, of the changes in statistics instruction over the last fifty years. The theme of the talk is that undergraduate statistics teaching, especially at the lower division level, has been too much a follower of textbooks and has not made sufficient use of guided immersive experience. Computer technology has reduced the role of textbooks in providing the technical aspects of statistics. But there are non-technical aspects of statistics that need instructional support. The conceptual ideas that underlie the techniques require emphasis. I suggest that a conceptual approach will improve the effectiveness of statistics instruction and will also broaden the audience. I outline some of the big ideas of statistics that are useful for a broad range of careers, and how textbook-based instruction, even with modern pedagogical software, has failed to emphasize these big ideas. Finally, I suggest an undergraduate teaching curriculum that might be a practical way to achieve the desired results.

J'apporte ici mon point de vue personnel, alimenté par mes expériences en enseignement, recherche et consultation, sur les changements dans l'enseignement de la statistique survenus au cours des 50 dernières années. J'avance que l'enseignement de la statistique au premier cycle, surtout dans les premières années, s'en est trop tenu aux manuels et n'a pas suffisamment fait appel à l'expérience immersive guidée. La technologie informatique a réduit le rôle des manuels en fournissant les aspects techniques de la statistique. Il demeure pourtant des aspects non techniques qui exige un soutien pédagogique. L'accent doit porter sur les idées conceptuelles qui sous-tendent les techniques. Je soumets qu'une approche conceptuelle de l'enseignement de la statistique en améliorera l'efficacité, tout en élargissant son public. Je souligne que certains des grands concepts statistiques sont utiles dans un large éventail de carrières et que l'enseignement fondé sur les manuels, même avec les outils logiciels pédagogiques modernes, n'a pas réussi à valoriser ces concepts importants. Enfin, je propose un programme d'enseignement au premier cycle qui pourrait être un moyen pratique d'obtenir les résultats désirés.

[16:00-16:30]

Bethany J.G. White (University of Toronto)

Leveraging Technology in Statistics Education: A Look at Developments since 2000

Tirer parti de la technologie dans l'enseignement des statistiques : un regard sur les développements depuis 2000

Technology has been playing an increasingly important role in our professional lives. The dramatic transformation the computing and educational technology landscapes have experienced over the last few decades, in particular, have completely revolutionized the way we do things in statistics and data science and have been presenting us with new challenges and opportunities for teaching and learning. Careful integration of technology

La technologie joue un rôle de plus en plus important dans notre vie professionnelle. Les transformations spectaculaires qu'ont connues les paysages de l'informatique et de la technologie éducative au cours des dernières décennies, en particulier, ont complètement révolutionné notre façon de faire dans le domaine de la statistique et de la science des données et nous ont présenté de nouveaux défis et de nouvelles opportunités pour l'enseignement et l'apprentissage. L'intégration minutieuse de la technologie

into our teaching has allowed us to engage our students in different and meaningful ways, and can go a long way to support our students' learning and educational experiences. In this talk, I'll share a literature-informed tour of advances in online and technology-enhanced teaching and learning in statistics in recent decades, highlight some of the digital tools and resources I have used to support student learning in my own teaching practice, and reflect on current technologies and those that show promise for statistics education going forward.

dans notre enseignement nous a permis d'impliquer nos étudiants de manière différente et significative, et peut grandement contribuer à soutenir l'apprentissage et les expériences éducatives de nos étudiants. Dans cet exposé, je partagerai un tour d'horizon, basé sur la littérature, des progrès réalisés dans l'enseignement et l'apprentissage en ligne et assisté par la technologie en statistique au cours des dernières décennies, je mettrai en évidence certains outils et ressources numériques que j'ai utilisés pour soutenir l'apprentissage des étudiants dans ma propre pratique d'enseignement, et je réfléchirai aux technologies actuelles et prometteuses pour l'enseignement de la statistique à l'avenir.

[16:30-17:00]

Jerald F. Lawless (University of Waterloo)

Statistics Education 1972-2022: Some Past Developments and A Look Ahead

L'enseignement des statistiques de 1972 à 2022 : évolution passée et perspectives d'avenir

I will review developments in undergraduate statistics education over the past 50 years. This will include comments on theory and methods, computing and areas of application, as well as the orientation and background of undergraduates in statistics. In looking ahead I'll consider whether statistics still has a set of core principles and ideas that should be taught to all students, and what those may be. I'll attempt to argue that in designing curricula and courses we should consider core material and areas of application, but also how to maximize the potential and impact of statistics for good in science and society.

Je passerai en revue l'évolution de l'enseignement de la statistique au cours des 50 dernières années. Je formulerai mes commentaires notamment sur la théorie et les méthodes, l'informatique et les domaines d'application, ainsi que l'orientation et le parcours des étudiants de premier cycle en statistique. Pour les perspectives d'avenir, je tenterai de déterminer si les statistiques ont encore un ensemble de principes et d'idées fondamentaux qui devraient être enseignés à tous les étudiants, et quels sont ces principes et idées. J'essaierai de démontrer que, lorsque nous concevons des programmes d'études et des cours, nous devons tenir compte du matériel et des domaines d'application fondamentaux, mais aussi de la manière de maximiser le potentiel et les effets positifs des statistiques sur la science et la société.

Change-point Detection
Détection des points de changement

Chair/Président: Bruno N Rémillard

Organizer/Responsable: Bruno N Rémillard

Date: Monday May 30 / lundi 30 mai

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-16:00]

Bouchra Nasri (Université de Montréal) **Bruno Rémillard** (HEC Montréal) **Tarik Bahroui**

Change-Point Problems for Multivariate Time Series Using Pseudo-Observations

Problèmes de points de rupture pour les séries chronologiques multivariées utilisant des pseudo-observations

In this talk, I show that under weak assumptions, the change-point tests designed for independent random vectors can also be used with pseudo-observations for testing change-point in the joint distribution of non-observable random vectors, the associated copula, or the margins, without modifying the limiting distributions. In particular, change-point tests can be applied to the residuals of stochastic volatility models or conditional distribution functions applied to the observations, which are prime examples of pseudo-observations. Since the limiting distribution of test statistics depends on the unknown joint distribution function or its associated unknown copula when the dimension is greater than one, I also show that iid multipliers and traditional bootstrap can be used with pseudo-observations to approximate P-values for the test statistics. Numerical experiments and examples of applications to change-point problems are given.

Dans cet exposé, je montre que sous des hypothèses faibles, les tests de point de rupture conçus pour des vecteurs aléatoires indépendants peuvent également être utilisés avec des pseudo-observations pour tester le point de rupture dans la distribution conjointe de vecteurs aléatoires non observables, la copule associée, ou les marges, sans modifier les distributions limites. En particulier, les tests de rupture peuvent être appliqués aux résidus des modèles de volatilité stochastique ou aux fonctions de distribution conditionnelle appliquées aux observations, qui sont de parfaits exemples de pseudo-observations. Étant donné que la distribution limite des statistiques de test dépend de la fonction de distribution conjointe inconnue ou de sa copule inconnue associée lorsque la dimension est supérieure à un, je montre également que les multiplicateurs iid et le bootstrap traditionnel peuvent être utilisés avec des pseudo-observations pour approximer les valeurs P pour le statistiques des tests. Des expériences numériques et des exemples d'applications aux problèmes de points de rupture sont donnés.

[16:00-16:30]

Sévérien Nkurunziza (University of Windsor)

Some Inference Problems in Generalized Ornstein-Uhlenbeck Processes with Change-Points

Quelques problèmes d'inférence dans les processus d'Ornstein-Uhlenbeck généralisés avec des points de rupture

We present inference methods in generalized Ornstein-Uhlenbeck processes with unknown change-points when the drift parameter may satisfy a restriction. A Salient feature of this research consists in the fact that the number of change-points and their locations are unknown. We generalize recent findings in five ways. First, our method incorporates the uncertain prior knowledge. Second, we derive the unrestricted estimator (UE) and the restricted estimator (RE) as well as their asymptotic properties. Third, we establish a test for testing the con-

Nous présentons des méthodes d'inférence dans les processus d'Ornstein-Uhlenbeck généralisés avec points de rupture inconnus lorsque le paramètre de dérive peut satisfaire à une restriction. Un fait saillant de cette recherche réside dans le fait que le nombre de points de rupture et leurs emplacements sont inconnus. Nous généralisons les résultats récents de cinq façons. Tout d'abord, notre méthode intègre de l'information à priori. Ensuite nous établissons l'estimateur sans restriction (UE), l'estimateur restreint (RE) ainsi que leurs propriétés asymptotiques. Troisièmement, nous établissons un test de restriction ainsi que

straint and we derive its asymptotic power. Fourth, we propose a class of shrinkage estimators (SEs) which includes as special cases the UE, RE. Fifth, we study the relative dominance of the proposed estimators, and we establish that SEs dominate the UE. The additional novelty of our methods consists in an established asymptotic result which is used to overcome the issue due to the fact that the dimensions of the estimators are random.

sa puissance asymptotique. Quatrièmement, nous proposons une classe d'estimateurs à rétrécissement (SE) qui comprend comme cas spéciaux l'UE, RE. Finalement, nous étudions la dominance relative de ces estimateurs et prouvons que les SE dominent l'UE. La nouveauté de plus réside dans un résultat asymptotique établie afin de surmonter le problème de dimensions aléatoires.

[16:30-17:00]

Zhou Zhou (University of Toronto) **Weichi Wu** (Tsinghua University)

Multiscale Jump Testing and Estimation Under Complex Temporal Dynamics

Test et estimation multi-échelles de sauts sous dynamique temporelle complexe

We consider the problem of detecting jumps in an otherwise smoothly evolving trend whilst the covariance and higher-order structures of the system can experience both smooth and abrupt changes over 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. For a wide class of non-stationary time series models and trend functions, the proposed method is shown to be able to detect all jump points within a nearly optimal range with a prescribed probability asymptotically under mild conditions. For a time series of length n, the computational complexity of the proposed method is $O(n)$ for each scale and $O(n\log^{1+\varepsilon} n)$ overall, where ε is an arbitrarily small positive constant.

Nous étudions des problèmes de détection des sauts dans une tendance qui par ailleurs évolue en douceur, tandis que la covariance et les structures d'ordre supérieur du système peuvent connaître des changements à la fois progressifs et brusques dans le temps. Le nombre de points de sauts peut diverger à l'infini avec une possibilité de rétrécissement des tailles de sauts à zéro. La méthode est basée sur une application multi-échelle d'un filtre passe-optimal de sauts à la série temporelle, application dans laquelle les échelles sont denses entre les bornes inférieure et supérieure admissibles. Pour une grande classe de séries temporelles non stationnaires et de fonctions tendances, la méthode proposée a montré sa capacité à détecter tous les points de sauts dans une étendue presque optimale avec une probabilité prescrite asymptotiquement sous faibles conditions. Pour une série temporelle d'une longueur n, la complexité computationnelle de la méthode proposée est de $O(n)$ pour chaque échelle et de $O(n\log^{1+\varepsilon} n)$ globalement lorsque ε est une constante positive arbitrairement petite.

Graduate Research in Actuarial Science 1
Recherche aux cycles supérieures en science actuarielle 1

Chair/Président: Yi Lu

Date: Monday May 30 / lundi 30 mai

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-15:45]

Sebastian F Calcetero (University of Toronto)

A functional Severity Regression Model for Applications in General Insurance

Modèle de régression fonctionnel de la sévérité pour des applications en assurance générale

Loss modeling is of critical concern in actuarial science for accurate pricing, reserving, and risk assessment. Claim sizes in a portfolio are heterogeneous due to attributes of policyholders, resulting in behaviors such as multi-modality, skewness, or heavy tails. Similarly, empirical evidence shows that the distribution of claim sizes is different for small, body, and large claim sizes. Hence, traditional parametric methods cannot account for all the behavior of losses distribution and give a poor predictive performance. That said, we present a semiparametric severity model in which the whole distribution of claim sizes is learned along with covariates based on a functional regression model for probability densities. We show desirable flexibility and consistency properties of the model and illustrate how it can capture both the distribution and the varying effect of covariates for small and large claims in an interpretative fashion in a real data set in automobile insurance.

La modélisation des pertes est une préoccupation importante en science actuarielle pour l'évaluation précise des prix, de la réserve de sinistres et des risques. Les tailles des réclamations dans un portefeuille sont hétérogènes en raison des attributs des détenteurs de police, ce qui entraîne des comportements comme la multimodalité, l'asymétrie et les queues lourdes. De façon similaire, des évidences empiriques indiquent que la distribution des tailles de réclamation diffère selon que le sinistre est mineur, corporel ou majeur. Ainsi, les méthodes paramétriques traditionnelles ne peuvent pas prendre en compte tous les comportements de la distribution de pertes et leur performance prédictive est donc faible. Cela dit, nous présentons un modèle semi-paramétrique de sévérité dans lequel la distribution complète des tailles de réclamation est apprise avec des covariables basées sur un modèle de régression fonctionnel pour les densités de probabilité. Un ensemble de données réelles en assurance automobile sert à montrer des propriétés désirables de flexibilité et de consistance du modèle et illustre comment celui-ci peut saisir à la fois la distribution et l'effet variable des covariables pour des petites et grosses réclamations.

[15:45-16:00]

Sébastien Jessup (Concordia University) **Mélina Mailhot** (Concordia University) **Mathieu Pigeon** (Université du Québec à Montréal)

On the Impact of Model Combination Methods on Extreme Precipitation Projections

L'impact de combinaisons de modèles sur les projections de précipitations extrêmes

In recent years, various model combination techniques have been considered in a wide range of applications. From non-parametric to Bayesian approaches, different methods rely on varying assumptions potentially leading to very different results. We apply multiple model combination methods to an ensemble of 24 experts in a pooling approach and use the differences in outputs from model combination methods to illustrate how one can gain additional insight from using multiple meth-

Dans les dernières années, plusieurs techniques de combinaison de modèles ont été considérées dans une large gamme d'applications. Que ce soit des approches non-paramétriques ou bayésiennes, différentes méthodes font appel à des hypothèses variées pouvant potentiellement mener à des résultats très différents. Nous appliquons plusieurs méthodes de combinaison de modèles à un ensemble de 24 experts dans une approche d'aggrégation et utilisons les différences entre les résultats de combinaison pour illustrer la valeur ajoutée d'utiliser plus d'une méthode. Les changements

ods. Areal reduction factor (ARF) and quantile projected changes are used to show that consistency, or lack thereof, across approaches reflects the uncertainty of combination techniques. This shows how one should use more than one combination method, seeing as a single method can lead to overconfidence in projections.

projetés de facteurs de réduction surfacique (ARF) et de quantiles sont utilisés pour montrer que le niveau de cohérence entre les résultats de chaque approche reflète le niveau d'incertitude entourant les résultats. Ceci démontre qu'on devrait utiliser plus d'une méthode de combinaison, étant donné qu'une seule méthode peut mener à un excès de confiance par rapport aux projections.

[16:00-16:15]

Mingren Yin (University of Waterloo)

Optimal Deductible Reinsurance with Model Uncertainty

Réassurance avec franchise optimale et incertitude du modèle

In the literature of reinsurance design problems, the distribution of the underlying risk is commonly assumed to be known. However, the estimation to the true distribution is prone to error. Thus, researchers are interested in the performance of reinsurance contracts in the worst-case scenario and designing the optimal contract with the consideration of distribution uncertainty. In this work, we first discuss the worst-case distributions from both perspectives of the insurer and the reinsurer, associated with a deductible insurance. Two parties of the reinsurance contract adopt general distortion risk measure to quantify their losses. We assume that an uncertainty sets includes all distributions having the same first two moments and being "close enough" from a given reference distribution. Depending on the insurer and reinsurer's choices of the reference distribution, mean and the variance, their uncertainty sets may be quite different. Furthermore, using various kinds of premium principles and risk measures, we can optimize the deductible level and determine the optimal stop-loss contract.

Dans la littérature des problèmes de conception de réassurance, la distribution du risque sous-jacent est généralement supposée connue. Cependant, l'estimation de la distribution réelle est sujette à erreur. Ainsi, les chercheurs s'intéressent à la performance des contrats de réassurance dans le pire des cas et à la conception du contrat optimal tout en tenant compte de l'incertitude de la distribution. Dans cette présentation, nous discutons d'abord des distributions dans le pire des cas, du point de vue de l'assureur et du réassureur, associées à une assurance avec franchise. Les deux parties du contrat de réassurance adoptent une mesure de risque de distorsion générale pour quantifier leurs pertes. Nous supposons qu'un ensemble d'incertitudes comprend toutes les distributions ayant les mêmes deux premiers moments et « assez proches » d'une distribution de référence donnée. Selon les choix de l'assureur et du réassureur concernant la distribution de référence, la moyenne et la variance, leurs ensembles d'incertitude peuvent être très différents. En outre, en utilisant différents types de principes de prime et de mesures de risque, nous pouvons optimiser le niveau de franchise et déterminer le contrat stop-loss optimal.

[16:15-16:30]

Meng Sun (Simon Fraser University) **Yi Lu** (Simon Fraser University)

Statistical Modeling of Data Breaches and its Application in Cyber Insurance

Modélisation statistique des fuites de données et son application dans la cyberassurance

Data breach incidents result in severe financial loss and reputational damage, which raises the importance of using insurance to manage and mitigate cyber-related risks. We analyze data breach chronology collected by Privacy Rights Clearinghouse (PRC) since 2005 and propose a generalized linear mixed model for data breach incidents. Our model captures the dependency between frequency and severity of cyber losses, as well as the behavior of cyber attacks on entities across time. Types of breach and organization and spatial location characteristics in chronology are taken into consider-

Les incidents de fuites de données entraînent de graves pertes financières et des dommages à la réputation, ce qui accroît l'importance de l'assurance pour gérer et mitiger les risques liés au cyberespace. Nous analysons une chronologie de fuites de données compilées par la Privacy Rights Clearinghouse (PRC) depuis 2005 et proposons un modèle mixte linéaire généralisé pour les incidents de fuites de données. Notre modèle rend compte de la dépendance entre la fréquence et la gravité des pertes liées au cyberespace et du comportement des attaques sur des entités qui s'y produisent avec le temps. Pour étudier les fréquences des fuites, nous prenons en compte les types de fuites et les caractéristiques

ation when investigating breach frequencies. Estimation of model parameters are presented under Bayesian framework using a combination of Gibbs sampler and Metropolis-Hastings algorithm. Predictions and applications of the proposed model in cyber insurance are discussed.

des organisations et des emplacements physiques dans la chronologie. L'estimation des paramètres du modèle est présentée dans un cadre bayésien combinant l'échantillonneur de Gibbs et l'algorithme de Metropolis-Hastings. Des prédictions et applications du modèle proposé en matière de cyberassurance sont aussi discutées.

[16:30-16:45]

Christopher Blier-Wong (Université Laval) **Hélène Cossette** (Université Laval) **Marceau Etienne** (Université Laval)

Risk Aggregation with FGM Copulas

Agrégation des risques avec copules FGM

This talk presents new results on risk aggregation when the dependence structure is a Farlie-Gumbel-Morgenstern (FGM) copula. Leveraging a new stochastic representation of the FGM copula, we provide expressions for the moments of aggregate random variables in terms of the order statistic moments of their marginals. When risks are mixed Erlang random variables, we show that the aggregate distribution is also mixed Erlang and develop convenient methods to compute the new parameters. We also develop allocation rules for conditional mean risk-sharing and Euler-based TVaR allocation. Finally, we present new results for the law of large numbers, the central limit theorem, a bound on the classical discrete-time ruin probability and large deviations for the FGM copula with the most positive dependence.

Cet exposé présente de nouveaux résultats sur l'agrégation des risques lorsque la structure de dépendance est une copule de Farlie-Gumbel-Morgenstern (FGM). En exploitant une nouvelle représentation stochastique de la copule FGM, nous fournissons des expressions pour les moments des variables aléatoires agrégées en termes des moments des statistiques d'ordre de leurs marginales. Lorsque les risques sont des mélanges de variables aléatoires Erlang, nous montrons que la distribution agrégée est également un mélange Erlang et développons des méthodes pratiques pour calculer les nouveaux paramètres. Nous développons également des règles d'allocation pour le partage du risque moyen conditionnel et l'allocation TVaR basée sur Euler. Enfin, nous présentons de nouveaux résultats pour la loi des grands nombres, le théorème central limite, une borne à la probabilité de ruine classique en temps discret et les grands écarts pour la copule FGM avec la dépendance la plus positive.

[16:45-17:00]

Pouya Faroughi (Western University) **Shu Li** (Western university) **Jiandong Ren** (Western university)

Generalized Poisson Distribution and Its Application in Actuarial Science

Distribution de Poisson généralisée et application en sciences actuarielles

In the paper, we introduce the Generalized Poisson (GP) and related distributions, and discuss their actuarial applications. After discussing some of its distributional properties, we present recursive methods for calculating the distribution function of the compound GP distribution, and then we explore computational methods for its risk measures (e.g., VaR and TVaR). Finally, we analyze a regression model for the GP distribution as well as its functional version, which we use to evaluate a real dataset for the number of claims in auto insurance. Illustrative comparisons with other common distributions for dealing with count data are also presented.

Dans cet article, nous présentons la distribution de Poisson généralisée (GP) et les distributions connexes et discutons de leurs applications actuarielles. Après avoir discuté de certaines propriétés de distribution, nous présentons des méthodes récursives pour calculer la fonction de distribution de la distribution GP composée, puis nous explorons des méthodes de calcul pour ses mesures de risque (par exemple, VaR et TVaR). Enfin, nous analysons un modèle de régression pour la distribution GP ainsi que sa version fonctionnelle, que nous utilisons pour évaluer un ensemble de données réelles pour le nombre de réclamations dans l'assurance automobile. Nous présentons également des comparaisons illustratives avec d'autres distributions courantes pour le traitement de données de dénombrement.

Chair/Président: Denis Talbot

Date: Monday May 30 / lundi 30 mai

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-15:45]

Shuo Sun (McGill University) **Johanna G. Nešlehová** (McGill University) **Erica E.M. Moodie** (McGill University)

Principal Stratification for Quantile Causal Effects under Partial Compliance: an Analysis of COVID-19 Case Counts

Stratification principale des effets causaux sur les quantiles en cas de conformité partielle : analyse du nombre de cas de COVID-19

Within the principal stratification framework in causal inference, the majority of the literature has focused on binary compliance with an intervention and modelling means. Yet in some research areas, compliance is partial, and research questions are concerned with causal effects on (possibly high) quantiles rather than on shifts in average outcomes. Modelling partial compliance is challenging because it can suffer from lack of identifiability. We develop an approach to estimate quantile causal effects within a principal stratification framework, where principal strata are defined by the bivariate vector of (partial) compliance to the two levels of a binary intervention. We propose a conditional copula approach to impute the missing potential compliance and estimate the principal quantile treatment effect surface at high quantiles, allowing the copula association parameter to vary with the covariates. A bootstrap procedure is used to estimate the parameter to account for inflation due to imputation of missing compliance. Moreover, we describe precise assumptions on which the proposed approach is based, and investigate the finite sample behaviour of our method by a simulation study. The proposed approach is used to study the 90th principal quantile treatment effect of executive stay-at-home orders on mitigating the risk of COVID-19 transmission in the United States.

Dans le cadre de la stratification principale de l'inférence causale, la littérature se concentre surtout sur la conformité binaire à une intervention et à modéliser les moyennes. Pourtant, dans certains domaines de recherche, la conformité est partielle et les recherches visent à déterminer les effets causaux sur les quantiles (éventuellement élevés) plutôt que sur les changements des résultats moyens. La modélisation de la conformité partielle représente un défi, car elle peut présenter un manque d'identifiabilité. Nous développons une approche pour estimer les effets causaux sur les quantiles dans un cadre de stratification principale, dans lequel les strates principales sont définies par le vecteur bivarié de la conformité (partielle) aux deux niveaux d'une intervention binaire. Nous proposons une approche de copule conditionnelle pour imputer la conformité manquante et estimer la surface quantile principale de l'effet de traitement à des quantiles élevés en permettant au paramètre d'association de la copule de varier avec les covariables. Nous utilisons une procédure bootstrap pour estimer le paramètre afin de tenir compte de l'inflation en raison de l'imputation de la conformité manquante. De plus, nous décrivons les hypothèses précises sur lesquelles l'approche proposée est basée, et nous étudions le comportement de notre méthode sur des échantillons de taille finie par une étude de simulation. Nous utilisons notre approche pour étudier l'effet de traitement au 90e quantile principal des mesures de confinement prises par les autorités pour réduire le risque de transmission de la COVID-19 aux États-Unis.

[15:45-16:00]

Yuliang Shi (University of Waterloo) **Yeying Zhu** (University of Waterloo) **Joel A. Dubin** (University of Waterloo)

Causal Inference on Missing Exposure via Triple Robust Estimation

Inférence causale avec données d'exposition manquante au moyen d'une estimation triplement robuste

How to deal with missing data in observational studies is a common problem for causal inference. However, few approaches exist if the data are missing at random

Les données manquantes dans les études d'observation posent souvent problème pour produire une inférence causale. Il existe cependant quelques approches lorsqu'il s'agit de données man-

(MAR) on the exposure of interest. The potential issue is that it causes extreme values on the estimated propensity scores and leads to biased estimates using imputation methods. In this article, a new method is provided called the weighted likelihood approach (WLA), which incorporates weights from both the missingness and treatment models to adjust for MAR and confounding issues. In addition, a new triple robust estimator is developed based on WLA, which only requires one of the treatment and outcome models to be correct even though the missingness model is misspecified. The simulation studies are conducted to compare WLA with multiple imputation and other missing data methods. Finally, an application is conducted to identify the causal effect of cardiovascular disease on the mortality of COVID-19.

quantes au hasard (MAR) relatives à l'exposition pertinente. Le problème potentiel, c'est qu'elles causent des valeurs extrêmes dans les scores de propension estimés et mènent à des estimations biaisées utilisant des méthodes d'imputation. Notre exposé présente une nouvelle méthode appelée l'approche de vraisemblance pondérée (WLA) qui incorpore les poids découlant à la fois de modèles de données manquantes et de traitement pour ajuster les données MAR et les problèmes de facteurs confondants. Basé sur l'approche WLA, un nouvel estimateur triplement robuste est développé qui requiert seulement qu'un des modèles de traitement et de résultats soit correct même si le modèle de données manquantes est mal spécifié. Des études en simulation visent à comparer l'approche WLA avec des méthodes d'imputation multiple et d'autres à données manquantes. Nous en faisons un application afin d'identifier l'effet causal de la maladie cardiovasculaire sur la mortalité liée à la COVID-19.

[16:00-16:15]

Jingyue Huang (University of Waterloo) **Leilei Zeng** (University of Waterloo) **Changbao Wu** (University of Waterloo)

Pseudo-Empirical Likelihood Approach for the Estimation of Average Treatment Effect

Approche de vraisemblance pseudo-empirique pour l'estimation de l'effet moyen du traitement

Propensity score (PS) based methods such as the inverse probability weighted (IPW) estimator or the doubly robust estimator which also involves the outcome regression models have been commonly used in practice for the estimation of the average treatment effect (ATE). Calibration methods have been shown to be effective in balancing the distributions of confounders explicitly among the treatment and control groups. We develop the pseudo-empirical likelihood (PEL) based methods for estimating the ATE with the model-calibration constraints constructed by the outcome regression models and establish the limiting distributions of the PEL ratio statistic. Our proposed point estimators are asymptotically equivalent to the IPW and the DR estimators, but the PEL ratio confidence intervals have several advantages over the traditional Wald-type intervals. Some preliminary results from simulation will be presented. This is a joint work with my PhD supervisors Dr. Leilei Zeng and Dr. Changbao Wu.

Les méthodes basées sur un score de propension (PS), comme l'estimateur de la pondération de probabilité inverse (IPW) ou l'estimateur doublement robuste (DR) qui fait aussi appel à des modèles de régression de l'issue, ont été couramment utilisées en pratique pour l'estimation de l'effet moyen du traitement (ATE). On a montré que les méthodes de calibrage sont efficaces pour équilibrer les distributions de facteurs confondants, explicitement parmi les groupes de traitement et témoins. Nous développons les méthodes basées sur la vraisemblance pseudo-empirique (PEL) pour l'estimation de l'ATE avec les contraintes de calibrage du modèle construites par les modèles de régression de l'issue et établissons les distributions limites des statistiques du rapport PEL. Les estimateurs ponctuels que nous proposons sont asymptotiquement équivalents avec les estimateurs IPW et DR, mais les intervalles de confiance du rapport PEL présentent plusieurs avantages par rapport aux intervalles de Wald traditionnels. Certains résultats préliminaires de simulation seront présentés. Ce travail est fait en collaboration avec mes directeurs de recherche au doctorat, le Dr Leilei Zeng et le Dr Changbao Wu.

[16:15-16:30]

Vanessa McNealis (McGill University) **Erica E.M. Moodie** (McGill University) **Nema Dean** (University of Glasgow)

Doubly Robust Estimation of Causal Effects in the Presence of Network Interference

Estimation doublement robuste d'effets causaux en présence d'interférence réseau

Causal inference on populations embedded in social networks poses technical challenges, since the typical no interference assumption may no longer hold. For in-

L'estimation d'effets causaux à partir de données de réseaux sociaux pose des défis méthodologiques, puisque les méthodes conventionnelles d'inférence causale présupposent l'absence d'in-

stance, in the context of infectious disease, the outcome of a study unit will likely be affected by the treatment of neighbours. While inverse probability weighted (IPW) estimators have been developed for this setting, they are often highly inefficient. In this work, we assume that the network is a union of connected subnetworks and propose doubly robust (DR) estimators combining models for treatment and outcome that are consistent and asymptotically normal if either model is correctly specified. We present empirical results that illustrate the DR property and the efficiency gain of DR over IPW estimators when both the outcome and the treatment are correctly modeled. Simulations are conducted under different scenarios of (latent) treatment dependence.

terférence entre les individus, une hypothèse qui ne tient peut-être plus. Par exemple, dans le contexte de maladies infectieuses, le risque d'infection d'un individu sera probablement affecté par le statut vaccinal de ses voisins. Bien que des estimateurs pondérés par l'inverse des probabilités ont été développées pour ce type de données, ils sont souvent inefficaces. Dans cette étude nous supposons qu'un réseau est l'union de sous-réseaux connexes et proposons des estimateurs doublement robustes combinant des modèles pour le traitement et la variable réponse, qui s'avèrent convergents et asymptotiquement normaux à condition que l'un des deux modèles soit correctement spécifié. Des résultats de simulations illustrent la propriété de double robustesse ainsi que l'efficacité supérieure de l'estimateur proposé lorsque les modèles de traitement et de réponse sont correctement spécifiés. Les simulations sont menées sous divers scénarios de dépendance (latente) des traitements.

[16:30-16:45]

Ian E. Waudby-Smith (Carnegie Mellon University) **David Arbour** (Adobe Research) **Ritwik Sinha** (Adobe Research) **Edward H. Kennedy** (Carnegie Mellon University) **Aaditya Ramdas** (Carnegie Mellon University)

Time-Uniform Central Limit Theory with Applications to Anytime-Valid Causal Inference

Théorie centrale limite uniforme dans le temps avec applications à l'inférence causale valide à tout moment

This work introduces time-uniform analogues of central limit theorem (CLT)-based confidence intervals. Our methods take the form of confidence sequences (CS) — sequences of confidence intervals that are uniformly valid over time. CSs provide valid inference at arbitrary stopping times, incurring no penalties for "peeking" at the data, unlike classical confidence intervals which require the sample size to be fixed in advance. Existing CSs in the literature are nonasymptotic, requiring strong assumptions on the data, while the classical (fixed-time) CLT is ubiquitous for the weak assumptions it imposes. Our work bridges the gap by introducing time-uniform CSs that only require CLT-like assumptions. We combine these with doubly robust estimators to derive nonparametric CSs for the average treatment effect (and other causal estimands). These allow randomized experiments and observational studies to be continuously monitored and adaptively stopped, all while controlling the type-I error.

Ce travail introduit des analogues uniformes dans le temps des intervalles de confiance basés sur le théorème central limite (TCL). Nos méthodes prennent la forme de séquences de confiance (SC) - des séquences d'intervalles de confiance qui sont uniformément valides dans le temps. Les SC permettent une inférence valide à des moments d'arrêt arbitraires, sans encourir de pénalités pour avoir « jeté un coup d'œil » aux données, contrairement aux intervalles de confiance classiques qui nécessitent que la taille de l'échantillon soit fixée à l'avance. Les SC existants dans la littérature sont non asymptotiques et nécessitent des hypothèses fortes sur les données, tandis que le TCL classique (à temps fixe) est universel pour les hypothèses faibles qu'il impose. Notre travail comble le fossé en introduisant des SC uniformes dans le temps qui ne nécessitent que des hypothèses de type TCL. Nous les combinons avec des estimateurs doublement robustes pour dériver des SC non paramétriques pour l'effet de traitement moyen (et autres paramètres causaux). Celles-ci permettent de surveiller en permanence les expériences randomisées et les études d'observation et de les arrêter de manière adaptative, tout en contrôlant l'erreur de type 1.

[16:45-17:00]

Zeyu Bian (McGill University) **Erica E.M. Moodie** (McGill University) **Susan Shortreed** (University of Washington) **Sahir R. Bhatnagar** (McGill University)

Variable Selection for Dynamic Treatment Regimes

Sélection de variables pour des régimes de traitement dynamique

With the aim of improving individual patients' health outcomes, dynamic treatment regimens (DTR) recommend effective treatments for individuals based on their characteristics. However, collected data often contain many irrelevant variables for tailoring treatment. Variable selection with the objective of optimizing patients' outcome by identifying useful tailoring variables is important. Many existing estimation methods are complicated and hard to implement, thus it is difficult to incorporate regularization within them. In this work, I show that with a suitable choice of weights, a weighted penalized regression model enjoys the desirable property of double robustness, and yet is straightforward to implement. The advantage of the newly proposed approach compared to alternative regularized DTR estimation methods lies in its simplicity of implementation using existing computationally efficient tools and the interpretability of the resulting DTRs.

Dans le but d'améliorer l'issue de santé d'un patient, un régime de traitement dynamique (DTR) préconise un traitement efficace basé sur ses caractéristiques individuelles. Cependant, les données collectées contiennent souvent de nombreuses variables non pertinentes pour concevoir un traitement adapté. Par conséquent, la sélection de variables est importante dans le but d'optimiser l'issue de santé du patient en identifiant des variables d'adaptation utiles. Comme bon nombre de méthodes d'estimation sont compliquées et difficiles à mettre en oeuvre, il devient ardu de leur incorporer de la régularisation. Dans cet exposé, je montre qu'avec un choix convenable de poids, un modèle de régression pénalisée pondérée possède la propriété désirable de double robustesse, tout en étant simple à mettre en oeuvre. Par comparaison avec d'autres méthodes d'estimation du DTR régularisées, l'avantage de cette approche nouvellement proposée réside dans la simplicité de sa mise en oeuvre à l'aide d'outils computationnels efficaces et de l'interprétabilité des régimes de traitement dynamique qui en résultent.

Chair/Président: Yanglei Song

Date: Monday May 30 / lundi 30 mai

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-15:45]

Alexander Shestopaloff (Memorial University of Newfoundland) **Radford M. Neal** (University of Toronto)

Bayesian Inference for Partially Observed Queueing Systems with Markov Chain Monte Carlo

Inférence bayésienne pour des systèmes de file d'attente partiellement observés avec la méthode Monte-Carlo par chaînes de Markov (MCMC)

Bayesian inference in partially observed queueing systems remains a challenging problem. We introduce an efficient MCMC sampling scheme to perform Bayesian inference in the M/G/1 queueing model given only observations of interdeparture times. Our MCMC scheme uses a combination of Gibbs sampling and simple Metropolis updates together with three novel "shift" and "scale" updates. We show that our novel updates improve the speed of sampling considerably, by factors of about 60 to about 180 on a variety of simulated data sets. We consider how the proposed MCMC method can be extended to perform Bayesian inference in more complex queueing systems.

L'inférence bayésienne pour des systèmes de file d'attente partiellement observés demeure tout un défi. Nous présentons un plan d'échantillonnage MCMC efficace pour faire une inférence bayésienne dans le modèle de file d'attente M/G/1 avec seulement des observations de temps entre les départs. Notre plan MCMC fait appel à une combinaison d'un échantillonnage de Gibbs et de simples mises à jour de Metropolis avec trois nouvelles mises à jour de « changement » et d'« échelle ». Nous montrons que nos nouvelles mises à jour améliorent considérablement la rapidité de l'échantillonnage par des facteurs d'environ 60 à environ 180 pour divers ensembles de données simulées. Nous tentons de voir comment la méthode MCMC proposée peut être étendue pour faire de l'inférence bayésienne pour des systèmes de file d'attente plus complexes.

[15:45-16:00]

Po Yang (Unviersity of Manitoba) **Shanika Basnayake** (University of Manitoba)

Bayesian Optimal Designs with High Prediction Efficiency

Plans optimaux bayésiens avec efficacité prédictive élevée

Design of experiments is a strategy used to identify the important factors which affect the response. A well-designed experiment plays a vital role in industry since it can provide information to conduct time- and cost-efficient process. For response surface experiments, the prediction of the response is an important task. We propose Bayesian optimality criteria for constructing optimal designs that have high prediction efficiency and less dependence on an assumed model. The constructed designs are compared with the designs obtained using different optimality criteria.

La planification d'expériences est une stratégie utilisée pour identifier les facteurs importants qui influent sur la réponse. Une expérience bien conçue joue un rôle essentiel dans l'industrie, car elle fournit l'information nécessaire pour exécuter un processus rentable en temps et en coût. Pour les expériences de surfaces de réponses, la prédiction de la réponse est une tâche importante. Nous proposons un critère d'optimalité bayésien pour la construction de plans optimaux à efficacité prédictive élevée et moins dépendants d'un modèle hypothétique. Les plans ainsi construits sont comparés à d'autres obtenus à l'aide de critères d'optimalité différents.

[16:00-16:15]

Sean Hellingman (Wilfrid Laurier University) **Zilin Wang** (Wilfrid Laurier University) **Mary E. Thompson** (University of Waterloo)

Markov Chain Models for Professional Soccer Tracking Data

Les modèles de chaîne de Markov pour les données de suivi du soccer professionnel

As soccer is widely regarded as the most popular sport in the world there is high interest in methods of improving team performances. To properly capture the dynamic actions of professional soccer, we propose Markov chains of greater complexity. These models allow for the inclusion of potential changes in the process caused by goals and substitutions, thus leading to a more complete and informative picture. Computer tracking data containing event descriptions and locations from La Liga games involving Futbol Club Barcelona have been used for model implementation and validation. Validations conducted through simulations show that the more complex models fit the data well. This better understanding should allow for more informed decisions surrounding substitutions in order to improve the probability of winning.

Le soccer est largement considéré comme le sport le plus populaire au monde et il existe un grand intérêt pour les méthodes d'amélioration des performances des équipes. Nous proposons des chaînes de Markov d'une plus grande complexité, pour bien saisir les actions dynamiques du soccer professionnel. Ces modèles permettent d'inclure les changements potentiels dans le processus causés par les buts et les substitutions, conduisant ainsi à une image plus complète et informative. L'estimation du modèle statistique et la validation du modèle ont été achevées grâce à l'utilisation de données de suivi informatique contenant des descriptions d'événements et des lieux des matchs de la Liga impliquant le Futbol Club Barcelona. Les validations effectuées par simulations montrent que les modèles les plus complexes s'ajustent bien aux données. Cette meilleure compréhension devrait permettre de prendre des décisions plus éclairées concernant les substitutions afin d'améliorer la probabilité de gagner.

[16:15-16:30]

Luke Hagar (University of Waterloo) **Nathaniel T. Stevens** (University of Waterloo)

A More Computationally Tractable Approach to Bayesian Interval-Based Sample Size Determination

Méthode de calcul simplifiée pour la détermination du nombre de sujets nécessaires en fonction d'un intervalle bayésien

Interval-based approaches to sample size determination aim to control the preciseness of posterior distributions used in Bayesian analyses. In particular, one may wish to control the length of highest density intervals (HDIs). The sufficient sample size distribution (SSSD) is defined to aid in this effort: when the sample size is taken to be a percentile p from this distribution, the HDI of the relevant posterior distribution will satisfy both the coverage and length criteria with probability p. Even though the SSSD is approximately normally distributed, existing methods for estimating the SSSD are inefficient in situations where conjugacy cannot be exploited. In this work, we propose using randomly generated idealized data, quasi-Markov chain Monte Carlo methods, Latin hypercube sampling, and maximum likelihood estimation to estimate the SSSD more efficiently. The proposed solution accurately estimates the SSSD and is an order of magnitude faster than existing estimation methods.

Les méthodes de détermination du nombre de sujets nécessaires en fonction d'un intervalle visent à contrôler la précision des distributions a posteriori utilisées dans les analyses bayésiennes. Plus particulièrement, on peut vouloir contrôler la longueur des intervalles de densité maximale. À cette fin, on définit la distribution de la taille suffisante du nombre de sujets nécessaires. Ainsi, lorsque la taille de l'échantillon est définie comme un percentile p de cette distribution, l'intervalle de densité maximale de la distribution a posteriori pertinente répond aux critères de couverture et de longueur avec une probabilité p. Même si la distribution de la taille suffisante du nombre de sujets nécessaires est approximativement distribuée normalement, les méthodes d'estimation actuelles de cette distribution sont inefficaces lorsque la conjugaison ne peut pas être exploitée. Dans le cadre de ces travaux, nous proposons d'utiliser des données idéalisées générées aléatoirement, des méthodes de quasi-Monte-Carlo par chaînes de Markov, un échantillonnage en hypercube latin et une estimation du maximum de vraisemblance pour estimer plus efficacement la distribution de la taille suffisante du nombre de sujets nécessaires. La solution que nous proposons permet d'estimer avec précision la distribution de la taille suffisante du nombre de sujets nécessaires et est un ordre de grandeur plus rapide que les méthodes d'estimation actuelles.

[16:30-16:45]

Xiaohua Liu (University of Manitoba) **Po Yang** (University of Manitoba)

A Bayesian Approach to Process Optimization On Data with Multi-Stratum Structure

Une approche bayésienne pour l'optimisation de processus pour des données avec structure à multistrates

Multi-stratum data arises naturally in industrial experiments due to the inconvenient and impractical complete randomization. The covariance matrix for such data is complicated because of the involvement of blocking factors and/or hard-to-change factors. Accounting for the uncertainty in the parameters of the model and in the forms of the model, we applied the Bayesian model averaging method and predictive approach to study the optimization problem for data with a multi-stratum structure. With the posterior probabilities of models as weights, we are not averaging the optimal levels of the controllable factors for each model but the predictive densities of the response over all the potential models. The goal of the optimization is to identify the values of the factors that result in the maximum probability of a response between a given region. This method is illustrated with two examples.

Les données multistrates surviennent naturellement dans des expériences industrielles à cause de la randomisation complète peu pratique. La matrice de covariance pour de telles données est complexe en raison de facteurs de blocage et de facteurs difficiles à changer. En tenant compte de l'incertitude dans les paramètres du modèle et de ses formes, nous avons appliqué la méthode de calcul de moyenne de modèle bayésien et l'approche prédictive dans le but d'étudier le problème d'optimisation de données avec structure à multistrates. En nous servant des probabilités postérieures des modèles en guise de poids, nous ne faisons pas la moyenne des niveaux optimaux des facteurs contrôlables pour chaque modèle, mais plutôt des densités prédictives de la réponse sur tous les modèles potentiels. L'objectif de l'optimisation est de repérer les valeurs des facteurs qui mènent à la probabilité maximale d'une réponse dans une région donnée. Nous illustrerons cette méthode avec deux exemples.

[16:45-17:00]

Hugh Chipman (Acadia University) **Derek Bingham** (Simon Fraser University)

Let's practice what we preach: Planning and interpreting simulation studies with design and analysis of experiments

Mettre en pratique ce que l'on prêche : planification et interprétation d'études de simulation avec conception et analyse d'expériences

Statisticians recommend the Design and Analysis of Experiments (DAE) for evidence-based research but often use tables to present their own simulation studies. Could DAE do better? We outline how DAE methods can be used to plan and analyze simulation studies. Tools for planning include fishbone diagrams, factorial and fractional factorial designs. Analysis is carried out via ANOVA, main-effect and interaction plots and other DAE tools. We also demonstrate how Taguchi Robust Parameter Design can be used to study the robustness of methods to a variety of uncontrollable population parameters.

Les statisticiens recommandent la conception et l'analyse d'expériences pour la recherche fondée sur des preuves, mais ils utilisent souvent des tableaux pour présenter leurs propres études de simulation. Les méthodes de conception et d'analyse d'expériences pourraient-elles faire mieux ? Nous décrivons la façon dont ces méthodes peuvent être utilisées pour planifier et analyser les études de simulation. Les outils de planification comprennent des diagrammes en arête de poisson, des plans factoriels et des plans factoriels fractionnaires. L'analyse est effectuée à l'aide de l'analyse de la variance, de diagrammes d'effets principaux et d'interactions, ainsi que d'autres outils de conception et d'analyse d'expériences. Nous montrons également la façon dont la conception robuste des paramètres de méthode de Taguchi permet d'analyser la robustesse des méthodes à plusieurs paramètres de population incontrôlables.

Recent Advances in Statistical Analysis of Event History Data
Progrès récents en analyse statistique des données d'historique des événements

Chair/Président: Leilei Zeng

Organizer/Responsable: Leilei Zeng

Date: Tuesday May 31 / mardi 31 mai

Time/Heure: 11:00-12:30

Abstract/Résumé

[11:00-11:30]

Hua Shen (University of Calgary)

Recurrent Event Analysis with Misclassified Covariate

Analyse d'événements récurrents avec covariable mal classée

Although we often encounter recurrent events data in public health study and medical research, standard methods for analyzing recurrent event data usually require the covariates to be completely and precisely observed though it is often not the case. Studies on the estimation of covariates effect on the risk of recurrent event in the presence of misclassified covariate remain sparse. We develop an expectation-maximization algorithm for fitting a semi-parametric regression model to recurrent event data with misclassified covariate in the absence of validation data. The likelihood-based algorithm is shown to yield estimators with small empirical bias in simulation studies.

Bien qu'on rencontre souvent des données d'événements récurrents dans les études de santé publique et la recherche médicale, les méthodes standard d'analyse de ces données exigent généralement que les covariables soient complètement et précisément observées, ce qui n'est souvent pas le cas. Il n'y a que peu d'études sur l'estimation de l'effet des covariables sur le risque d'événement récurrent en présence de covariables mal classées. Nous développons un algorithme de maximisation de l'espérance pour adapter un modèle de régression semi-paramétrique aux données d'événements récurrents avec des covariables mal classées en l'absence de données de validation. Nous montrons par des études de simulation que l'algorithme basé sur la vraisemblance produit des estimateurs avec un faible biais empirique.

[11:30-12:00]

Yan Yuan (University of Alberta) **Zhe Lu** (University of Alberta)

Age-Specific Risk Prediction, a Case Study of Early Menopause in Childhood Cancer Survivors

Prédiction du risque en fonction de l'âge : une étude de cas sur la ménopause précoce chez les survivants d'un cancer infantile

Age-specific disease risk prediction is often of clinical interest for counseling and follow-up management, especially for childhood cancer survivors (CCS) who have increased risk for many chronic conditions later in life due to the lifesaving but toxic cancer treatment at a young age. The classic model for time-to-event outcome is the Cox model. It models the hazard function, from which the estimated risk can then be derived. When the interest is age-specific disease risk, we propose to use inverse probability censoring weights to handle censoring in combination with modern binary classifiers for better prediction and calibration performance. In this talk, we present a case study for predicting early menopause risk in CCS as they aging from 21 to 40

La prédiction du risque de maladie en fonction de l'âge présente souvent un intérêt clinique pour les conseils et la gestion du suivi, surtout pour les enfants survivants d'un cancer, qui présentent un risque accru de souffrir de nombreuses maladies chroniques plus tard dans leur vie, en raison du traitement anticancéreux toxique, mais salvateur qu'ils ont reçu à un jeune âge. Le modèle classique utilisé pour modéliser le temps écoulé avant que ne survienne un événement est le modèle de Cox. Il modélise la fonction de risque, à partir de laquelle le risque estimé peut ensuite être obtenu. Lorsque la question porte sur le risque de maladie en fonction de l'âge, nous proposons d'utiliser des poids de censure à probabilité inverse pour gérer la censure en association avec des classificateurs binaires modernes pour obtenir de meilleurs résultats de prédiction et de calage. Dans cette présentation, nous

years. We compare the Cox model, the elastic-net penalized logistic regression, and the XGboost algorithm using a nested cross-validation framework. Model performance was assessed with scaled Brier score, AUC, AP and Spiegelhalter-z.

présentons une étude de cas pour prédire le risque de ménopause précoce chez les survivants d'un cancer infantile au cours de leur vieillissement (de 21 à 40 ans). Nous comparons le modèle de Cox, la régression logistique pénalisée par filet élastique et l'algorithme XGboost dans un cadre de validation croisée imbriquée. Nous évaluons l'efficacité du modèle à l'aide du score de Brier pondéré, de l'aire sous la courbe (AUC), de la précision moyenne et du test z de Spiegelhalter.

[12:00-12:30]

Liqun Diao (University of Waterloo) **Richard J. Cook** (University of Waterloo) **Ce Yang** (Harvard University)

Survival Trees for Current Status Data

Arbres de survie pour des données d'état actuel

Current status data arise when the exact time of an event of interest is not known and the only available information about the time is whether the time is beyond a single assessment. When interest lies in prediction based on such data, we define observed data loss functions through censoring unbiased transformations and pseudo-observations to construct unbiased estimates of complete data loss functions, and we use these to fit regression trees and make predictions using current status data. The trees grown based on these methods are found to have good properties empirically in terms of recovery of the true tree structure and event time prediction.

On parle de données d'état actuel lorsque le moment exact d'un événement d'intérêt n'est pas connu et que la seule information disponible sur le moment est de savoir si le moment est au-delà d'une seule évaluation. Lorsque l'intérêt réside dans la prédiction basée sur de telles données, nous définissons des fonctions de perte des données observées en censurant des transformations non biaisées et des pseudo-observations pour construire des estimations non biaisées des fonctions de perte des données complètes, et nous les utilisons pour ajuster des arbres de régression et faire des prédictions en utilisant les données d'état actuel. Il s'avère que les arbres construits sur la base de ces méthodes ont de bonnes propriétés empiriques en termes de récupération de la véritable structure de l'arbre et de prédiction du moment de l'événement.

Ensemble Learning via Diverse and Random Projections of Features
Apprentissage d'ensemble via des projections diverses et aléatoires de caractéristiques

Chair/Président: William J. Welch

Organizer/Responsable: Jabed Tomal

Date: Tuesday May 31 / mardi 31 mai

Time/Heure: 11:00-12:30

Abstract/Résumé

[11:00-11:30]

S. Ejaz Ahmed (Brock University)

Post Shrinkage Strategy in High Dimensional Data Analysis

Stratégie post-rétrécissement dans l'analyse de données de grande dimension

In high-dimensional settings where number of variables is greater than observations, or when number of variables are increasing with the sample size, many penalized strategies were studied for simultaneous variable selection and estimation. Penalty estimation strategy yields good results when the model at is assumed to be sparse. However, in a real scenario a model may include both sparse signals and weak signals. In this setting variable selection methods may not distinguish predictors with weak signals and sparse signals and will treat weak signals as sparse signals. The prediction based on a selected submodel may not be preferable. We suggest a high-dimensional shrinkage estimation strategy to improve the prediction performance of a submodel. We demonstrate that the proposed strategy performs uniformly better than the full estimator. Interestingly, it improves the prediction performance of the selected submodel. The relative performance is appraised by both simulation studies and the real data analysis.

Dans le contexte de problèmes en haute dimension où le nombre de variables est supérieur au nombre d'observations, ou lorsque le nombre de variables croît avec la taille de l'échantillon, plusieurs stratégies de pénalisation ont été étudiées pour l'estimation et la sélection simultanées des variables. Les stratégies d'estimation penalisées donnent de bons resultats quand il est supposé que le modèle est parcimonieux. Cependant, dans le cadre d'une véritable expérience, un modèle peut comprendre à la fois des signaux faibles et clairsemés. Dans ce contexte, il est possible que les méthodes de selection des variables ne puissent distinguer les prédicteurs des signaux faibles de ceux provenant de signaux clairsemés et traitent les signaux faibles comme des signaux clairsemés. La prediction basée sur un sous-modèle sélectionné pourrait ne pas convenir. Nous suggérons une stratégie d'estimation de retrécissement pour données de haute dimension afin d'améliorer les performances de prediction d'un sous-modèle. Nous démontrons que la stratégie proposée fonctionne uniformément mieux que l'estimateur complet. Fait intéressant, cela améliore la performance du sous-model selectionné. La performance relative des estimateurs est analysée de par des études de simulation de même qu'en faisant usage de données réelles.

[11:30-12:00]

Timothy I. Cannings (University of Edinburgh) **Richard J. Samworth** (University of Cambridge)

Random-projection ensemble classification

Classification d'ensembles de projections aléatoires

Random-projection ensembles offer general approaches for high-dimensional statistical problems. I will first present our proposed framework for classification problems, based on careful combination of the results of applying an arbitrary base classifier to random projections of the feature vectors into a lower-dimensional space. The random projections are divided into non-

Les ensembles de projections aléatoires offrent des approches générales pour les problèmes statistiques de grande dimension. Je présenterai tout d'abord le cadre que nous proposons pour les problèmes de classification, qui repose sur une combinaison minutieuse des résultats de l'application d'un classificateur de base arbitraire à des projections aléatoires des vecteurs caractéristiques dans un espace de dimension inférieure. Nous divisons les pro-

overlapping blocks, and within each block we select the projection yielding the smallest estimate of the test error. Our random projection ensemble classifier then aggregates the results of applying the base classifier on the selected projections, with a data-driven voting threshold to determine the final assignment. I will also present some preliminary results of ongoing work which aims to adapt our general framework for the problem of sufficient dimension reduction.

jections aléatoires en blocs non superposés, et dans chaque bloc nous sélectionnons la projection qui donne la plus petite estimation de l'erreur de test. Notre classificateur d'ensemble de projections aléatoires regroupe ensuite les résultats de l'application du classificateur de base sur les projections sélectionnées, avec un seuil de vote basé sur les données pour déterminer la répartition finale. Je présenterai également quelques résultats préliminaires d'une étude en cours qui vise à adapter notre cadre général au problème de la réduction de dimension suffisante.

[12:00-12:30]

Jabed Tomal (Thompson Rivers University) **William J. Welch** (University of British Columbia) **Ruben H. Zamar** (University of British Columbia)

Robust Ranking by Ensembling of Diverse Models and Assessment Metrics

Classement robuste par ensemble de divers modèles et métriques d'évaluation

We propose an ensemble of classification models formed using different assessment metrics. For a given metric, a classifier performs feature selection and combines models based on different subsets of feature variables which we call phalanxes. This first step, which employs the algorithm of phalanx formation, identifies strong and diverse subsets of feature variables. A second phase of ensembling aggregates classifiers across diverse assessment metrics. The proposed method is applied to protein homology data to mine homologous proteins, where the feature variables used for developing classifiers are various measures of similarity scores of proteins, and found robust for ranking both evolutionary close and distant homologous proteins.

Nous proposons un ensemble de modèles de classification formé à partir de différentes métriques d'évaluation. Pour une métrique donnée, un classificateur effectue une sélection de caractéristiques et combine les modèles selon différents sous-ensembles de variables de caractéristique que l'on appelle phalanges. Premièrement, au moyen de l'algorithme de formation de phalange, on repère les sous-ensembles robustes et diversifiés des variables de caractéristique. Deuxièmement, on assemble les classificateurs agrégés à travers diverses métriques d'évaluation. La méthode proposée est appliquée à des données d'homologie de protéine pour extraire les protéines homologues, où les variables de caractéristiques adoptées pour le développement des classificateurs sont diverses mesures de scores de similitudes de protéines. Cette méthode s'avère robuste dans le classement de protéines homologues évolutives rapprochées ou éloignées.

Chair/Président: Silvana Manuela Pesenti

Organizer/Responsable: Silvana Manuela Pesenti

Date: Tuesday May 31 / mardi 31 mai

Time/Heure: 11:00-12:30

Abstract/Résumé

[11:00-11:30]

Mélina Mailhot (Concordia University) **Fatima Palacios Rodriguez** (Universidad de Sevilla) **Elena Di Bernardino** (Université Côte D'Azur)

Smooth Copula-based Generalized Extreme Value model

Modèle lisse d'extrémum généralisé à l'aide de copules

A smooth copula-based Generalized Extreme Value (GEV) model will be presented, using a two-steps approach combining GEV parameters' smooth functions in space through the use of spatial covariates and a flexible hierarchical copula-based model to take into account dependency between the locations. The hierarchical copula structure is detected via a clustering algorithm implemented with an adapted version of the copula-based dissimilarity measure recently introduced in the literature. The considered data contains a large portion of missing values, and one observes several non-concomitant record periods at different locations. We compare the classical GEV parameter interpolation approaches with the proposed smooth copula-based GEV modeling approach, and apply the model to extreme rainfall in eastern Canada.

Nous présentons un modèle lisse d'extrémum généralisé à l'aide de copules avec une approche en deux étapes combinant les fonctions lisses des paramètres d'un extrémum généralisé dans l'espace par l'utilisation de covariables spatiales et d'un modèle hiérarchique souple à l'aide de copules pour tenir compte de la dépendance entre les emplacements. La structure de copule hiérarchique est détectée par un algorithme de groupement implémenté avec une version adaptée de la mesure de dissimilarité à l'aide de copules récemment présentée dans la littérature. Nous tenons compte de données qui contiennent un grand nombre de valeurs manquantes, et nous observons plusieurs périodes d'enregistrement non concomitantes à différents endroits. Enfin, nous comparons les approches classiques d'interpolation des paramètres d'extrémum généralisé à l'approche proposée de modélisation lisse d'extrémum généralisé basée sur une copule, et nous appliquons le modèle aux précipitations extrêmes dans l'est du Canada.

[11:30-12:00]

Menglin Zhou (The University of British Columbia) **Natalia Nolde** (University of British Columbia)

Reverse Stress Testing and Multivariate Extremes

Tests de résistance inversés et extrêmes multivariés

Reverse stress testing of a financial portfolio aims to identify scenarios for risk factors that lead to a specified adverse portfolio outcome, typically a large portfolio loss. The stress scenarios of interest naturally need to be probable yet extreme. In order to capture movements of risk factors that result in large portfolio losses, we propose a method to estimate stress scenarios using extrapolation based on techniques from multivariate extreme value theory. Such a method effectively addresses data scarcity in the joint tail regions while allowing for more flexible model assumptions focused on extremes.

Les tests de résistance inversés d'un portefeuille financier visent à déterminer des scénarios pour les facteurs de risque qui engendrent un résultat défavorable dans un portefeuille, c'est-à-dire une perte importante. Les scénarios de résistance en question doivent bien sûr être probables, mais aussi extrêmes. Afin de capturer les mouvements des facteurs de risque qui entraînent d'importantes pertes dans le portefeuille, nous proposons une méthode permettant d'estimer les scénarios de résistance par une extrapolation reposant sur des techniques issues de la théorie des valeurs extrêmes multivariées. Cette méthode résout efficacement le problème de rareté des données dans les régions de queues conjointes tout en permet-

We study the asymptotic behaviour of the proposed estimator, investigate its finite-sample performance in simulation studies and apply it to real data in a case study.

tant des hypothèses de modèle plus souples axées sur les extrêmes. Nous étudions le comportement asymptotique de l'estimateur proposé, nous examinons son efficacité pour des échantillons de taille finie dans des études de simulation, puis nous l'appliquons à des données réelles dans une étude de cas.

[12:00-12:30]

Mathieu Boudreault (Université du Québec à Montréal)

A Global Flood Risk Modeling Framework Built with Climate Models and Machine Learning

Cadre mondial de modélisation des risques d'inondation construit avec des modèles climatiques et l'apprentissage automatique

In this presentation, we introduce a data-driven, global, fast, flexible, and climate-consistent flood risk modeling framework for applications that do not necessarily require high-resolution flood mapping. We use statistical and machine learning methods to examine the relationship between historical flood occurrence and impact from the Dartmouth Flood Observatory (1985-2017), and climatic, watershed, and socioeconomic factors for 4734 watersheds globally. Using bias-corrected output from the NCAR CESM Large Ensemble (1980-2020), and the fitted statistical relationships, we simulate one million years of events worldwide along with the population displaced in each event. During the presentation, we discuss potential applications of the model, notably for the international (re)insurance industry, including global flood hazard and risk maps, the impacts of El Nino on flood risk and the contribution of climate (change) and urbanization to flood risk over the past 40 years.

Dans cette présentation, nous introduisons un cadre de modélisation des risques d'inondation axé sur les données, global, rapide, flexible et cohérent avec le climat pour des applications qui ne nécessitent pas nécessairement une cartographie des inondations à haute résolution. Nous utilisons des méthodes statistiques et d'apprentissage automatique pour examiner la relation entre l'occurrence et l'impact des inondations historiques provenant de l'Observatoire des inondations de Dartmouth (1985-2017), et les facteurs climatiques, hydrologiques et socio-économiques pour 4 734 bassins versants dans le monde. En utilisant les résultats corrigés du biais du NCAR CESM Large Ensemble (1980-2020), et les relations statistiques ajustées, nous simulons un million d'années d'événements dans le monde entier ainsi que la population déplacée dans chaque événement. Nous discutons des applications potentielles du modèle, notamment pour l'industrie internationale de la (ré)assurance : cartes mondiales des risques et des dangers d'inondation, impacts d'El Nino sur les risques d'inondation et contribution du changement climatique et de l'urbanisation aux risques d'inondation au cours des 40 dernières années.

Improving Robust High-dimensional Causal Inference and Prediction Modelling
Amélioration de l'inférence causale robuste à haute dimension et modélisation de la prédiction

Chair/Président: Celia M.T. Greenwood

Organizer/Responsable: Celia M.T. Greenwood, Gabriela Cohen Freue

Date: Tuesday May 31 / mardi 31 mai

Time/Heure: 11:00-12:30

Abstract/Résumé

[11:00-11:25]

Sahir R. Bhatnagar (McGill University)

Variable Selection in Parametric Hazard Models

Sélection de variables dans les modèles de risque paramétriques

The semiparametric Cox model has become the default approach to survival analysis, even though Cox himself later suggested he would prefer to model the hazard function directly to do things like predict the outcome for a particular patient. Methods relying on time matching or risk-set sampling require a separate estimation of the baseline hazard for survival or cumulative incidence curves. Extending these methods to more complex settings, such as penalized regression, require specialized implementations. In this talk, we first introduce case-base sampling; a parametric approach where hazard functions can be estimated in continuous-time using logistic regression. This approach naturally leads to estimates of the survival or risk functions that are smooth-in-time. We then show how case-base sampling can be used for variable selection through regularized estimation of the hazard function. We contrast our approach with Coxnet, which regularizes the Cox partial likelihood.

Le modèle semi-paramétrique de Cox est devenu l'approche par défaut de l'analyse de survie, même si Cox lui-même a suggéré plus tard qu'il préférait modéliser directement la fonction de risque dans certains cas, comme pour prédire le résultat pour un patient particulier. Les méthodes reposant sur l'appariement temporel ou l'échantillonnage par ensembles de risques nécessitent une estimation séparée du risque de base pour les courbes de survie ou d'incidence cumulative. L'extension de ces méthodes à des contextes plus complexes, comme la régression pénalisée, exige des implémentations spécialisées. Dans cet exposé, nous présentons d'abord l'échantillonnage « case-base », une approche paramétrique où les fonctions de risque peuvent être estimées en temps continu au moyen d'une régression logistique. Cette approche conduit naturellement à des estimations des fonctions de survie ou de risque qui sont lissées dans le temps. Nous montrons ensuite comment utiliser l'échantillonnage « case-base » pour la sélection de variables par estimation régularisée de la fonction de risque. Nous comparons notre approche à Coxnet, qui régularise la vraisemblance partielle de Cox.

[11:25-11:50]

Xinyi Zhang (University of Toronto)

Fighting Noise with Noise: Causal Inference with Many Candidate Instruments

Combattre le bruit par le bruit : inférence causale à l'aide de nombreux instruments candidats

Instrumental variable methods provide useful tools for inferring causal effects in the presence of unmeasured confounding. To apply these methods with large-scale data sets, a major challenge is to find valid instruments from a possibly large candidate set. In practice, most of the candidate instruments are often not relevant for studying a particular exposure of interest. Moreover, not all relevant candidate instruments are valid as they may directly influence the outcome of interest. In this

Les méthodes des variables instrumentales constituent des outils utiles pour déduire les effets causaux en présence de facteurs de confusion non mesurés. Pour appliquer ces méthodes à des ensembles de données à grande échelle, l'un des principaux défis consiste à trouver des instruments valides à partir d'un vaste ensemble de candidats. Dans la pratique, la plupart des instruments candidats ne sont souvent pas adaptés à l'étude d'une exposition présentant un intérêt particulier. De plus, tous les instruments candidats adaptés ne sont pas valides, car ils peuvent influencer

article, we propose a data-driven method for causal inference with many candidate instruments that addresses these two challenges simultaneously. A key component of our proposal is a novel resampling method that constructs pseudo variables to identify and remove irrelevant candidate instruments having spurious correlations with the exposure. Theoretical and synthetic data analyses show that the proposed method performs favourably compared to existing methods. We apply our method to a Mendelian randomization study estimating the effect of obesity on health-related quality of life.

directement le résultat recherché. Dans cette présentation, nous proposons une méthode axée sur les données pour déterminer l'inférence causale à l'aide de nombreux instruments candidats, qui traite ces deux problèmes simultanément. Nous nous appuyons sur une nouvelle méthode de rééchantillonnage qui permet de créer des pseudo-variables afin de repérer et d'éliminer les instruments candidats non adaptés présentant des corrélations parasites avec l'exposition. Grâce à des analyses théoriques et de données synthétiques, nous constatons que notre méthode donne de bons résultats par rapport aux méthodes actuelles. Nous appliquons notre méthode à une étude de randomisation mendélienne estimant l'effet de l'obésité sur la qualité de vie liée à la santé.

[11:50-12:15]

Eric Tchetgen Tchetgen (Uniiversity of Pennsylvania)

Doubly Robust Calibration of Prediction Sets under Covariate Shift

Calibration doublement robuste d'ensembles de prédiction selon un décalage de covariables

Conformal prediction has received tremendous attention in recent years and has offered new solutions to problems in missing data and causal inference; yet these advances have not leveraged modern semiparametric efficiency theory for more robust and efficient uncertainty quantification. In this paper, we consider the problem of obtaining distribution-free prediction regions accounting for a shift in the distribution of the covariates between the training and test data. Under an explainable covariate shift assumption analogous to the standard missing at random assumption, we propose three variants of a general framework to construct well-calibrated prediction regions for the unobserved outcome in the test sample. Our approach is based on the efficient influence function for the quantile of the unobserved outcome in the test population combined with an arbitrary machine learning prediction algorithm, without compromising asymptotic coverage. Next, we extend our approach to account for departure from the explainable covariate shift assumption in a semiparametric sensitivity analysis for potential latent covariate shift. In all cases, we establish that the resulting prediction sets eventually attain nominal average coverage in large samples. This guarantee is a consequence of the product bias form of our proposal which implies correct coverage if either the propensity score or the conditional distribution of the response is estimated sufficiently well. Our results also provide a framework for construction of doubly robust prediction sets of individual treatment effects, under unconfoundedness conditions as well as allowing for some degree of unmeasured confounding.

La prédiction conforme a suscité beaucoup d'intérêt dans les dernières années et a offert de nouvelles solutions à des problèmes de données manquantes et d'inférence causale. Cependant, ces avancées n'ont pas exploité la théorie d'efficacité semi-paramétrique moderne pour rendre plus efficace et robuste la quantification de l'incertitude . Dans cet article, nous abordons le problème d'obtention de régions de prévision sans distribution, en tenant compte d'un décalage dans la distribution des covariables entre les données d'apprentissage et de test. Selon une hypothèse de décalage de covariables explicable analogue à l'hypothèse des données manquantes aléatoirement, nous proposons trois variants d'un cadre général afin de construire des régions de prévision pour le résultat non observé dans l'échantillon test. Notre approche est fondée sur la fonction d'influence efficace pour le quantile du résultat non observé dans la population test combinée à un algorithme de prédiction par apprentissage automatique arbitraire, sans compromettre la couverture asymptotique. Ensuite, nous élargissons notre approche afin de s'écarter de l'hypothèse de décalage de covariables explicable dans une analyse de sensibilité semi-paramétrique pour la possibilité de décalage de covariables latentes. Dans tous les cas, nous établissons que les ensembles de prévisions obtenus atteignent à la longue une couverture moyenne nominale dans de grands échantillons. Cette garantie est causée par le biais des produits de notre méthode qui supposent une couverture adéquate si le score de propension ou la distribution conditionnelle de la réponse sont suffisamment bien estimés. Nos résultats procurent aussi un cadre pour la construction d'ensembles de prévisions doublement robustes des effets thérapeutiques individuels, selon des conditions d'ignorabilité et en permettant un certain degré de variables confondantes non mesurées. Enfin, nous abordons l'agrégation des ensembles de

Finally, we discuss aggregation of prediction sets from different machine learning algorithms for optimal prediction and illustrate the performance of our methods in both synthetic and real data.

prévisions à partir de différents algorithmes d'apprentissage automatique pour prédire de façon optimale et illustrer la performance de notre méthode avec des données réelles et synthétiques.

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

Date: Tuesday May 31 / mardi 31 mai

Time/Heure: 11:00-12:30

Abstract/Résumé

[11:00-12:00]

Mark S Handcock (University of California, Los Angeles) **Ian E. Fellows Krista J. Gile Henry F. Raymond**

Sampling Hard-to-Reach Populations

Échantillonnage de populations difficiles à rejoindre

In many situations, standard survey sampling strategies fail because the target populations cannot be accessed through well-defined sampling frames. Typically, a sampling frame for the target population is not available, and its members are rare or stigmatized in the larger population so that it is prohibitively expensive to contact them through the available frames. We discuss statistical issues in studying hard-to-reach or otherwise "hidden" populations. These populations are characterized by the difficulty in survey sampling from them using standard probability methods. Examples in a demographic setting include unregulated workers and migrants. Examples of such populations in a behavioral and social setting include injection drug users, men who have sex with men, and female sex workers. Hard-to-reach populations are under-served by current sampling methodologies mainly due to the lack of practical alternatives to address these methodological difficulties. We will focus on populations where some form of social network information can be used to assist the data collection. In such situations sophisticated statistical methods are needed to allow the characteristics of the population to be inferred from the collected data. We review time-location sampling, adaptive network sampling, including respondent-driven sampling, as well as indirect and meta-mathods. We also discuss model-assisted methods and and capture-recapture ideas. This is joint work with Ian E. Fellows, Krista J. Gile, and Henry F. Raymond.

Dans plusieurs situations, les stratégies d'échantillonnage par enquête habituelles ne suffisent pas, car les populations cibles ne peuvent pas être rejointes par l'entremise de bases de sondage bien définies. Généralement, une base sondage pour la population cible n'est pas disponible et ses membres sont rares ou stigmatisés dans l'ensemble de la population. Il devient donc beaucoup trop coûteux de les contacter en utilisant les bases disponibles. Nous discutons des problèmes statistiques concernant l'étude de populations difficiles à rejoindre ou ≪cachées≫. Il est ardu d'échantillonner ces populations par enquête au moyen de méthodes probabilistes classiques. Dans un cadre démographique, cela comprend par exemple les travailleurs non réglementés et les migrants. Des exemples de telles populations dans un cadre social et comportemental incluent les utilisateurs de drogues injectables, les hommes homosexuels et les travailleuses du sexe. Les populations difficiles à rejoindre sont mal desservies par les méthodologies d'échantillonnage en raison principalement du manque d'autres moyens pratiques pour résoudre ces difficultés méthodologiques. Nous nous concentrerons sur des populations ayant une forme de réseau social dont l'on pourra tirer de l'information pour soutenir la collecte de données. Dans de telles situations, des méthodes statistiques sophistiquées sont requises pour permettre l'inférence des caractéristiques de la population à partir des données recueillies. Nous passons en revue l'échantillonnage lieux-moments, l'échantillonnage adaptatif par réseau, y compris l'échantillonnage fondé sur les répondants, et les méta-analyses indirectes. Nous abordons aussi les méthodes assistées par un modèle et certaines idées de capture-recapture. Il s'agit d'un travail conjoint avec E. Fellows, Krista J. Gile et Henry F. Raymond.

Chair/Président: Alessandro Maria Maria Selvitella

Date: Tuesday May 31 / mardi 31 mai

Time/Heure: 11:00-12:30

Abstract/Résumé

[11:00-11:15]

Gabriel Oppong Afriyie (University of Calgary) **Meng Wang** (University of Calgary, Canada) **Na Li** (University of Calgary, Canada) **Chel Hee Lee** (University of Calgary, Canada) **Alberto Nettel Aguirre** (University of Wollongong, Australia) **Anita Brobbey** (University of Calgary, Canada) **David Hughes** (University of Liverpool, United Kingdom) **Tolulope Sajobi** (University of Calgary, Canada)

Longitudinal Discriminant Aanalysis for Dementia Risk Prediction

Analyse discriminante longitudinale pour prédire le risque de démence

Discriminant analysis models based on multivariate mixed effects (DA-MM) and multivariate generalized estimating equations (DA-GEE) models have been developed for classification in multivariate longitudinal data. This study compares the accuracy of these models for predicting dementia in a longitudinal registry of patients with mild cognitive impairment followed over a 3-year period and by computer simulations. The overall accuracy of DA-MM and DA-GEE classifiers in predicting 3-year dementia risk were 82.8% and 80.6%, respectively. More detailed results from the simulation study that examined the performance of these models under a variety of data analytic conditions will be also discussed in this talk.

Des modèles d'analyse discriminante basés sur des effets mixtes multivariés (DA-MM) et des équations d'estimation généralisée multivariée (DA-GEE) ont été conçus à des fins de classification pour des données longitudinales multivariées. Cette étude compare la précision de ces modèles relative à la prédiction de la démence à partir d'un registre longitudinal de patients souffrant de déficit cognitif modéré sur une période de trois ans et au moyen de simulations informatiques. La précision globale des classificateurs DA-MM et DA-GEE pour prédire le risque de démence sur une période de trois ans était de 82,8 % et 80,6 % respectivement. Lors de cet exposé, nous aborderons en détail les résultats tirés de l'étude en simulation sur la performance de ces modèles selon plusieurs conditions d'analyse de données.

[11:15-11:30]

Gansen Deng (Western University) **Ryan Koh** (Toronto Rehabilitation Institute) **Wenqing He** (Western University) **Samah Hassan** (Toronto Rehabilitation Institute) **Shoba Subramaniam** (Toronto Rehabilitation Institute) **Dinesh Kumbhare** (Toronto Rehabilitation Institute)

Self-reported Data Analysis of a Chronic Pain Study

Analyses de données autodéclarées d'une étude sur la douleur chronique

Chronic Pain (CP) is a complicated condition and is highly variable. A mechanism-based classification has been proposed to group CP patients into neuropathic, nociceptive or nocioplastic classes. However, little is known about the class profiles. The current study aimed to explore the ability to use the routinely collected data to classify patients into the desired classes. Both the Latent Class Analysis (LCA) and the Hierarchical Clustering (HC) methods were applied to partition the patients, and multiple supervised learning methods were invoked to predict the clusters. The occurrence frequency and

La douleur chronique (DC) est une condition complexe et très variable. Une classification basée sur des mécanismes a été proposée pour grouper les patients souffrant de DC dans les classes neuropathique, nociceptive ou nociplastique. Cependant, on en sait peu sur les profils de classe. La présente étude cherche à trouver un moyen d'utiliser les données recueillies couramment pour grouper les patients dans les classes pertinentes. Nous appliquons l'analyse de classe latente (ACL) et le regroupement hiérarchique (RH) pour diviser les patients, et avons recours à de multiples méthodes d'apprentissage supervisées pour prédire les groupes. Nous nous servons de la fréquence d'occurrence et la distance

correlation distance were used in HC to account for the nature of self-reported data. Network analysis was conducted to investigate the correlations between variables. It is shown that the self-reported data were insufficient in predicting the mechanism-based classes. Additional domain and temporal info may be needed in the class profiles.

de corrélation dans le RH pour tenir compte de la nature des données autodéclarées. Une analyse de réseau a aussi été menée pour étudier la corrélation entre les variables. Les données autodéclarées se sont avérées insuffisantes pour prédire les classes basées sur les mécanismes. Des informations supplémentaires se rapportant sur le domaine et de nature temporelle pourraient être requises pour les profils de classe.

[11:30-11:45]

Antonio Peruzzi (Ca' Foscari University of Venice) **Roberto Casarin** (Ca' Foscari University of Venice)

Media Bias and Polarization via a Markov-Switching Latent Space Model: an Application to the Media Environment of France, Germany, and Italy.

Les biais et la polarisation médiatiques à l'aide d'un modèle latent à changement d'espaces de Markov : application à l'environnement médiatique en France, Allemagne et Italie

The news consumption landscape has drastically changed in the last decades and several old issues return to the fore. One of these is whether and to which extent news outlets bias information. We propose a new dynamic latent-space model (LS) for news outlets in which we exploit both time-varying online audience duplication-network data as well as textual contents from published articles to measure media bias over time. Our model, estimated within the Bayesian framework, recovers the latent coordinates of news outlets in a 2-dimensional euclidean space, while providing a proper interpretation respectively in terms of media slant and online engagement. The aim is twofold: making advancements both concerning the analysis of the timely evolution of audience duplication networks and concerning the determination of media slant and polarization. The developed model is applied to a Facebook dataset regarding news outlets from France, Germany and Italy.

La consommation des nouvelles a nettement changé au cours des dernières décennies et de nombreux problèmes anciens refont surface, notamment si et dans quelle mesure les sources de nouvelles biaisent l'information. Nous proposons un nouveau modèle dynamique d'espaces latents (LS) pour les sources de nouvelles, lequel exploite les données de réseaux de duplication d'audience en ligne à temps variés de même que les contenus textuels d'articles publiés pour mesurer le biais médiatique au fil du temps. Estimé dans le cadre bayésien, notre modèle reconstitue les coordonnées latentes de sources de nouvelles dans un espace euclidien à deux dimensions, tout en fournissant une interprétation adéquate respectivement du biais médiatique et de l'engagement en ligne. Le but est de faire des progrès à deux égards : l'analyse de l'évolution opportune des réseaux de duplication d'audience et la détermination du biais et de la polarisation médiatiques. Le modèle développé est appliqué à un ensemble de données Facebook au sujet des sources de nouvelles de France, d'Allemagne et d'Italie.

[11:45-12:00]

Amin Kharaghani (University of Toronto: Dalla Lana School of Public Health) **Milos Milic** (Krembil Centre for Neuroinformatics, Centre for Addiction and Mental Health) **Earvin Tio** (Krembil Centre for Neuroinformatics, Centre for Addiction and Mental Health) **David A. Bennett** (Rush University Medical Center) **Philip L. De Jager** (Columbia University Medical Center) **Julie A. Schneider** (Rush University Medical Center) **Lei Sun** (University of Toronto) **Daniel Felsky** (Krembil Centre for Neuroinformatics, Centre for Addiction and Mental Health; University of Toronto)

Association of Novel Whole-person Eigen-polygenic Scores with Alzheimer's disease

Association à la maladie d'Alzheimer de nouveaux scores polygéniques eigen de soins intégraux de la personne

Late-Onset Alzheimer's Disease (LOAD) is a heterogeneous disorder with complex etiology and high, largely unexplained, heritability. We used Pan-UK Biobank Consortium GWAS summary statistics to calculate polygenic scores (PGS) for 2,312 heritable traits in a sample of 2,044 elderly with clinical and post-mortem autopsy data. Weighted gene co-expression network

La maladie d'Alzheimer d'apparition tardive (LOAD) est un trouble hétérogène avec une étiologie complexe et une héritabilité élevée, encore largement inexpliquée. Nous avons utilisé des statistiques sommaires d'une étude d'association pangénomique (GWAS) d'un consortium de la Pan-UK Biobank pour calculer chez 2 044 personnes âgées les scores polygéniques (PGS) de 2 312 traits héritables avec un échantillonnage de données d'autop-

analysis (WGCNA) was used to identify discrete correlated clusters of PGS. Eigen-PGS (ePGS) were calculated as the first principal component of PGS within each cluster. We identified between 11 to 35 clusters (modules) depending on PGS p-value inclusion thresholds. Multiple ePGS, composed primarily of PGS related to vascular health, were associated with a diagnosis of LOAD (p=4x10-12) and both β-amyloid (Aβ) (p=1.2x10-26) and tau neuropathology (p=3.1x10-21). The addition of ePGS to gold standard LOAD PGS models significantly improved performance (for Aβ deposition; delta-R2=2.4%, LRT p=1.69×10^{-9}). This novel application of WGCNA offers improvements over existing single-PGS approaches and will aid in the generation of new etiological hypotheses of LOAD.

sies cliniques et post-mortem. Une analyse de réseau de coexpression de gènes pondérée (WGCNA) a été utilisée pour identifier des groupes corrélés discrets de PGS. Les Eigen-PGS (ePGS) ont été calculés comme première composante principale des PGS à l'intérieur de chaque groupe. Nous avons identifié entre 11 et 35 groupes (modules) dépendamment des seuils d'inclusion de la valeur p des PGS. De multiples ePGS composés surtout de PGS liés à la santé vasculaire ont été associés avec un diagnostic de maladie d'Alzheimer d'apparition tardive (p = 4 x10-12) et à une neuropathologie à la fois avec peptide β-amyloïde (Aβ) (p = 1,2 x10-26) et tau (p = 3,1 x 10-21). L'ajout d'ePGS aux modèles des scores polygéniques étalon-or de la LOAD a sensiblement amélioré la performance (pour un dépôt Aβ : delta-R2 = 2,4 %, LRT p = $1,69 \times 10^{-9}$). En plus d'apporter des améliorations aux approches avec un seul PGS, cette nouvelle application de la WGCNA contribuera à générer de nouvelles hypothèses étiologiques quant à la maladie d'Alzheimer d'apparition tardive.

[12:00-12:15]

Mohammad Kaviul Anam Khan (University of Toronto) **Rafal Kustra** (University of Toronto)

Understanding the Properties of Permutation Based Importance for Inputs of Black Box Machine Learning Methods

Comprendre les propriétés de l'importance basée sur des permutations pour les entrées des méthodes d'apprentissage automatique «boîte noire»

The goal of this study is to identify important predictors for black box machine learning methods, where the prediction function is highly non-linear and cannot be represented by statistical parameters. Thus such black-box models lack interpretability and it is very difficult to identify "important" or significant inputs for an outcome from such models. The main target is to investigate applicability of permutation based approach, proposed by Breiman (2001) and then modified by Fisher et.al (2019) to such non-linear non-additive methods. Another aim is to decompose the proposed variable importance metric (VIM) to obtain a causal parameter which is a function of the expected conditional average treatment effect over the distribution of treatments for multinomial and continuous treatments. A simulation study was then conducted to check the performance of the estimated VIM using split-sampling techniques using multiple known machine learning methods. The estimation technique of VIM was also evaluated under model misspecification.

L'objectif de cette étude est de repérer les prédicteurs importants pour les méthodes d'apprentissage automatique «boîte noire», lorsque la fonction de prédiction est hautement non linéaire et ne peut être représentée par des paramètres statistiques. Ces modèles «boîte noire» sont complexes à interpréter et il est très difficile de repérer les entrées pertinentes ou «importantes» pour un résultat tiré de ce genre de modèles. L'objectif principal est d'étudier l'applicabilité d'une approche de permutation, proposée par Breiman (2001) puis ensuite modifiée par Fisher et coll. (2019) pour ce genre de méthodes non additives et non linéaires. Un autre objectif est de décomposer la mesure d'importance de variables (VIM) proposée pour obtenir un paramètre causal qui est une fonction de l'effet de traitement moyen conditionnel espéré par rapport à la distribution de traitements pour les traitements continus et multinomiaux. Une étude de simulation a été menée afin de vérifier la performance de la VIM estimée grâce à des techniques d'échantillonnage divisé à partir de plusieurs méthodes d'apprentissage automatique connues. La technique d'estimation de la VIM a aussi été évaluée sous erreurs de spécification du modèle.

[12:15-12:30]

Mengying Lei (McGill University) **Aurélie Labbe** (HEC Montreal) **Lijun Sun** (McGill University)

Scalable Spatiotemporally Varying Coefficient Modelling with Bayesian Kernelized Tensor Regression

Modélisation des coefficients extensibles à variation spatio-temporelle avec régression tensorielle noyautée bayésienne

As a regression technique in spatial statistics, the spatiotemporally varying coefficient model (STVC) is an important tool for discovering nonstationary and interpretable response-covariate associations over both space and time. However, it is difficult to apply STVC for large-scale spatiotemporal analyses due to the high computational cost. To address this challenge, we summarize the spatiotemporally varying coefficients using a third-order tensor structure and propose to reformulate the spatiotemporally varying coefficient model as a special low-rank tensor regression problem. The low-rank decomposition can effectively model the global patterns of the large data sets with a substantially reduced number of parameters. To further incorporate the local spatiotemporal dependencies, we use Gaussian process (GP) priors on the spatial and temporal factor matrices. We refer to the overall framework as Bayesian Kernelized Tensor Regression (BKTR). For model inference, we develop an efficient Markov chain Monte Carlo (MCMC) algorithm, which uses Gibbs sampling to update factor matrices and slice sampling to update kernel hyperparameters. We conduct extensive experiments on both synthetic and real-world data sets, and our results confirm the superior performance and efficiency of BKTR for model estimation and parameter inference.

En tant que technique de régression en statistiques spatiales, le modèle de coefficients à variation spatio-temporelle (STVC) est un outil important dans la découverte d'associations de réponse-covariables non stationnaires et interprétables dans l'espace et dans le temps. Il est cependant difficile d'appliquer le STVC à des analyses spatio-temporelles à grande échelle en raison du coût de calcul élevé. Afin relever ce défi, nous résumons les coefficients à variation spatio-temporelle à l'aide d'une structure tensorielle de troisième ordre et proposons de reformuler ce modèle comme un problème de régression tensorielle à rang réduit. La décomposition à rang réduit peut modéliser efficacement les schémas globaux des grands ensembles de données avec un nombre de paramètres substantiellement réduit. Pour mieux intégrer les dépendances spatio-temporelles locales, nous utilisons une loi a priori gaussienne (GP) sur les matrices de facteurs spatiaux et temporels. Nous nous rapportons au cadre général en tant que régression tensorielle noyautée bayésienne (BKTR). Pour l'inférence de modèle, nous développons un algorithme efficace de Monte Carlo par chaîne de Markov (MCMC), qui se sert de l'échantillonnage Gibbs dans le but d'actualiser les matrices factorielles et de l'échantillonnage par tranches pour mettre à jour les hyperparamètres de noyau. Nous menons des expériences approfondies sur des ensembles de données synthétiques ainsi que réelles, dont les résultats confirment la performance supérieure et l'efficacité de la BKTR pour l'estimation des modèles et l'inférence des paramètres.

Chair/Président: Grace S. Chiu

Date: Tuesday May 31 / mardi 31 mai

Time/Heure: 11:00-12:30

Abstract/Résumé

[11:00-11:15]

Pankaj Uttam Bhagwat (University of Sherbrooke) **Eric Marchand** (University of Sherbrooke)

Bayesian Inference and Prediction for Mean-Mixtures of Normal Distributions

Inférence bayésienne et prédiction pour des mélanges de lois normales sur la moyenne

We study frequentist risk properties of predictive density estimators for mean mixtures of multivariate normal distributions, involving an unknown location parameter $\theta \in \mathbf{R}^d$, and which include multivariate skew normal distributions. We provide explicit representations for Bayesian posterior and predictive densities, including the benchmark minimum risk equivariant (MRE) density, which is minimax and generalized Bayes with respect to an improper uniform density for θ. For four dimensions or more, we obtain Bayesian densities that improve uniformly on the MRE density under Kullback-Leibler loss. We also provide plug-in type improvements, investigate implications for certain type of parametric restrictions on θ, and illustrate and comment the findings based on numerical evaluations.

Nous abordons l'étude de l'efficacité de densités prédictives, dans un cadre décisionnel avec la perte Kullback-Leibler, pour des mélanges de lois normales sur la moyenne avec un paramètre de position inconnu $\theta \in \mathbf{R}^d$. Nous mettons en évidence des représentations explicites pour des densités a posteriori et prédictives bayésiennes, y incluant la meilleure densité prédictive équivariante (MDPÉ) qui est aussi minimax et Bayes généralisée associée à une densité uniforme pour θ. Pour quatre dimensions ou plus, nous obtenons des densités bayésiennes et des densités par substitution dominant uniformément la MDPÉ sous la perte Kullback-Leibler, y compris certains cas où il existe une contrainte paramétrique sur θ. Nous fournissons également des améliorations de type plug-in, étudions les implications de certains types de restrictions paramétriques sur θ, et illustrons les résultats théoriques à l'aide d'évaluations numériques.

[11:15-11:30]

Ziming Chen (University of Toronto) **Jeffrey Berger** (New York University School of Medicine) **Lana Castellucci** (The Ottawa Hospital) **Michael Farkouh** (Toronto General Hospital, University Health Network, Toronto) **Ewan Goligher** (Toronto General Hospital, University Health Network, Toronto) **Beverley Hunt** (King's College, London) **Lucy Kornblith** (University of California San Francisco) **Patrick Lawler** (Peter Munk Cardiac Centre, University Health Network, Toronto) **Eric Leifer** (National Heart, Lung, and Blood Institute) **Matthew Neal** (University of Pittsburgh Medical Center) **Ryan Zarychanski** (University of Manitoba) **Anna Heath** (The Hospital for Sick Children, Toronto, University of Toronto, University College London)

A Comparison of Methods for Bayesian Inference in Clinical Trials

Comparaison de méthodes d'inférence bayésienne pour les essais cliniques

Bayesian analysis updates inference as more data becomes available. Typically, Bayesian inference uses simulation approaches such as Markov Chain Monte Carlo (MCMC) but an approximation approach, the Integrated Nested Laplace Approximation (INLA), is also available. Although the simulation-based methods are theoretically accurate, they can be computationally expensive. The goal of the study is to compare INLA

L'analyse bayésienne met à jour l'inférence au fur et à mesure que des données supplémentaires sont disponibles. Généralement, l'inférence bayésienne utilise des approches de simulation telles que la méthode de Monte-Carlo par chaîne de Markov (MCMC), mais il existe également une approche par approximation, l'approximation de Laplace imbriquée et intégrée (INLA). Bien que les méthodes basées sur la simulation soient théoriquement correctes, elles peuvent être coûteuses en termes de calcul. L'objec-

and two MCMC algorithms (in the software JAGS and STAN) using ATTACC/ACTIV-4a trial data of patients who were hospitalized for Covid-19 but not critically ill. By fitting Bayesian hierarchical generalized mixed models with categorical, binary and time-to-event outcomes, the posterior distributions of the treatment effect are compared. INLA requires noticeably less computational time compared to STAN and JAGS (seconds compared to hours). All the 95% CIs for the treatment effect estimated using INLA overlapped with the simulation-based methods.

tif de l'étude est de comparer INLA et deux algorithmes MCMC (dans les logiciels JAGS et STAN) en utilisant les données de l'essai ATTACC/ACTIV-4a de patients hospitalisés pour la Covid-19 mais non gravement malades. En ajustant des modèles mixtes généralisés hiérarchiques bayésiens à des réponses catégorielles, binaires et temporelles, nous comparons les distributions a posteriori de l'effet du traitement. INLA nécessite nettement moins de temps de calcul que STAN et JAGS (quelques secondes contre quelques heures). Tous les IC à 95 % pour l'effet du traitement estimé à l'aide d'INLA recoupent ceux des méthodes basées sur la simulation.

[11:30-11:45]

James Willard (McGill University)

Interim Analysis Covariate Adjustment for Bayesian Group Sequential Designs

Analyse intermédiaire avec covariables d'ajustement pour des plans bayésiens séquentiels de groupe

In conventionally randomized controlled trials, adjustment for baseline prognostic information (BPI) is used to increase power. However, its performance hasn't been formally characterized within the context of more flexible designs, such as Bayesian group sequential designs (BGS). BGS are sequentially randomized and allow for early stopping at interim analyses based on predefined stopping rules, which are typically a function of the posterior probability of the treatment effect. Adjustment for BPI at each interim analysis improves the posterior estimation of the treatment effect, so its use is shown to be beneficial for BGS. The present research investigates the impact of BPI adjustment on BGS with continuous, binary and time-to-event outcomes through a simulation study. Several scenarios for the interim analysis adjustment models are used. The impact of these adjustment models on power, expected sample size, probability of stopping the trial early, and bias is quantified.

Pour accroître l'efficacité des essais contrôlés randomisés conventionnels, on utilise l'ajustement de l'information pronostique de référence (BPI). Sa performance n'est toutefois pas formellement caractérisée dans le contexte de plans plus souples, comme les plans bayésiens séquentiels de groupe (BGS). Les BGS sont randomisés séquentiellement et ils permettent un arrêt précoce dans les analyses intermédiaires, basé sur des règles d'arrêt prédéterminées qui sont généralement une fonction de probabilité a posteriori de l'effet du traitement. L'ajustement de la BPI au cours de chaque analyse intermédiaire améliore l'estimation a posteriori de l'effet du traitement, et par conséquent son utilisation est bénéfique dans les BGS. À l'aide d'une étude en simulation, la présente recherche étudie l'impact de l'ajustement de la BPI sur les BGS lorsque la variable réponse est continue, binaire ou représente un temps d'attente avant un événement. Plusieurs scénarios sont utilisés pour les modèles d'ajustement de l'analyse intermédiaire. L'impact de ces modèles d'ajustement sur l'efficacité, la taille d'échantillon attendu, la probabilité d'arrêt précoce de l'essai et le biais est quantifié.

[11:45-12:00]

Shamsia Sobhan (University of Manitoba) **Mahmoud Torabi** (University of Manitoba)

Spatial Survival Analysis in Presence of Semi-Competing Risks

Analyse de survie spatiale en présence de risques semi-concurrents

Semi-competing risks data arise where a non-terminal event (e.g., lung cancer) is censored by a terminal event (e.g., death). In some applications, semi-competing risks data are arranged in clusters such as geographic regions. Incorporating the cluster effect on the risk of events not only improves the accuracy and efficiency of parameter estimation, but also investigate spatial pattern of events over the study period and identify high-risk

On parle de données sur les risques semi-concurrents lorsqu'un événement non terminal (par exemple, cancer du poumon) est censuré par un événement terminal (par exemple, le décès). Dans certaines applications, ces données sur les risques semi-concurrents sont disposées en grappes, par exemple des régions géographiques. L'inclusion de l'effet de grappe sur le risque d'événements permet non seulement d'améliorer la précision et l'efficacité de l'estimation des paramètres, mais également

areas. The commonly used spatial-survival models are mostly restricted to single-event or competing risks settings. This work proposes a spatial semi-competing risk model in a Bayesian setting that allows for spatial variation while estimating risks of terminal and non-terminal events. The performance of the proposed model is evaluated in a simulation study, and a comparison of models with or without spatial effect is provided to investigate the cost of ignoring spatial variation. We also illustrate the proposed model in real data examples.

d'étudier le schéma spatial des événements sur la période d'étude et d'identifier les zones à haut risque. Or les modèles de survie spatiale couramment utilisés sont pour la plupart limités à un seul événement ou à des risques concurrents. Ce travail propose un modèle spatial de risque semi-concurrent dans un cadre bayésien qui permet une variation spatiale tout en estimant les risques d'événements terminaux et non terminaux. Nous évaluons la performance du modèle proposé via une étude de simulation, et nous comparons des modèles avec ou sans effet spatial pour étudier le coût de l'exclusion de la variation spatiale. Nous illustrons également le modèle proposé sur des exemples de données réelles.

[12:00-12:15]

Victoire Michal (McGill University) **Laís Picinini Freitas** (Fundação Oswaldo Cruz) **Alexandra M. Schmidt** (McGill University)

A Bayesian Hierarchical Model for Disease Mapping that Accounts for Scaling and Heavy-tailed Latent Effects

Un modèle hiérarchique bayésien en cartographie des maladies tenant compte de l'échelle et d'effets latents à queue épaisse

In disease mapping, we estimate the relative risk of a disease across different areas within a region of interest. The number of cases in an area is often modelled through a Poisson distribution with mean given by the product between an offset and the logarithm of the relative risk of the disease. The Besag, York and Mollié model, commonly used to account for potential overdispersion and a spatial correlation structure among the counts, does not accommodate outliers. We define outliers in two ways: areas with extreme risks and areas with different latent behaviours compared to the region of interest. We build on the Bayesian hierarchical model proposed by Riebler et al. (2016) and assume a scale mixture structure wherein the variance of the latent process changes across areas and allows for outlier identification. We compare our approach with that proposed by Congdon (2017), in an analysis of cases of Zika during the 2015-2016 epidemic in Rio de Janeiro.

En cartographie des maladies, le risque relatif d'une maladie est estimé pour différentes zones d'une région d'intérêt. Le nombre de cas par zone est souvent modélisé par une loi Poisson dont la moyenne est le produit entre un facteur et le logarithme du risque relatif de la maladie. Le modèle Besag, York et Mollié, fréquemment utilisé afin de prendre en compte une potentielle sur-dispersion et une corrélation spatiale des nombres de cas, ne tient pas compte de possibles valeurs aberrantes. Nous parlons de valeurs aberrantes dans deux cas : des zones avec des risques extrèmes et des zones avec un comportement latent différent du reste de la région d'intérêt. Nous développons le modèle hiérarchique bayésien proposé par Riebler et al. (2016) et supposons un mélange d'échelle, où la variance du processus latent est différente entre les régions et permet l'identification de zones aberrantes. Notre approche est comparée à celle de Congdon (2017) dans l'analyse des cas de Zika durant l'épidémie de 2015-2016 à Rio de Janeiro.

[12:15-12:30]

Nikola Surjanovic (University of British Columbia) **Saifuddin Syed** (University of British Columbia) **Alexandre Bouchard-Côté** (University of British Columbia) **Trevor Campbell** (University of British Columbia)

Parallel Tempering With a Variational Reference

Atténuation parallèle avec référence variationnelle

Sampling from multi-modal and high-dimensional target distributions is a challenging task that is often required in order to perform Bayesian inference. Parallel tempering (PT) methods address this problem by constructing a Markov chain on an expanded state space

L'échantillonnage à partir de lois cibles de haute dimension et plurimodales représente un défi de taille qu'il faut souvent relever afin de réaliser une inférence bayésienne. L'atténuation parallèle (AP) aborde ce problème en construisant une chaîne de Markov sur un espace-état étendu qui échantillonne simultanément à partir d'une

that simultaneously samples from a sequence of distributions lying on an annealing path from the prior to the target. In this work we consider generalized annealing paths that start from a variational reference. The reference distribution is tuned to minimize an appropriate notion of distance to the target distribution, maximizing a quantity related to the effective sample size. We apply the method to several posterior inference problems, finding that PT with a variational reference can greatly improve performance. The proposed methodology is particularly useful in cases where the prior and posterior are almost mutually singular and the geometry of the posterior is complex.

séquence de lois fondée sur un chemin hybride tiré de l'a priori de la cible. Dans le cadre de ce travail, nous examinons les chemins hybrides généralisés débutant à partir d'une référence variationnelle. Cette loi de référence est réglée afin de minimiser une notion de distance appropriée par rapport à la loi cible, et maximiser une quantité relative à la taille d'échantillon efficace. Nous appliquons la méthode à plusieurs problèmes d'inférence à posteriori, et démontrons que l'AP avec référence variationnelle peut nettement améliorer la performance. La méthodologie proposée est tout particulièrement pratique dans les cas où l'a priori et l'a posteriori sont presque mutuellement singulier et que la géométrie du postérieur est complexe.

Statistical Analysis of Covid-19 Data
Analyse statistique des données Covid-19

Chair/Président: Lengyi Spectrum Han

Date: Tuesday May 31 / mardi 31 mai

Time/Heure: 11:00-12:30

Abstract/Résumé

[11:00-11:15]

Haoyu Wu (McGill University)

Estimating COVID-19 Incidence using Deaths, Cases, Tests, Surveys, and Vaccinations

Estimation de l'incidence de la COVID-19 en répertoriant les décès, cas, tests, sondages et taux de vaccination

Estimating COVID-19 infections and disease severity is crucial to public health surveillance but often challenging due to inconsistencies in available data sources. Furthermore, with the ongoing vaccination campaign, vaccine-induced immunity substantially impacts the infection transmission rate and death rate, largely parallel to the natural epidemic development. Recently, a Bayesian model based on a latent SIR compartmental model was proposed. It includes likelihood components for deaths, cases, tests, and surveys to estimate the incidence and infection fatality rate (IFR), accounting for biases and delays in the data. We first test the robustness of the model under various simulated scenarios, including age- and time-dependent transmission rate and IFR in an SEIR data generating mechanism. We next propose an extended model to account for the effect of vaccination on IFR and validate it using synthetic data. Both models are then applied to estimate the incidence and IFR in Quebec.

L'estimation du nombre d'infections causées par la COVID-19 et de la gravité de la maladie est essentielle pour la surveillance de la santé publique mais elle est souvent difficile en raison des divergences dans les sources de données disponibles. De plus, avec la campagne de vaccination en cours, l'immunité vaccinale influe sensiblement sur les taux de transmission de l'infection et de décès qui sont largement parallèles au développement naturel de l'épidémie. Un modèle bayésien basé sur un modèle compartimental latent SIR a été proposé récemment. Il comprend des composants de vraisemblance pour les décès, cas, tests et sondages afin d'estimer l'incidence de l'infection et le taux de mortalité (IFR), qui prend en compte les biais et les délais dans les données. Nous évaluons d'abord la robustesse du modèle sous divers scénarios de simulation, y compris le taux de transmission dépendant de l'âge et du temps ainsi que l'IFR pour un mécanisme de génération de données SEIR. Nous proposons ensuite un modèle étendu pour prendre en compte l'effet de la vaccination sur l'IFR et le validons à l'aide de données synthétiques. Nous appliquons ensuite les deux modèles à l'estimation de l'incidence et de l'IFR au Québec.

[11:15-11:30]

Justin James Ian Slater (University of Toronto) **Ayuish Bansal** (Centre for Global Health Research) **Jeffrey S. Rosenthal** (University of Toronto) **Harlan Campbell** (University of British Columbia) **Paul Gustafson** (University of British Columbia) **Patrick E. Brown** (University of Toronto)

An Almost Bayesian Approach to Estimating COVID-19 Incidence and Infection Fatality Rates

Approche quasi-bayésienne pour l'estimation de l'incidence de la COVID-19 et des taux de mortalité par infection

Naive estimates of incidence and infection fatality rates (IFR) of COVID-19 suffer from a variety of biases, many of which relate to preferential testing. This has motivated epidemiologists from around the globe to conduct serosurveys that measure the immunity of individuals by testing for the presence of SARS-CoV-2 antibodies in the blood. These quantitative measures (titre values) are then used as a proxy for previous or current infection. In previous work, this data has not been used

L'estimation naïve de l'incidence de la COVID-19 et des taux de mortalité par infection (IFR) souffre de divers biais dont plusieurs sont liés à des tests préférentiels. C'est ce qui a incité des épidémiologistes du monde entier à mener des enquêtes sérologiques pour mesurer l'immunité des sujets en testant la présence dans le sang d'anticorps du SARS-CoV-2. Ces mesures quantitatives (valeurs de titres) sont utilisées comme approximation d'une infection antérieure ou en cours. Le plein potentiel de ces données n'a pas été exploité dans une étude précédente. Dans

to it's full potential. In this talk, we demonstrate how multivariate mixture models can be used in combination with poststratification to estimate cumulative incidence and IFR in an approximate Bayesian framework without discretization. In doing so, we propagate error from both the estimated number of infections and the incomplete deaths data to provide estimates of IFR. This method is demonstrated using data from the Action to Beat Coronavirus (Ab-C) serosurvey in Canada.

notre exposé, nous montrons comment des modèles de mélanges multivariés peuvent être utilisés en conjugaison avec une poststratification pour estimer l'incidence cumulative et l'IFR dans un cadre bayésien approximatif sans discrétisation. De cette façon, des estimations de l'IFR sont fournies en utilisant une propagation des erreurs à partir du nombre estimé des infections et des données incomplètes sur les décès. Cette méthode est illustrée à l'aide de données tirées de l'enquête sérologique canadienne Action to Beat Coronavirus (Ab-C).

[11:30-11:45]

Yasin Khadem Charvadeh (University of Western Ontario) **Grace Y. Yi** (University of Western Ontario)

Comparing the Effectiveness of Virus Control Policies for COVID-19 with the Q-Learning Method

Comparer l'efficacité des politiques antivirus pour la COVID-19 avec la méthode d'apprentissage par renforcement (Q-learning)

The case fatality rate of COVID-19 is one of the useful measures to compare the disease severity for different countries. As mitigation policies on controlling the spread of COVID-19 vary from country to country, it is interesting to study how the COVID-19 case fatality rate of a country may be associated with its mitigation policies as well as risk factors. In this talk, we investigate this problem by examining the COVID-19 data from 175 countries for a period of nine months. We embed the problem into a dynamic framework and utilize the Q-learning method to explore how different preventive policies may be related to lowering the case fatality rate.

Le taux de létalité de la COVID-19 est l'une des mesures les plus pratiques pour comparer la gravité de la maladie pour différents pays. Vu que les politiques d'atténuation relatives au contrôle de la propagation de la COVID-19 varient d'un pays à l'autre, il est intéressant d'étudier comment le taux de létalité de la COVID-19 peut être relié à ses politiques d'atténuation ainsi que ses facteurs de risque. Lors de cet exposé, nous étudions ce problème en examinant les données de la COVID-19 provenant de 175 pays dans une période de neuf mois. Nous avons intégré le problème dans un cadre dynamique et utilisons la méthode Q-learning pour découvrir comment différentes politiques de prévention pourraient être liées à la réduction du taux de létalité.

[11:45-12:00]

Yijia Weng (Western University)

Meta-Analysis for Estimating the COVID-19 Average Incubation Time

Méta-analyse pour l'estimation du temps moyen d'incubation de la COVID-19

Many studies have been carried out to estimate the average incubation time for COVID-19. However, available studies do not reveal comparable estimates of the mean incubation time, and they vary considerably from 1.8 days to 14 days. It is difficult to assess which estimate more reasonably reflects the average incubation time of the population because different studies are carried out for different subjects under different conditions. In this talk, we explore synthetic estimates of the average incubation time of COVID-19 by capitalizing on the report estimates in the literature, and assess heterogeneity involved with the reported studies on COVID-19 as well as the publication bias. We take different angles to estimate the mean incubation time, and our analyses provide estimates which range from 5.68 days to 8.30

De nombreuses études ont été réalisées pour estimer le temps d'incubation moyen de la COVID-19. Cependant, les études actuelles ne révèlent pas d'estimations comparables de la durée moyenne d'incubation, et elles varient considérablement de 1,8 jour à 14 jours. Il est ainsi difficile d'évaluer quelle estimation reflète le plus raisonnablement le temps d'incubation moyen de la population, car les différentes études sont réalisées chez différents sujets dans différentes conditions. Dans cette présentation, nous explorons des estimations synthétiques du temps d'incubation moyen de la COVID-19 en exploitant les estimations rapportées dans la littérature, et nous évaluons l'hétérogénéité associée aux études rapportées sur la COVID-19, ainsi que le biais de publication. Nous adoptons différents angles pour estimer la durée moyenne d'incubation. Nos analyses fournissent des estimations qui vont de 5,68 jours à 8,30 jours.

days.

[12:00-12:15]

Yuan Bian (University of Western Ontario) **Yasin Khadem Charvadeh** (University of Western Ontario) **Grace Y. Yi** (University of Western Ontario) **Wenqing He** (University of Western Ontario)

Is 14-Days a Sensible Quarantine Length for COVID-19? A Case Study of COVID-19 Incubation Times

Une période de quarantaine de 14 jours pour la COVID-19 est-elle raisonnable ? Étude de cas sur les périodes d'incubation de la COVID-19

To confine the spread of an infectious disease, setting a sensible quarantine time is crucial, which is however not trivial. It depends on various underlying factors, including a good understanding of the distribution of incubation times of the disease. Regarding the ongoing COVID-19 pandemic, 14-days is commonly taken as a quarantine time. In this talk, we examine the distribution of the COVID-19 incubation time using likelihood-based methods. Our study is carried out on a case study which includes 178 COVID-19 cases with the information of exposure periods and dates of symptom onset collected. We employ different models to describe incubation times of COVID-19. Our findings suggest that statistically, the 14-day quarantine time may not be long enough to control the probability of an early release of infected individuals to be small.

Pour limiter la propagation d'une maladie infectieuse, il est crucial de fixer une période de quarantaine raisonnable, ce qui n'est toutefois pas trivial. Elle dépend de plusieurs facteurs sous-jacents, notamment d'une bonne compréhension de la distribution des périodes d'incubation de la maladie. Concernant la pandémie actuelle de COVID-19, on considère en général que la période de quarantaine est de 14 jours. Dans cette présentation, nous examinons la distribution des périodes d'incubation de la COVID-19 à l'aide de méthodes fondées sur la vraisemblance. Nos travaux reposent sur une étude de cas qui comprend 178 cas de COVID-19, dont les informations sur les périodes d'exposition et les dates d'apparition des symptômes ont été recueillies. Nous utilisons différents modèles pour décrire les périodes d'incubation de la COVID-19. Nos résultats suggèrent que, sur le plan statistique, la période de quarantaine de 14 jours n'est peut-être pas assez longue pour limiter le risque d'une libération précoce des personnes infectées.

[12:15-12:30]

Jingxue Feng (Simon Fraser University) **Jie Wang** (Simon Fraser University) **Jiarui Zhang** (Simon Fraser University) **Liangliang Wang** (Simon Fraser University)

Clustering and Identification of SARS-CoV-2 Mutations Associated with Clinical Severity

Regroupement et identification des mutations du SRAS-CoV-2 associées à la sévérité clinique

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is a novel coronavirus emerged in December 2019 in China that causes the outbreak of COVID-19 worldwide. The genetic cluster analysis of SARS-CoV-2 variants is crucial to characterize the virus and has been widely studied. However, the existing genetic clustering methods are merely based on whole genome sequencing data without giving consideration to clinical features. In our work, with the involvement of both genome sequencing data and clinical data, we developed a model-based clustering method to group SARS-CoV-2 mutations that share similar relationship to the clinical features, with a focus in disease severity. Parameters in the model are estimated via Bayesian inference that takes model uncertainty into account. Our analysis will facilitate the process of identifying clusters of SARS-CoV-2 mutations, and simultaneously provides insights

Le coronavirus du syndrome respiratoire aigu sévère 2 (SRAS-CoV-2), un nouveau coronavirus apparu en décembre 2019 en Chine, est à l'origine de l'épidémie de la COVID-19 dans le monde entier. L'analyse par groupement génétique des variants du SRAS-CoV-2 est cruciale pour caractériser le virus et a fait l'objet de nombreuses études. Cependant, les méthodes de groupement génétique actuelles reposent uniquement sur les données de séquençage du génome entier et ne tiennent pas compte des caractéristiques cliniques. Dans le cadre de nos travaux, nous mettons au point une méthode de groupement basée sur un modèle, en faisant appel à la fois aux données de séquençage du génome et aux données cliniques. Notre objectif est de regrouper les mutations du SRAS-CoV-2 qui présentent une relation semblable aux caractéristiques cliniques, en accordant une attention particulière à la gravité de la maladie. Nous estimons les paramètres du modèle par inférence bayésienne en tenant compte de l'incertitude du modèle. Notre analyse facilite le processus d'identification des

into the association between mutations and clinical features.

groupes de mutations du SRAS-CoV-2 et permet en même temps de mieux comprendre l'association entre les mutations et les caractéristiques cliniques.

SSC 2021 Gold Medal Address
Allocution du récipiendaire de la Médaille d'or 2021 de la SSC

Chair/Président: Bruce Smith

Date: Tuesday May 31 / mardi 31 mai

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-14:30]

Art Owen (Stanford University) **Hal Varian** (Google) **Dan Kluger** (Stanford University) **Harrison Li** (Stanford University) **Tim Morrison** (Stanford University)

Tie-breaker Designs

Plans d'échantillonnage de bris d'égalité

Companies may offer incentives to their best customers and philanthropists may offer scholarships to the strongest students. They can evaluate the impact of these treatments later using a regression discontinuity analysis. Unfortunately, regression discontinuity analyses have high variance. It is possible to get much more statistical efficiency using a tie-breaker design that works by triage: top subjects get the treatment, bottom subjects do not, and those in between have their treatment randomized. Statistical efficiency increases monotonically with the amount of randomization, causing an exploration versus exploitation tradeoff. This holds in a simple two line regression model and also with nonparametric kernel regression based methods. We have found D-optimal treatment probability functions for scalar and vector data. The conclusion is that when it is possible (and ethical) to randomize for a group of subjects, it is wise to do so. Some portions of this work were done as a paid consultant for Google and were not part of Art Owen's Stanford responsibilities. Subsequent work was done as a Stanford faculty member.

Les entreprises offrent parfois des primes à leurs meilleurs clients et les philanthropes offrent des bourses d'études aux meilleurs étudiants. Ils peuvent vouloir évaluer l'impact de ces traitements par la suite via une analyse de discontinuité de régression. Malheureusement, les analyses de discontinuité de régression présentent une variance élevée. Il est possible d'obtenir une efficacité statistique beaucoup plus grande en utilisant un plan d'échantillonnage de bris d'égalité qui fonctionne par triage : les sujets du haut de l'échelle reçoivent le traitement, les sujets du bas de l'échelle ne le reçoivent pas, et ceux qui se trouvent entre les deux voient leur traitement randomisé. L'efficacité statistique augmente de façon monotone avec la quantité de randomisation, ce qui entraîne un compromis entre exploration et exploitation. Ceci est valable dans un modèle de régression simple à deux lignes et également avec des méthodes non paramétriques basées sur la régression à noyau. Nous avons trouvé des fonctions de probabilité de traitement D-optimales pour les données scalaires et vectorielles. La conclusion est que lorsqu'il est possible (et éthique) de randomiser pour un groupe de sujets, il est sage de le faire. Certaines parties de ce travail ont été réalisées à titre de consultant rémunéré pour Google et ne font pas partie des responsabilités d'Art Owen à Stanford. Les travaux ultérieurs ont été réalisés à titre de membre de la faculté de Stanford.

New Advances in Microbiome Data Science
Nouvelles avancées en science des données sur le microbiome

Chair/Président: Depeng Jiang

Organizer/Responsable: Pingzhao Hu

Date: Tuesday May 31 / mardi 31 mai

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-16:00]

Longhai Li (University of Saskatchewan) **Wei Bai** (University of Saskatchewan) **Mei Dong** (University of Toronto) **Longhai Li** (University of Saskatchewan) **Wei Xu** (University of Toronto)

Randomized Quantile Residuals for Diagnosing Zero-Inflated Generalized Linear Mixed Models with Applications to Microbiome Count Data

Résidus quantiles randomisés pour le diagnostic des modèles linéaires généralisés mixtes avec excès de zéros et applications aux données de comptage du microbiome

For differential abundance analysis, zero-inflated generalized linear models, typically zero-inflated NB models, have been increasingly used to model microbiome and other sequencing count data. A common assumption in estimating the false discovery rate is that the p values are uniformly distributed under the null hypothesis, which demands that the postulated model fit the count data adequately. Therefore, model checking is critical to control the FDR at a nominal level in differential abundance analysis. We conduct large-scale simulation studies to investigate the performance of the RQRs for zero-inflated GLMMs. The simulation studies show that the type I error rates of the GOF tests with RQRs are very close to the nominal level. We also apply the RQRs to diagnose six GLMMs to a real microbiome dataset. The results show that the OTU counts at the genus level of this dataset can be modelled well by zero-inflated and zero-modified NB models.

Dans le cadre de l'analyse de l'abondance différentielle, les modèles linéaires généralisés avec excès de zéros, communément les modèles BN avec excès de zéros, sont largement utilisés pour modéliser les données de comptage du microbiome et d'autres données de séquençage. Une hypothèse courante lorsqu'on estime le taux de fausses découvertes est que les valeurs-p sont uniformément distribuées sous l'hypothèse nulle, ce qui exige que le modèle proposé s'ajuste correctement aux données de comptage. Par conséquent, la vérification du modèle est essentielle pour contrôler le taux de fausses découvertes à un taux nominal dans les analyses d'abondance différentielle. Nous réalisons des études de simulation à grande échelle afin d'étudier l'efficacité des résidus quantiles randomisés pour les modèles linéaires généralisés mixtes avec excès de zéros. Les études de simulation montrent que les taux d'erreur de type I des tests de qualité d'ajustement avec les résidus quantiles randomisés sont très proches du seuil nominal. Nous appliquons également les résidus quantiles randomisés pour diagnostiquer six modèles linéaires généralisés mixtes à un ensemble de données réelles sur le microbiome. Les résultats montrent que le nombre d'unités taxonomiques opérationnelles au niveau du genre de cet ensemble de données peut être bien modélisé par des modèles binomiaux négatifs avec excès de zéros et avec des zéros modifiés.

[16:00-16:30]

Wei Xu (Princess Margaret Cancer Centre)

Model Development on Longitudinal Microbiome Sequencing Data using Machine-learning Methodology

Développement de modèle pour les données longitudinales de séquençage de microbiome au moyen d'une méthodologie d'apprentissage automatique

Human microbiome is dynamic in nature, attributing to the presence of interactions among microbes, host, and

Le microbiome humain est de nature dynamique, en raison de la présence d'interactions entre les microbes, l'hôte et l'environne-

the environment. Researchers have shown that the microbiome can change over time, transiently or in long term, by infections or due to medical interventions such as antibiotics. In this presentation, I will introduce some of the machine learning models for longitudinal data prediction tasks and their application in disease prediction using microbiome sequencing data with repeated measures. I will show how advanced neural networks such as stratified Convolutional Neural Networks (CNNs) and Long Short-Term Memory networks (LSTMs) can be used for feature extraction and temporal dependency analysis in longitudinal microbiome data. I will also discuss about the challenges and future directions in this research area, along with the performance comparison with other machine learning models.

ment. Les chercheurs ont démontré que le microbiome peut changer avec le temps, de façon éphémère ou à long terme, à cause d'infections ou d'interventions médicales comme les antibiotiques. Lors de cet exposé, je vous présenterai certains modèles d'apprentissage automatique pour les tâches relatives aux données longitudinales ainsi que leur application dans la prédiction de maladie à l'aide de données de séquençage du microbiome avec mesures répétées. Je montrerai comment les réseaux neuronaux avancés comme le réseau de neurones convolutif (CNN) et les réseaux de longue mémoire à court terme (LSTM) peuvent servir à extraire la caractéristique et analyser la dépendance temporelle dans des données longitudinales de microbiome. J'aborderai aussi les défis et les orientations à venir dans ce domaine de recherche, ainsi que la comparaison de performance avec d'autres modèles d'apprentissage automatique.

[16:30-17:00]

Pingzhao Hu (University of Manitoba)

Computational meta-analyses of oral microbiome studies

Méta-analyses computationnelles des études sur le microbiome buccal

Dental caries is the term used for tooth decay and cavities and it is the most prevalent infectious disease in the oral cavity. Caries in children with primary dentition is known as early childhood caries (ECC) and it affects about half of the children worldwide. The microorganisms found in the oral cavity are collectively called the oral microbiome. The dysbiosis of microbiomes causes several oral diseases including caries. Although some oral microbiome-based studies have been performed to identify potential microbiome biomarkers to predict ECC status, the results were not reproducible due to small sample sizes. Machine learning (ML) can help in identifying the pattern in heterogeneous data obtained from different cohorts for ECC. In this study, we explored state-of-the-art ML models to identify ECC markers for the classification of ECC across multiple cohorts.

Le terme « carie » désigne la destruction progressive d'une dent, qui engendre la formation d'une cavité. Il s'agit de la maladie infectieuse la plus répandue dans la cavité buccale. La carie chez les enfants qui ont une dentition primaire est connue sous le nom de « carie de la petite enfance » et touche près de la moitié des enfants dans le monde. Les micro-organismes présents dans la cavité buccale sont communément appelés le microbiome buccal. La dysbiose des microbiomes est à l'origine de plusieurs maladies buccales, dont les caries. Bien que certaines études axées sur le microbiome buccal aient été réalisées pour identifier des biomarqueurs potentiels du microbiome afin de prédire l'état des caries de la petite enfance, les résultats n'étaient pas réplicables en raison de la petite taille des échantillons. L'apprentissage automatique peut aider à déterminer le schéma des données hétérogènes obtenues à partir de différentes cohortes concernant les caries de la petite enfance. Dans cette étude, nous examinons les derniers modèles d'apprentissage automatique afin d'identifier les marqueurs de caries de la petite enfance pour la classification de celles-ci dans plusieurs cohortes.

Functional Data Analysis for Complex Data
Analyse fonctionnelle de données complexes

Chair/Président: Liangliang Wang

Organizer/Responsable: Liangliang Wang

Date: Tuesday May 31 / mardi 31 mai

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-16:00]

Jiguo Cao (Simon Fraser University) **Shu Jiang** (University of Waterloo) **Graham Colditz Bernard Rosner**

Predicting the Onset of Breast Cancer using Mammogram Imaging Data with Irregular Boundary

Prévision de l'apparition du cancer du sein à l'aide de données d'imagerie mammaire à limites irrégulières

With mammography being the primary breast cancer screening strategy, it is essential to make full use of the mammogram imaging data to better identify women who are at higher and lower than average risk. Our primary goal is to extract mammogram-based features that augment the well-established breast cancer risk factors to improve prediction accuracy. In this talk, I will introduce a supervised functional principal component analysis (sFPCA) over triangulations method for extracting features that are ordered by the magnitude of association with the failure time outcome. The proposed method accommodates the irregular boundary issue posed by the breast area within the mammogram imaging data with flexible bivariate splines over triangulations.

La mammographie étant la principale stratégie de dépistage du cancer du sein, il est essentiel d'exploiter pleinement les données d'imagerie mammaire pour mieux identifier les femmes qui présentent un risque supérieur ou inférieur à la moyenne. Notre objectif principal est d'extraire des caractéristiques de mammographie qui augmentent les facteurs de risque de cancer du sein bien établis, afin d'améliorer la précision de la prédiction. Dans cet exposé, je présenterai une méthode d'analyse en composantes principales fonctionnelles supervisée (sFPCA) sur triangulations pour extraire des caractéristiques qui sont ordonnées par l'ampleur de l'association avec le résultat du temps de défaillance. La méthode proposée tient compte du problème des limites irrégulières posées par la zone du sein dans les données d'imagerie mammaire avec des splines bivariées flexibles sur triangulations.

[16:00-16:30]

Tianyu Guan (Brock University)

Exploring Pre-launch Movie Electronic Word of Mouth Time Series by Functional Data Analysis

Exploration des critiques de films avant sortie par analyse de données fonctionnelles sur séries chronologiques

Online product reviews, commonly conceptualized as electronic Word of Mouth, are essentially a time series of multinomial distributions with two dimensions of interests, namely the time dimension and the rating dimension. In this research, we apply functional data analysis to study the data stream of online product reviews so as to find the most efficient way to summarize online product reviews in both the time and rating dimensions in predicting the subsequent sales. We observe that most online product review ratings exhibit a positivity bias and extremity, therefore, we apply the functional principal component analysis to explore the major variations among the quantile curves of the movies. The functional principal component (FPC) scores at various quantile

Les critiques de produits en ligne, ou bouche-à-oreille électronique, constituent essentiellement en une série chronologique de distributions multinomiales avec deux dimensions d'intérêt, à savoir la dimension temporelle et la dimension de notation. Dans cette recherche, nous appliquons l'analyse de données fonctionnelles pour étudier le flux de données des critiques de produits en ligne et trouver la manière la plus efficace de les résumer dans les deux dimensions de temps et d'évaluation pour prédire les ventes ultérieures. Nous observons que la plupart des critiques de produits en ligne présentent un biais de positivité et d'extrémisme. Par conséquent, nous appliquons l'analyse en composantes principales fonctionnelle pour explorer les principales variations entre les courbes quantiles des films. Nous utilisons ensuite les scores de la composante principale fonctionnelle (CPF) à divers niveaux de quantile

levels are then used to predict the box office revenues in the opening week. We use the group LASSO method to select the quantile levels at which the FPC scores make significant contributions to the prediction. In addition, this research shows that top-end percentiles would be better summary statistics compared to the mean in capturing the relations between the pre-launch product ratings time pattern and launch sales.

pour prédire les recettes du box-office lors de la semaine de sortie. Nous utilisons la méthode LASSO de groupe pour sélectionner les niveaux de quantile auxquels les scores CPF contribuent de manière significative à la prédiction. En outre, cette recherche montre que les percentiles supérieurs seraient de meilleures statistiques sommaires par rapport à la moyenne pour refléter les relations entre le schéma temporel des critiques de produit avant sortie et les ventes lors de la sortie.

[16:30-17:00]

Jinhan Xie (University of Alberta)

Optimal Functional Logistic Regression Under Case-Control Design

Régression logistique fonctionnelle optimale planifiée sous un cas-témoins

It is well-known that a case-control study is a more promising approach for studying the relationship of existing factors and rare disease incidence in many medical studies or epidemiology, contrary to prospective studies. This paper studies the functional logistic regression model with a functional predictor under case-control study. The coefficient function together with the intercept parameter are estimated through the penalized likelihood approach. Theoretically, we establish the minimax rates of convergence for estimating coefficient function under mild conditions. Extensive empirical studies including simulations and the real data example are conducted to examine the finite-sample performance of the proposed method.

L'étude cas-témoins est bien connue en tant qu'approche prometteuse pour étudier le lien entre des facteurs existants et des cas de maladie rare dans un grand nombre d'études médicales ou épidémiologiques, contrairement aux études prospectives. Cet article étudie le modèle de régression logistique fonctionnel avec un prédicteur fonctionnel dans le cadre d'une étude cas-témoins. La fonction de coefficient avec le paramètre de constante est estimée à l'aide de l'approche de vraisemblance pénalisée. Théoriquement, nous établissons les taux minimax de convergence afin d'estimer la fonction de coefficient selon les conditions modérées. Nous menons des études empiriques approfondies comprenant des simulations et des exemples à partir de données réelles afin d'examiner la performance de l'échantillon fini de la méthode proposée.

Reflection and Outlook of Statistical Sciences – Celebration of the 50th Anniversary of the Canadian Statistical Community
Réflexion et perspectives des sciences statistiques – Célébration du 50e anniversaire de la communauté statistique canadienne

Chair/Président: Grace Y. Yi

Organizer/Responsable: Grace Y. Yi

Date: Tuesday May 31 / mardi 31 mai

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-17:00]

Charmaine B. Dean (University of Waterloo) **Richard Lockhart** (Simon Fraser University) **Johanna G. Nešlehová** (McGill University) **Bruno Rémillard** (HEC Montréal) **Thérèse A. Stukel** (ICES/ University of Toronto) **Lei Sun** (University of Toronto)

Reflection and Outlook of Statistical Sciences – Celebration of the 50th Anniversary of the Canadian Statistical Community

Réflexion et perspectives des sciences statistiques - Célébration du 50e anniversaire de la communauté statistique canadienne

In celebrating the 50th anniversary of the Canadian statistical community, this panel session brings together several prominent statisticians to share their insights and perspectives about statistical sciences from different angles. Topics include (1). tackling the challenges of beginning a career in academia, (2). developing leadership skills, (3). training the next generation of statistical scientists with EDI, (4). establishing collaborative networks, (5). reflecting the evolvement of statistical sciences in the past, and (6). outlooking the future directions of the profession.

En célébrant le 50e anniversaire de la communauté statistique canadienne, cette séance d'experts réunit plusieurs statisticiens éminents qui partageront leurs idées et leurs perspectives au sujet des sciences statistiques sous différents angles. Les sujets abordés sont les suivants : (1). relever les défis associés au début d'une carrière universitaire, (2). développer des compétences en leadership, (3). former la prochaine génération de statisticiens avec l'EDI, (4). établir des réseaux de collaboration, (5). refléter l'évolution des sciences statistiques dans le passé, et (6). prévoir les orientations futures de la profession.

Data Fairness and Ethics
Équité des données et éthique

Chair/Président: Nathan A. Taback

Organizer/Responsable: Nathan A. Taback

Date: Tuesday May 31 / mardi 31 mai

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-16:00]

Fanny Chevalier (University of Toronto)

Don't Look. See! Are we Blinded by Data (Visualization)?

Sommes-nous aveuglés par les (visualisations de) données ?

We are constantly required to make decisions about the world we live in. But are we good judges of how things work and what's best to do in each situation? To help with this process, we often rely on data to inform our decision, but is it enough? This talk will explore why we may not always make well-informed decisions, even with best intentions, and even when our motivations are driven by careful examination of data. I will challenge the ways we leverage data for analysis and communication, and will propose strategies that embrace the imperfect, subjective nature of human's perception.

Chaque jour, il nous faut prendre des décisions qui ont trait au monde dans lequel nous évoluons. Mais sommes-nous de bons juges, quand il s'agit de comprendre comment les choses fonctionnent et quelle est la meilleure chose à faire dans chaque situation ? Pour aider notre raisonnement, nous avons souvent recours aux données et à leur analyse, pour apprécier le problème et envisager des solutions. Dans cette présentation, je vais passer en revue certaines raisons pour lesquelles il nous est difficile, voire impossible de prendre des décisions éclairées, même avec les meilleures intentions, et ce, même quand nos motivations sont dictées par une analyse profonde des données à notre disposition. Je discuterai les verrous associés à la façon dont nous utilisons les données pour l'analyse et la communication, et proposerai des stratégies qui tiennent compte des imperfections et de la nature subjective de la perception humaine.

[16:00-16:30]

Lauren Klein (Emory University)

Data Feminism

Féminisme et données

What is feminist data science? How is feminist thinking being incorporated into data-driven work? And how are scholars in the humanities and social sciences, in particular, bringing together data science and feminist theory in their research? Drawing from her recent book, Data Feminism (MIT Press), coauthored with Catherine D'Ignazio, Klein will present a set of principles for doing data science that are informed by the past several decades of intersectional feminist activism and critical thought. In order to illustrate these principles, as well as some of the ways that scholars and designers have begun to put them into action, she will discuss a range of recent research projects including several of her own.

Comment définir la science des données féministe ? Comment la pensé féministe est-elle incorporée aux travaux axés sur les données ? Et comment les spécialistes, en particulier dans les sciences humaines et sociales, allient-ils science des données et théorie féministe dans leurs recherches ? En s'appuyant sur son récent ouvrage Data Feminism (MIT Press), coécrit avec Catherine D'Ignazio, Lauren Klein présente un ensemble de principes pour faire de la science des données sous l'éclairage de plusieurs décennies d'activisme et de pensée critique féministes intersectionnels. Afin d'illustrer ces principes de même que certains des moyens déjà mis en œuvre par des spécialistes et concepteurs, elle fera état d'un certain nombre de projets de recherche récents, y compris plusieurs des siens. Ensemble, ces exemples montre la

Taken together, these examples demonstrate how feminist thinking can be operationalized into more ethical, more intentional, and more capacious data practices.

façon dont la pensée féministe peut être opérationnalisée en pratiques axées sur les données plus éthiques, plus intentionnelles et de plus grande ampleur.

Chair/Président: Jinko Graham

Organizer/Responsable: Jinko Graham

Date: Tuesday May 31 / mardi 31 mai

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-16:00]

Brad McNeney (Simon Fraser University) **Pulindu Ratnasekera** (Simon Fraser University) **Jinko Graham** (Simon Fraser University)

Robust Inference of Gene-Environment Interaction from Heterogeneous Samples of Case-Parent Trios

Inférence robuste de l'interaction gène-environnement à partir d'échantillons hétérogènes

In a case-parent trio study we collect genotypes on affected children and their parents. Information may also be collected on the child's environmental exposures. The design permits estimation and testing of genetic effects and gene-by-environment interaction. Inference of genetic effects is robust to population structure, but when genotypes are measured at a non-causal test locus, population stratification can create spurious interaction. That is, the exposure can appear to modify the disease risk of genotypes at the test locus without actually modifying the disease risk of genotypes at the causal locus. We review previous methods to reduce bias from population stratification and propose a new method in which we adjust the risk model by principal components computed from a genome-wide panel of markers. The method is illustrated on simulated data and on data from a study of genetic modifiers of exposures known to affect the risk of cleft palate.

Dans une étude de trio cas-parents, nous recueillons des génotypes provenant d'enfants atteints et leurs parents. Les renseignements peuvent aussi être recueillis à partir des expositions environnementales de l'enfant. Le modèle permet l'estimation et le test des effets génétiques et des interactions gène-environnement. L'inférence d'effets génétiques est robuste par rapport à la structure de la population, mais lorsque les génotypes sont mesurés selon un locus test non causal, la stratification de la population peut produire des interactions fausses. C'est-à-dire que l'exposition peut survenir pour modifier le risque de maladie des génotypes relatif chez le locus test sans pour autant modifier le risque de maladie des génotypes chez le locus causal. Nous faisons le point sur les méthodes précédentes servant à réduire le biais de la stratification de la population et proposons une nouvelle méthode dans laquelle nous ajustons le modèle de risque par des composés principaux calculés à partir d'un panneau de marqueurs à l'échelle du génome. Nous illustrons la méthode à l'aide de données simulées et de données tirées de modificateurs génétiques des expositions reconnues pour influencer le risque de fente palatine.

[16:00-16:30]

Qingrun Zhang (University of Calgary)

cLD: Rare-Variant Disequilibrium Between Genomic Regions Identifies Novel Genomic Interactions

Déséquilibre de liaison cumulatif (cLD) : un déséquilibre lié à un variant rare entre des régions génomiques pour identifier de nouvelles interactions génomiques

Linkage disequilibrium (LD) is a fundamental concept in genetics; critical for studying genetic associations and molecular evolution. However, LD measurements are only reliable for common genetic variants, leaving low-frequency variants unanalyzed. In this work, we introduce cumulative LD (cLD), a stable statistic that captures the rare-variant LD between genetic regions

Le déséquilibre de liaison (LD), un concept fondamental en génétique, est essentiel à l'étude des associations génétiques et de l'évolution moléculaire. Les mesures LD sont toutefois seulement fiables pour les variants génétiques répandus, avec pour résultat l'absence d'analyse de variants à faible fréquence. Nous présentons un LD cumulatif (cLD), une statistique stable qui décrit le déséquilibre de liaison lié à un variant rare entre des

and opens the door for furthering biological knowledge using rare genetic variants. In application, we find cLD reveals an increased genetic association between genes in 3D chromatin interactions, a phenomenon recently reported negatively by calculating standard LD between common variants. Additionally, we show that cLD is higher between gene pairs reported in interaction databases, identifies unreported protein-protein interactions, and reveals interacting genes distinguishing case/control samples in association studies.

régions génétiques, tout en ouvrant la voie à un approfondissement du savoir biologique à l'aide de variants génétiques rares. Dans son application, nous découvrons que le cLD révèle une association génétique accrue entre les gènes dans les interactions de la chromatine de structure 3D, un phénomène dont on a fait état négativement récemment en calculant le LD standard entre des variants répandus. Nous montrons aussi que le cLD est plus élevé entre des paires de gènes signalées dans des bases de données sur les interactions, qu'il identifie des interactions protéines-protéines non signalées et révèle des gènes interactifs distinguant des échantillons cas/témoins dans des études d'associations.

[16:30-17:00]

Lloyd T. Elliott (Simon Fraser University)

Brain Imaging Genetics with 40,000 Subjects and 3,000 Phenotypes

Génétique de l'imagerie cérébrale avec 40 000 sujets et 3 000 phénotypes

UK Biobank is a major prospective epidemiological study that is carrying out detailed multimodal brain imaging on 100,000 participants, and includes genetics and ongoing health outcomes. We present a new open resource of GWAS summary statistics, resulting from a greatly expanded set of genetic associations with brain phenotypes, using the 2020 UK Biobank imaging data release of approximately 40,000 subjects, 3,000 phenotypes, and 10 million variants with MAF greater than 1 percent. We include associations on the X chromosome, and several new classes of image-derived phenotypes (primarily, more fine-grained subcortical volumes, and cortical grey-white intensity contrast). We develop a method to identify clusters of associations across phenotypes (Peaks) and we find 692 replicating clusters of associations, including 12 on the X chromosome. Our novel associations implicate pathways involved in the rare X-linked syndrome STAR (syndactyly, telecanthus and anogenital and renal malformations), Alzheimer's disease and mitochondrial disorders.

Le projet UK Biobank est une importante étude épidémiologique prospective qui consiste à effectuer une imagerie cérébrale multimodale détaillée sur 100 000 participants, et qui inclut la génétique et les résultats de santé en cours. Nous présentons une nouvelle ressource ouverte de statistiques descriptives d'études d'association pangénomiques, résultant d'un ensemble considérablement élargi d'associations génétiques avec des phénotypes cérébraux, à l'aide de la version 2020 des données d'imagerie de la base de données UK Biobank qui contient environ 40 000 sujets, 3 000 phénotypes et 10 millions de variants ayant une fréquence d'allèle mineur supérieure à 1 %. Nous incluons des associations sur le chromosome X et plusieurs nouvelles classes de phénotypes issus d'images (principalement des volumes sous-corticaux plus fins et un contraste d'intensité gris-blanc du cortex). Nous avons mis au point une méthode permettant d'identifier les groupes d'associations entre les phénotypes (Peaks) et nous avons trouvé 692 groupes d'associations répliqués, dont 12 sur le chromosome X. Nos nouvelles associations sont en lien avec des mécanismes jouant un rôle dans le rare lien entre le chromosome X et le syndrome de syndactylie-télécanthus-malformations rénale et anogénitale (STAR), dans la maladie d'Alzheimer et dans les troubles mitochondriaux.

Chair/Président: Fangda Liu

Date: Tuesday May 31 / mardi 31 mai

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-15:45]

Ramin Eghbalzadeh (Concordia University)

A discrete-time version of the arbitrage-Free Nelson-Siegel term structure model

Version en temps discret du modèle de structure des termes Nelson-Siegel sans arbitrage

The term structure is a function that associates each maturity with a spot rate. Investigating the dynamics of the term structure is essential in the financial markets and has applications in different areas including the pricing of financial derivatives and assets, portfolio management, risk measuring, etc. The arbitrage-free Nelson-Siegel (AFNS) model is one of the most attractive term structure models since Its dynamics mimic the behavior of the celebrated and well-interpretable Nelson-Siegel model while ensuring that prices produced by the model are arbitrage-free. In this presentation, I introduce a discrete-time version of the arbitrage-free Nelson-Siegel (DTAFNS) model which not only has good fitting power and high forecasting performance but also is a theoretical rigorous term structure model. The calibration is performed with a Kalman filter.

La structure des termes est une fonction qui associe chaque échéance à un taux au comptant. L'étude de la dynamique de la structure des termes est essentielle sur les marchés financiers et a des applications dans différents domaines, notamment la fixation du prix des produits dérivés et des actifs financiers, la gestion de portefeuille, la mesure du risque, etc. Le modèle de Nelson-Siegel sans arbitrage (NSSA) est l'un des modèles de structure des termes les plus attrayants car sa dynamique imite le comportement du célèbre et bien interprétable modèle de Nelson-Siegel tout en garantissant que les prix produits par le modèle sont sans arbitrage. Dans cette présentation, j'introduis une version en temps discret du modèle de Nelson-Siegel sans arbitrage (NSSTD) qui non seulement possède un bon pouvoir d'ajustement et une haute performance de prévision mais qui est aussi un modèle de structure des termes rigoureux sur le plan théorique. L'étalonnage est effectué à l'aide d'un filtre de Kalman.

[15:45-16:00]

Mathilde Bourget (UQAM)

Statistical Modeling of Flood Risk in Climate Change

Modélisation statistique du risque d'inondation en contexte de changements climatiques

According to the latest IPCC report, extreme precipitations are more likely in the future, thus leading to a potential increase in flood occurrences over North America. Stakeholders like insurance companies will need to adapt their risk management to prepare for such changes. The following thesis will explore statistical models, such as a random forest, a generalized additive model, and a generalized linear model, for flood risk in Canada and in the United States. Model training is executed on historical flood data from 2007 to 2020 in the United States. Regional climate models are then used to predict flood probabilities in Canada and in the United States from 2020 to 2060. These models will then be used to analyze the financial impact of climate change

Le plus récent rapport du GIEC affirme que l'Amérique du Nord continuera d'observer des changements climatiques, ce qui peut mener à une hausse potentielle du risque d'inondation. Ces changements nécessiteront une adaptation au niveau de la gestion des risques des parties prenantes telles les compagnies d'assurance. L'ouvrage suivant portera sur la modélisation statistique des inondations aux États-Unis et au Canada en contexte de changement climatique. L'objectif est de valider, à l'aide de différents modèles climatiques régionaux, si la probabilité d'inondations causées par la pluie augmentera sur l'horizon 2020-2060. Les modèles proposés sont une forêt aléatoire, un modèle additif généralisé et un modèle linéaire généralisé. Ceux-ci sont entrainés à l'aide de données historiques d'inondations aux États-Unis et sont ensuite utilisés pour prédire les probabilités d'inondations au Canada et

on fictional portfolios.

aux États-Unis. Une analyse des impacts financiers sera ensuite effectuée à l'aide de portefeuilles fictifs.

[16:00-16:15]

Emma Kroell (University of Toronto) **Silvana Manuela Pesenti** (University of Toronto) **Sebastian Jaimungal** (University of Toronto)

Reverse Sensitivity Testing for Stochastic Processes

Test de sensibilité inverse des processus stochastiques

One way to quantify uncertainty in risk evaluations in a financial or actuarial setting is using sensitivity analysis, where one studies the relationship between the variability in model outputs and uncertainty in model inputs. We build on a particular type of sensitivity analysis called reverse sensitivity testing, which has recently been introduced in the literature. We generalize the reverse sensitivity framework by developing a methodology applicable to Levy-Ito processes, which proceeds as follows: First, we introduce a stress to a stochastic process by increasing a risk measure evaluated at the process's terminal time. Second, we derive the stressed probability measure under which the stochastic process fulfils the stress and that has minimal Kullback-Leibler divergence. We study the characteristics of the stochastic process under the stressed probability measure and illustrate them using numerical experiments.

Pour quantifier l'incertitude de l'évaluation des risques dans un contexte financier ou actuariel, on peut utiliser une analyse de sensibilité, par laquelle on étudie la relation entre la variabilité des sorties du modèle et l'incertitude des entrées du modèle. Nous nous basons sur un type particulier d'analyse de sensibilité appelée « test de sensibilité inverse », qui a récemment été présentée dans la littérature. Nous généralisons le cadre de sensibilité inverse en mettant au point une méthodologie applicable aux processus de Lévy–Itô. Pour ce faire, nous introduisons une contrainte dans un processus stochastique en augmentant une mesure de risque évaluée au temps final du processus. Ensuite, nous déterminons la mesure de probabilité contrainte pour laquelle le processus stochastique présente une divergence de Kullback-Leibler minimale et satisfait la contrainte. Nous examinons les caractéristiques du processus stochastique selon la mesure de probabilité contrainte et les illustrons à l'aide d'expériences numériques.

[16:15-16:30]

Spark Tseung (University of Toronto) **Tsz Chai Fung** (Georgia State University) **Ian Weng Chan** (University of Toronto) **Andrei L. Badescu** (University of Toronto) **X. Sheldon Lin** (University of Toronto)

Modelling Heterogeneous Risks with Random Effects in the Mixture-of-Experts Model

Modélisation des risques hétérogènes à l'aide d'effets aléatoires dans le modèle de mélange d'experts

In statistical applications, mixed (or random effects) models are often used for modelling unobserved effects. Several restrictive assumptions of the classical (generalized) linear mixed models have rendered them unsuitable for insurance data, which typically exhibit multimodality and rather different body and tail. In this talk, we present an extension to a class of the mixture-of-experts model by incorporating random effects. This non-trivial extension preserves the desirable property of denseness, mathematical tractability and interpretability. Besides, the addition of random effects accounts for unobserved effects such as heterogeneous risks without over-complicating the model with too many parameters. Estimation of model parameters and realizations of random effects can be accomplished by a modification of the Expected-Conditional-Maximization algorithm. Finally, we present numerical simulations and case studies

Dans les applications statistiques, les modèles mixtes (ou à effets aléatoires) sont souvent utilisés pour modéliser les effets non observés. Plusieurs hypothèses restrictives des modèles linéaires mixtes (généralisés) classiques les rendent inadaptés aux données d'assurance, qui présentent en général une multimodalité, ainsi qu'une distribution avec un corps et une queue assez différents. Dans cette présentation, nous présentons une extension à une classe du modèle de mélange d'experts par l'insertion d'effets aléatoires. Cette extension non triviale préserve les propriétés souhaitables de densité, de tractabilité et d'interprétabilité mathématiques. De plus, l'ajout d'effets aléatoires permet de tenir compte des effets non observés, tels que les risques hétérogènes, sans compliquer excessivement le modèle avec un trop grand nombre de paramètres. L'estimation des paramètres du modèle et les effets aléatoires peuvent être réalisés par une modification de l'algorithme d'espérance-maximisation conditionnelle. Finalement, nous présentons des simulations numériques et des études

on real insurance datasets.

de cas sur des ensembles de données d'assurance réelles.

[16:30-16:45]

Liyuan Lin (University of Waterloo) **Hirbod Assa** (Kent Business School) **Ruodu Wang** (University of Waterloo)

On technical properties and calibrations of PELVE

Propriétés techniques et calages du PELVE

Recently Li and Wang (2022) introduced a new risk measure called Probability Equivalent Level of VaR-ES (PELVE) to derive the equivalent probability level when replacing ES with VaR. In this paper, we take a closer look at the properties of PELVE and explore more about a random variable and its PELVE. First, we study the monotonicity and convergence properties of PELVE. We present sufficient conditions that show these properties are closely related to the hazard rate function of a random variable, which makes us able to identify propertied of PELVE for important examples. Second, we study a PELVE calibration problem where we set to find a distribution that yields a given PELVE. We find that a general solution to the calibration problem can be represented as the answer to an advanced differential equation, that makes us able to find general numerical solutions, and also characterize all the distributions that have a constant PELVE.

Récemment, Li et Wang (2022) ont présenté une nouvelle mesure de risque appelée niveau équivalent de probabilité (PELVE) de la valeur à risque (VaR) et du déficit attendu (ES) pour obtenir le niveau de probabilité équivalent lorsqu'on remplace le déficit attendu par la valeur à risque. Dans cette présentation, nous nous penchons sur les propriétés du PELVE et nous explorons plus en détail une variable aléatoire et le PELVE de celle-ci. Tout d'abord, nous examinons les propriétés de monotonicité et de convergence du PELVE. Nous présentons des conclusions qui montrent que ces propriétés sont étroitement liées à la fonction du taux de hasard d'une variable aléatoire, ce qui nous permet de déterminer les propriétés du PELVE pour des exemples significatifs. Ensuite, nous analysons un problème de calage du PELVE pour lequel nous cherchons à trouver une distribution qui produit un certain PELVE. Nous constatons qu'une solution générale au problème de calage peut être représentée comme la réponse à une équation différentielle avancée, ce qui nous permet de trouver des solutions numériques générales, et aussi de caractériser toutes les distributions qui ont un PELVE constant.

[16:45-17:00]

Zhenzhen Huang (University of Waterloo) **Pengyu Wei** (Nanyang Technological University) **Chengguo Weng** (University of Waterloo)

Statistical Classification Methods for the Combining Portfolio Strategy

Méthodes de classification statistique pour la stratégie de combinaison de portfolios

Due to the well-known parameter uncertainty problem in modern portfolio theory, the combining portfolio strategy utilizes the 1/N rule as a shrinkage point to improve the performance of a sophisticated portfolio strategy, where the optimal combining coefficient is determined under the normal assumption for asset returns. To generalize the combining portfolio strategy without the normal assumption, we propose a statistical classification framework that uses logistic regression or random forest to find the combining coefficient according to prudently selected market features. Empirical studies with many market datasets show that our methods constantly generate better out-of-sample Sharpe ratio and comparable out-of-sample expected utility in most scenarios. Meanwhile, our methods yield a classification model with transparent interpretability of important features for evaluating the combining coefficient.

En raison de problèmes bien connus d'incertitude de paramètres dans la théorie de portfolio moderne, la stratégie de combinaison de portfolios utilise la règle 1/N en guise de point de rétrécissement pour améliorer la performance d'une stratégie de portfolio sophistiquée, où le coefficient de combinaison optimal est déterminé selon l'hypothèse normale des retours d'actif. Afin de généraliser les stratégies de combinaison de portfolios en évitant l'hypothèse de normalité, nous proposons une cadre de classification statistique qui emploie une régression logistique de forêts aléatoires pour trouver le coefficient de combinaison selon les caractéristiques de marché soigneusement sélectionnées. Des études empiriques avec des jeux de données du marché démontrent que notre méthode génère continuellement un taux Sharpe hors de l'échantillon supérieur et une utilité prévue hors de l'échantillon similaire dans la plupart des cas. De plus, notre méthode procure un modèle de classification avec possibilité d'interprétation transparente de caractéristiques importantes pour évaluer le coefficient

de combinaison.

Control Chart and Statistical Methods for Clinical Trials
Carte de contrôle et méthodes statistiques pour les essais cliniques

Chair/Président: Luke Hagar

Date: Tuesday May 31 / mardi 31 mai

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-15:45]

Armando Turchetta (McGill University) **Nicolas Savy** (Institut de Mathématiques de Toulouse) **Erica E.M. Moodie** (McGill University) **David A. Stephens** (McGill University)

A Time-Dependent Poisson-Gamma Model for Recruitment Forecasting in Multicenter Clinical Trials

Modèle Poisson-Gamma dépendant du temps pour la prévision du recrutement dans les essais cliniques multicentriques

Estimating the recruitment time in multicenter clinical trials is a key component of the feasibility assessment. Yet, deterministic models mainly based on trial investigators' recruitment assumptions are still used. A Bayesian approach built on a doubly stochastic Poisson process, known as the Poisson-Gamma model, was introduced to address the lack of a strong and consistent statistical methodology in this field. This approach is based on the modeling of enrollments as a Poisson process where the recruitment rates are assumed to be constant over time and to follow a common Gamma prior distribution. However, the constant-rate assumption is a restrictive limitation that is rarely appropriate for applications in real clinical trials. In this presentation, we illustrate a flexible generalization of this methodology which allows the enrollment rates to vary over time by modeling them through B-splines, and we show the suitability of this approach for a wide range of recruitment behaviors.

L'estimation du temps de recrutement dans les essais cliniques multicentriques est un élément clé de l'évaluation de la faisabilité. Pourtant, on utilise encore des modèles déterministes basés principalement sur les hypothèses de recrutement des investigateurs de l'essai. Une approche bayésienne construite sur un processus de Poisson doublement stochastique, connue sous le nom de modèle Poisson-Gamma, a été introduite pour remédier à l'absence d'une méthodologie statistique solide et cohérente dans ce domaine. Cette approche est basée sur la modélisation des inscriptions comme un processus de Poisson où les taux de recrutement sont supposés être constants dans le temps et suivre une distribution de probabilités a priori Gamma commune. Cependant, l'hypothèse de taux constants est une limitation restrictive qui est rarement appropriée pour les applications aux essais cliniques réels. Dans cette présentation, nous illustrons une généralisation souple de cette méthode qui permet une variation dans le temps des taux de recrutement en les modélisant par des B-splines ; puis nous montrons la pertinence de cette approche pour un large éventail de comportements de recrutement.

[15:45-16:00]

Fatemeh Mahmoudi (University of Calgary) **Xuewen Lu** (University of Calgary)

Variable Selection in Semiparametric Shared Frailty Illness-death Models for Semi-competing Risks Data

Sélection des variables dans des modèles maladie-décès semi-paramétriques à fragilité partagée pour des données de risques semi-concurrents

Semi-competing risks data arise when both non-terminal and terminal events are considered in a model. In this framework, terminal event can censor the non-terminal event, but not vice versa. It is known that variable selection is practical in identifying significant risk factors in high-dimensional data while some recent works on penalized variable selection deal with these competing risks separately without incorporating possible dependence between them. We perform variable selection in

Les données de risques semi-concurrents sont présentes lorsqu'un modèle prend en compte à la fois les événements non-terminaux et terminaux. Dans un tel cadre, un événement terminal peut censurer l'événement non terminal, mais pas l'inverse. On sait que la sélection des variables est pratique pour identifier des facteurs de risque importants dans les données à haute dimension, tandis que des études récentes sur la sélection de variables pénalisées prend en compte séparément ces risques concurrents sans incorporer de dépendance possible entre eux. Nous procédons à une sélection

an illness-death model using shared frailty where semi-parametric hazard regression models are used to model the effect of covariates. We propose a broken adaptive ridge (BAR) penalty to encourage sparsity and conduct extensive simulation studies to compare its performance with other methods. The oracle property and the grouping effect of the proposed BAR procedure are also investigated using simulation studies. The proposed method is then applied to the real-life data arising from a Colon Cancer study.

des variables d'un modèle maladie-décès en utilisant une fragilité partagée pour laquelle nous utilisons des modèles de régression des risques semi-paramétriques pour modéliser l'effet des covariables. Nous proposons une pénalité de crête adaptative brisée (BAR) pour favoriser la dispersion et pour mener des études en simulation approfondies afin de comparer sa performance à celle d'autres méthodes. À l'aide d'études en simulation, nous examinons la propriété oracle et l'effet de regroupement de la procédure BAR proposée dont nous verrons aussi l'application à des données réelles tirées d'une étude sur le cancer du côlon.

[16:00-16:15]

Qi Lyu (University of Regina)

Modified Economic Model of Hotelling's T^2 Control Chart with Variable Sampling Interval

Modèle économique modifié de la carte de contrôle T^2 d'Hotelling avec intervalle d'échantillonnage variable

In quality control field, the cost parameters usually implement constants in traditional economic model. In order to decrease the cost of economic statistical design and acquire more accurate results, we apply the modified economic model with Taguchi's loss function on Hotelling's T^2 control chart. Variable Sampling Interval scheme is utilized as the sampling method. With the help of Taguchi's loss function for multivariate quality characteristic, a clear expression of the cost coefficient matrix is derived. This is a part of my joint research with Dr. M. Tavakoli (Birjand University of Technology, Iran), Dr. R. Pourtaheri (Allameh Tabataba'i University, Iran), and A. Volodin (University of Regina, Canada). Keywords: Economic model, Taguchi's loss function, Hotelling's T^2 control chart, Variable Sampling Interval, Artificial bee colony algorithm, Monte Carlo method

Dans le domaine du contrôle de la qualité, les paramètres de coût intègrent généralement des constantes dans le modèle économique traditionnel. Afin de réduire le coût du plan statistique économique et d'obtenir des résultats précis, nous appliquons le modèle économique modifié avec fonction de perte de Taguchi à la carte de contrôle T^2 d'Hotelling. On utilise un schéma d'intervalle d'échantillonnage variable en guise de méthode d'échantillonnage. À l'aide de la fonction de perte de Taguchi pour la caractéristique de qualité multivariée, on peut dériver une expression claire de la matrice à coefficient de coût. Ceci fait partie de mon programme conjoint de recherche avec M. Tavakoli (Birjand University of Technology, Iran) R. Pourtaheri (Allameh Tabataba'i University, Iran) et A. Volodin (Université de Regina, Canada). Mots clés : Modèle économique, fonction de perte de Taguchi, carte de contrôle T^2 d'Hotelling, intervalle de sondage de variable, algorithme de colonie d'abeilles artificielles, méthode Monte Carlo

[16:15-16:30]

Apsara Pathum Jayasooriya (Memorial University of Newfoundland) **Asokan Mulayath Variyath** (Memorial University of Newfoundland) **Yanqing Yi** (Memorial University of Newfoundland)

Statistical Inference for Multiple Stage Randomized Clinical Trials with Binary Responses

Inférence statistique pour des essais cliniques randomisés en plusieurs étapes avec des réponses binaires

This study focuses on computing the critical values, type I error rate and the power of multiple-stage randomized clinical trials with binary responses. Different methods such as Pocock, Peto, O'Brien & Fleming are used to construct the alpha spending functions in order to control the overall type I error rate. Critical values are obtained using an iterative Markov chain approach to satisfy the alpha spending at each stage. Considering the discrete nature of the data, a likelihood ratio test statistic is used to calculate the type I error rate and the sta-

Cette étude s'intéresse au calcul des valeurs critiques, du taux d'erreur de type I et de la puissance des essais cliniques randomisés en plusieurs étapes avec des réponses binaires. Diverses méthodes comme Pocock, Peto et O'Brien & Fleming sont utilisées pour la construction des fonctions de dépense du risque alpha afin de contrôler le taux d'erreur de type I global. Les valeurs critiques sont obtenues à l'aide d'une approche par chaînes de Markov satisfaisant la dépense du risque alpha à chaque étape. En prenant en compte la nature discrète des données, une statistique du test du rapport de vraisemblance est utilisée pour calculer le taux d'erreur

tistical power, and the alpha spending function methods are compared. Results are obtained for testing one sample and two samples, both cases with three stages. However, the method is generalized for any number of stages. Keywords: type I error, likelihood ratio test, group sequential analysis, alpha spending function

de type I de même que la puissance statistique et les méthodes avec fonctions de dépense du risque alpha sont comparées. Des résultats de tests avec un échantillon et deux échantillons sont obtenus, en trois étapes dans les deux cas. La méthode est cependant généralisée pour un nombre quelconque d'étapes. Mots-clés : erreur de type I, test du rapport de vraisemblance, analyse séquentielle de groupe, fonction de dépense du risque alpha

[16:30-16:45]

Junwei Shen (McGill University) **Shirin Golchi** (McGill University) **Erica E.M. Moodie** (McGill University) **David Benrimoh** (McGill University, Aifred Health)

New designs for Bayesian adaptive cluster-randomized trials

Nouveaux plans pour les essais adaptatifs bayésiens randomisés en grappes

Adaptive approaches, allowing for more flexible trial design, have been proposed for individually randomized trials to save time or reduce sample size. However, adaptive designs for cluster-randomized trials in which groups of participants are randomized to treatment arms are less common. Motivated by a potential real-world cluster-randomized trial, two Bayesian adaptive designs for cluster-randomized trials are proposed to allow for early stopping for efficacy at pre-planned interim analyses. The difference between the two designs lies in the way that participants are sequentially recruited. The design operating characteristics are explored via simulations for a variety of scenarios and two outcome types for the two designs. The simulation results show that for different outcomes the design choice may be different. We make recommendations for designs of Bayesian adaptive cluster-randomized trial based on the simulation results.

On a proposé des approches adaptatives (permettant des plans d'essai plus souples) pour les essais randomisés individuels afin de gagner du temps ou de réduire la taille de l'échantillon. Cependant, les plans d'essais adaptatifs pour les essais randomisés en grappes, dans lesquels des groupes de participants sont randomisés dans des bras de traitement, sont moins courants. Sur la base d'un essai randomisé en grappes potentiel réalisé dans des conditions réelles, nous proposons deux plans adaptatifs bayésiens pour les essais randomisés en grappes afin que les analyses intermédiaires planifiées à l'avance puissent être arrêtées rapidement pour en garantir l'efficacité. La différence entre les deux modèles réside dans la manière dont les participants sont successivement recrutés. Nous examinons les caractéristiques de fonctionnement des plans par des simulations pour divers scénarios et deux types de résultats pour les deux plans. Les résultats des simulations montrent que pour divers résultats, le choix du plan peut être différent. Sur la base des résultats des simulations, nous formulons des recommandations pour les plans d'essais randomisés en grappes adaptatifs bayésiens.

[16:45-17:00]

Hira Nadeem (University of Regina)

Special Case of Direct-Inverse Sampling Scheme for the Cross Product Ratio - Participant Enrollment in Clinical Trials

Cas spécial du plan d'échantillonnage direct inversé pour le rapport de produits croisés – recrutement de participants à des essais cliniques

A successful clinical trial requires efficient strategies for enrolling and retaining the study participants. However, experts of clinical trials worldwide find it extremely challenging to recruit and retain participants. Therefore, this paper focuses on the simple idea of participant enrollment using Bernoulli samples. The participants are enrolled in the clinical trials using the Direct or Inverse Binomial sampling schemes and the point estimate for the cross-product ratio ρ=(p1(1-p2))/(p2(1-p1)) is calculated. Prior studies in this domain indicate that

Un essai clinique réussi requiert des stratégies efficaces pour recruter et retenir les participants dans l'étude. Cependant, les experts des essais cliniques du monde entier trouvent qu'il est extrêmement difficile de recruter et de retenir les participants. Ainsi, cet article porte sur le concept simple de l'inscription des participants à l'aide d'échantillons de Bernoulli. On recrute les participants aux essais cliniques à l'aide de plans d'échantillonnage binomiaux directs ou inversés et on calcule l'estimation ponctuelle du rapport des produits croisés ρ=(p1(1-p2))/(p2(1-p1)). Des études antérieures dans ce domaine indiquent

the special case of the Direct-Inverse sampling scheme works the best, where the number of successes in the Direct sampling scheme is used in the second sampling scheme of the Inverse binomial scheme. Asymptotic confidence intervals are constructed. Monte-Carlo method is used to investigate the key probability characteristics of intervals. CYP-GUIDES case study is discussed to determine the estimate of ρ for the special case.

que le cas particulier du schéma d'échantillonnage direct inversé est celui qui fonctionne le mieux lorsque le nombre de succès du plan d'échantillonnage direct est utilisé dans le deuxième plan d'échantillonnage du plan binomial inversé. On crée des intervalles de confiance asymptotiques, puis on utilise la méthode de Monte-Carlo pour étudier les principales probabilités caractéristiques des intervalles. Enfin, on se penche sur l'étude de cas CYP-GUIDES (génotypage des cytochromes psychotropes dans le cadre de l'enquête pour l'aide à la décision) afin de déterminer l'estimation de ρ pour le cas spécial.

Chair/Président: Hon-Yiu So

Date: Tuesday May 31 / mardi 31 mai

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-15:45]

Alex Stringer (University of Waterloo)

New Results in Modelling Dependent Data

Nouveaux résultats dans la modélisation des données dépendantes

We discuss models for dependent data where the intractable marginal likelihood is approximated using adaptive (Gaussian) quadrature. The integration error is quantified for GLMMs with "regular" response and random effects distributions, beyond the exponential family and Gaussian cases. The integration and statistical errors are balanced leading to a firm recommendation for how many quadrature points to use, recovering the results of ten previous simulation studies and data analyses. Results are applicable to the modelling of dependent binary and survival data. Based on joint work with Blair Bilodeau.

Nous examinons les modèles pour les données dépendantes dans lesquels la vraisemblance marginale intraitable est approximée avec la quadrature (gaussienne) adaptative. L'erreur d'intégration est quantifiée pour les modèles linéaires généralisés mixtes avec des distributions de réponses et d'effets aléatoires « régulières », qui ne se limitent pas à la famille exponentielle ni aux cas gaussiens. Les erreurs d'intégration et statistiques sont équilibrées. Il en résulte ainsi une solide recommandation quant au nombre de points de quadrature à utiliser par la récupération des résultats de dix études de simulation et d'analyses de données antérieures. Les résultats sont applicables à la modélisation de données binaires dépendantes et de survie. Il s'agit de travaux conjoints avec Blair Bilodeau.

[15:45-16:00]

Glen McGee (University of Waterloo) **Alex Stringer** (University of Waterloo)

Flexible Marginal Models for Dependent Data

Modèles marginaux flexibles pour données dépendantes

Models for dependent data are distinguished by their targets of inference: in particular, marginal models are useful when interest lies in quantifying associations averaged across a population of clusters. Moreover, when the functional form of covariate-outcome associations is unknown, more flexible regression methods are needed to allow for potentially non-linear relationships. We propose a novel marginal additive model (MAM) for modelling cluster-correlated data with non-linear population-averaged associations. The proposed methods allow for estimation, uncertainty quantification, characterization of variability across clusters, as well as cluster-specific prediction, all within a unified likelihood framework. We further introduce a fitting algorithm that allows for efficient calculation of standard errors.

Les modèles pour données dépendantes se distinguent par leurs cibles d'inférence : en particulier, les modèles marginaux sont utiles lorsque l'intérêt réside dans la quantification d'associations moyennées sur une population de grappes. De plus, lorsque la forme fonctionnelle des associations covariable-résultat est inconnue, des méthodes de régression plus souples sont nécessaires pour permettre des relations potentiellement non linéaires. Nous proposons un nouveau modèle additif marginal (MAM) pour modéliser les données corrélées par grappes avec des associations non linéaires moyennées sur la population. Les méthodes proposées permettent l'estimation, la quantification de l'incertitude, la caractérisation de la variabilité entre les grappes, ainsi que la prédiction spécifique aux grappes, le tout dans un cadre de vraisemblance unifié. Nous introduisons également un algorithme d'ajustement qui permet un calcul efficace des erreurs types.

[16:00-16:15]

Mohamad Elmasri (McGill University) **Aurélie Labbe** (HEC Montreal) **Denis Larocque** (HEC Montreal) **Laurent Charlin** (HEC Montreal)

Predictive Inference for Travel Time on Transportation Networks

Inférence prédictive pour le temps de trajet sur les réseaux de transport

Recent statistical methods fitted on large-scale GPS data are getting close to answering the proverbial question "Are we there yet?" Most current methods focus on predicting the expected travel time. Little is known about its distribution, which is key for decision-making and downstream applications. We develop a novel statistical approach to this problem, where we show that, under general conditions, without assuming a distribution of speed, travel time divided by route distance follows a Gaussian distribution with route-invariant population mean and variance. We develop efficient inference methods for such parameters and propose asymptotically tight population prediction intervals for travel time. Using road-level information (e.g. traffic density), we further develop a trip-specific Gaussian-based predictive distribution, resulting in tight prediction intervals for short and long trips. Compared to alternative approaches, our trip-specific predictive distribution achieves (a) the theoretical coverage at every level of significance, (b) tighter prediction intervals, (c) less predictive bias, and (d) more efficient estimation and prediction procedures. This makes our approach promising for low latency large-scale transportation applications.

De récentes méthodes statistiques ajustées sur des données GPS à grande échelle sont très près de répondre à la question proverbiale « On est bientôt arrivés ? ». La plupart des méthodes actuelles se concentrent sur la prédiction du temps de trajet prévu. On sait peu de choses sur sa distribution, qui est pourtant essentielle pour la prise de décision et les applications en aval. Nous développons une nouvelle approche statistique de ce problème, dans laquelle nous montrons que, dans des conditions générales, sans supposer une distribution de la vitesse, le temps de trajet divisé par la distance de l'itinéraire suit une distribution gaussienne avec une moyenne et une variance de population invariante de l'itinéraire. Nous développons des méthodes d'inférence efficaces pour ces paramètres et proposons des intervalles de prédiction de population asymptotiquement serrés pour le temps de trajet. En utilisant des informations au niveau de la route (par exemple, densité du trafic), nous développons une distribution prédictive gaussienne spécifique au trajet, ce qui permet d'obtenir des intervalles de prédiction serrés pour les trajets courts et longs. Par rapport à d'autres approches, notre distribution prédictive spécifique aux trajets atteint (a) la couverture théorique à chaque niveau de signification, (b) des intervalles de prédiction plus étroits, (c) moins de biais de prédiction et (d) des procédures d'estimation et de prédiction plus efficaces. Cela rend notre approche prometteuse pour les applications de transport à grande échelle à faible latence.

[16:15-16:30]

Kexin Luo (Western University) **Myriam Brossard** (Lunenfeld-Tanenbaum Research Institute, Sinai Health) **Shelley B. Bull** (Lunenfeld-Tanenbaum Research Institute, Sinai Health; Dalla Lana School of Public Health, University of Toronto)

Estimation of Genome-wide Significance Thresholds for Multi-variant Region-level Genetic Association Testing of Complex Traits

Estimation des niveaux de signification pangénomiques pour tester l'association de régions génétiques à multi-variant pour des caractères complexes

Multiple testing criteria for genome-wide family-wise error control are well-established for single-variant association test statistics, but less so for multi-variant region-based tests. Although region tests can reduce multiple test burden and better capture signals under complex genetic architectures, within- and between-region correlations and variant frequencies may affect the significance threshold. We develop permutation methods to estimate empirical thresholds for region statistics and apply them to compare the performance of region and single-variant association tests in lipid traits from Canadian Longitudinal Study on Aging. We

Les critères de tests multiples pour le contrôle du taux d'erreur familial pangénomique sont bien établis pour les statistiques de tests d'association à variant unique, mais le sont beaucoup moins pour les tests de régions génétiques à multi-variant. Même si les tests au niveau des régions peuvent réduire le fardeau associé aux tests multiples et mieux saisir les signaux sous des architectures génétiques complexes, les corrélations à l'intérieur et entre les régions ainsi que la fréquence des variants peuvent affecter le niveau de signification. Nous développons des méthodes de permutation pour estimer les niveaux empiriques des statistiques au niveau des régions et nous les appliquons afin de comparer la performance des tests d'association à variant unique au niveau

partition the genome in two ways: quasi-independent regions based on linkage disequilibrium versus coding gene regions alone, consider two levels of minor allele frequency filtering, and test for association in each region. Application of empirical genome-wide significance thresholds at fixed family-wise error reveals regions missed by single tests.

des régions pour des caractéristiques lipidiques tirées de l'Étude longitudinale canadienne sur le vieillissement. Nous partitionnons le génome de deux façons : par régions quasi-indépendantes basées sur le déséquilibre de liaison et par régions de codage génétique seulement, tout en considérant deux niveaux de filtrage de la fréquence de l'allèle mineur et en testant l'association dans chaque région. L'application de niveaux de signification empiriques pangénomiques avec un taux d'erreur familial fixe révèle des régions que des tests individuels ont manqué.

[16:30-16:45]

Kevin Granville (University of Western Ontario) **Douglas G. Woolford** (University of Western Ontario) **Charmaine B. Dean** (University of Waterloo) **Colin B. McFayden** (Ontario Ministry of Northern Development, Mines, Natural Resources and Forestry, Aviation, Forest Fire and Emergency Services) **Den Boychuk** (Ontario Ministry of Northern Development, Mines, Natural Resources and Forestry, Aviation, Forest Fire and Emergency Services)

On the Selection of an Interpolation Method with an Application to the Fire Weather Index in Ontario, Canada

Sélection d'une méthode d'interpolation avec application à l'Indice forêt-météo en Ontario, au Canada

Studies in the environmental sciences frequently rely on the presence of spatially dense climatological data. However, such data are often available only at a fixed set of locations across a region. Interpolation enables the approximation of variables of interest at locations between those sites. When collaborating with an end user, mutual knowledge exchange allows for greater insight on what is required of an interpolation method since each method may have different pros and cons. We discuss several key considerations one should make in an interpolation study, such as the purpose of the variable and the goals of the end user, including how the variable is used to inform decisions. We illustrate in a case study of the province of Ontario, Canada, our considerations when contrasting several methods with the goal of interpolating the Fire Weather Index. This work is in collaboration with the Ontario Ministry of Northern Development, Mines, Natural Resources and Forestry.

Les études en sciences de l'environnement s'appuient souvent sur la présence de données climatologiques spatiales densément observées. De telles données sont cependant souvent disponibles seulement pour un ensemble fixe de sites dans une région. L'interpolation permet l'approximation des variables d'intérêt dans des emplacements entre ces sites. En collaborant avec un utilisateur final, l'échange mutuel de connaissances permet une meilleure compréhension de ce qui est requis d'une méthode d'interpolation, puisque chaque méthode comporte des avantages et des désavantages. Nous abordons plusieurs points importants à considérer dans une étude d'interpolation, comme le but de la variable et les objectifs de l'utilisateur final, y compris le mode d'utilisation de la variable pour éclairer les décisions. Une étude de cas dans la province de l'Ontario au Canada sert à illustrer nos considérations lorsque nous comparons plusieurs méthodes dans un but d'interpolation de l'Indice forêt-météo. Ce travail est fait en collaboration avec le Ministère des Richesses naturelles de l'Ontario.

[16:45-17:00]

François A. Marshall (Boston Univeristy)

Inferring Driver Nonlinearity in Physical Systems using Cyclostationary Signal Processing

Déduire la nonlinéarité du moteur dans les systèmes physiques à l'aide du traitement du signal cyclostationnaire

Even when the stochastic driver of a dynamical system is known (e.g., Earth's rotation in the presence of extraneous noise), specifying the nonlinear response of a detection system can be challenging. Often, the fixed-lag autocorrelation functions of the observable process each exhibit harmonic periodicity. i.e., the process is said to be cyclostationary, and characterized by correlation between some of its frequency-offset spectral jumps. De-

Même lorsque le moteur stochastique d'un système dynamique est connu (par exemple, la rotation de la Terre en présence de bruits étrangers), la spécification de la réponse nonlinéaire d'un système de détection peut être difficile. Souvent, les fonctions d'autocorrélation à retard fixe du processus observable présentent chacune une périodicité harmonique. C'est-à-dire que le processus est dit cyclostationnaire et caractérisé par une corrélation entre certains de ses sauts spectraux de retard de fréquence. Décrire la non-

scribing nonlinearity by coupling between certain driver states provides an easier interpretation of the system dynamics than that obtained only considering said spectral correlations of the observable process. To this end, a two-layer latent-space model is used to explain the spectral jumps of the observable process, from which can be inferred the nonlinear coupling events. In an analysis of a cyclostationary process driven by Earth's rotational modes, spectral correlations are identified using a novel, high-power, likelihood-ratio test.

linéarité par couplage entre certains états du moteur permet une interprétation plus aisée de la dynamique du système que celle obtenue en ne considérant que lesdites corrélations spectrales du processus observable. À cet effet, un modèle d'espace latent à deux couches est utilisé pour expliquer les sauts spectraux du processus observable, à partir desquels peuvent être déduits les événements de couplage nonlinéaires. Dans une analyse d'un processus cyclostationnaire entraîné par les modes de rotation de la Terre, les corrélations spectrales sont identifiées à l'aide d'un nouveau test rapport des maximums de vraisemblance à haute puissance.

2021 SSC Impact Award Address
Allocution du récipiendaire du prix pour impact de la SSC 2021

Chair/Président: Tolu Sajobi

Date: Wednesday June 1 / mercredi 1 juin

Time/Heure: 11:00-12:15

Abstract/Résumé

[11:00-12:00]

Thérèse A. Stukel (ICES/ University of Toronto)

Innovative Uses of Health Administrative Data for Health Policy Research

Utilisation innovatrice de données administratives de santé pour la recherche sur les politiques en santé

The emergence of Big Data has been a catalyst for the world of Data Science. Novel statistical and computer science methods hold promise for extracting information from new data sources. ICES holds one of the largest repositories of health administrative data from physician and hospital billing records to unstructured clinical data from electronic medical records.We illustrate innovative linkages of administrative data to inform health policy applications ranging from Ontario Local Health Integration Networks (LHIN) boundaries to Multispecialty Physician Networks, on which Ontario Health Teams (OHTs) were modelled. We demonstrate the use of instrumental variables to remove unmeasured confounding, survival bias and reverse causality in health care studies, one in the US showing that higher healthcare spending did not lead to better outcomes, and a parallel study in Ontario showing that higher-spending hospitals had better outcomes. Finally, we illustrate the use of statistical learning methods to predict high need, high cost users of the health care system.

L'émergence de mégadonnées a été un catalyseur pour le monde de la science des données. Les nouvelles méthodes scientifiques en statistique et en informatique promettent d'extraire de l'information à partir de nouvelles sources de données. ICES contient l'un des plus grands dépôts de données administratives de santé, comprenant entre autres des registres de facturation de médecins et d'hôpitaux et des données cliniques non structurées tirées de dossiers médicaux électroniques. Nous illustrons les liens innovateurs des données administratives pour informer des applications de politiques en santé comme les réseaux locaux d'intégration des services de santé de l'Ontario (RLISS) et le réseau de médecins multispécialistes, à partir desquels les équipes Santé Ontario (OHT) ont été conçues. Nous démontrons l'utilisation de variables pour éliminer les confondants non mesurés, les biais de survie et la causalité inversée dans les études sur les soins de santé. Par exemple, l'une de ces études aux États-Unis a montré qu'une hausse de dépense en soins de santé ne mène pas à de meilleurs résultats, tandis qu'une étude parallèle en Ontario indique le contraire. Enfin, nous illustrons l'utilisation de méthodes d'apprentissage statistique pour prédire le nombre d'utilisateurs à besoins et à coûts élevés dans le système de soins de santé.

Input Privacy Preserving Technologies for Official Statistics
Technologies de préservation de la confidentialité des données pour les statistiques officielles

Chair/Président: Abel C. Dasylva

Organizer/Responsable: Abel C. Dasylva

Date: Wednesday June 1 / mercredi 1 juin

Time/Heure: 11:00-12:30

Abstract/Résumé

[11:00-11:30]

Teresa Scassa (University of Ottawa)

Legal Dimensions of Privacy Preserving Technologies for Official Statistics

Dimensions légales de technologies protégeant la confidentialité pour les statistiques officielles

National Statistics Offices are increasing looking towards administrative data as a source for national statistics. The use of such data can raise significant privacy concerns. The use of input privacy-preserving technologies is being studied domestically and international as a means of enabling the use of administrative data while also protecting privacy. This presentation explores whether these technologies adequately respond to privacy concerns and whether they raise any new privacy considerations. It examines whether existing Canadian legislation presents any barriers to the adoption or use of these technologies to facilitate the use of administrative data. It also considers whether there are other legal/policy considerations that need to be taken into account in this context.

Les organismes nationaux de statistiques cherchent de plus en plus à adopter les données administratives comme source de statistiques nationales. Cependant, l'utilisation de telles données peut causer des inquiétudes significatives concernant la confidentialité. C'est pourquoi l'utilisation de technologies protégeant la confidentialité est étudiée tant à l'intérieur du pays qu'à l'extérieur afin de permettre l'exploitation des données administrative tout en conservant la confidentialité. Cette présentation cherche à déterminer si ces technologies répondent adéquatement aux problèmes de confidentialité et si elles soulèvent de nouvelles préoccupations relatives à la confidentialité. Elle s'interroge sur les barrières que pourrait représenter la législation canadienne suite à l'adoption ou l'utilisation de ces technologies pour faciliter l'exploitation de données administratives. Elle aborde aussi les préoccupations juridiques et politiques à considérer dans ce contexte.

[11:30-12:00]

Saeid Molladavoudi (Statistics Canada)

Privacy Enhancing Technologies at Statistics Canada

Technologies d'amélioration de la confidentialité à Statistique Canada

Privacy Enhancing Technologies (PET) are an emerging class of technologies with a promise to protect the privacy and confidentiality of data throughout its lifecycle, while maintaining its utility. PETs provide Statistical Offices opportunities to facilitate collaborative analytics on less-accessible data to derive valuable insights. Statistics Canada has started experimenting with PETs a few years ago. To this end, multiple research projects have successfully been completed, such as the application of homomorphic encryption on training a machine learning classifier, privacy preserving record linkage with secure Multi-Party Computation and ap-

Les technologies d'amélioration de la confidentialité (TAC) représentent une nouvelle catégorie de technologies prometteuses pour la protection de la vie privée et la confidentialité des données tout au long du cycle de vie de ces dernières, tout en conservant leur utilité. Les technologies d'amélioration de la confidentialité permettent aux bureaux de statistique de faciliter l'analyse collaborative de données moins accessibles afin d'en tirer de précieux renseignements. Statistique Canada a commencé à expérimenter les TAC il y a quelques années. À cette fin, de nombreux projets de recherche ont été menés à bien, tels que l'application du chiffrement homomorphe à la formation d'un classificateur d'apprentissage automatique, le couplage d'enregistrements préservant la

plying Federated Learning in the context of privacy preserving crowd sourcing. The agency is also actively participating in international working groups, e.g. the UN Global Task Team and the Input Privacy Preserving project within the High-Level Group for Modernization of Official Statistics. In this presentation, we will discuss some of these activities and share insights on potential opportunities and challenges of adopting PETs in the Official Statistics.

confidentialité avec le calcul multipartite sécurisé et l'application de l'apprentissage fédéré dans le contexte de la production participative préservant la confidentialité. L'agence participe aussi activement à des groupes de travail internationaux, comme le Groupe de travail mondial des Nations Unies et le Groupe de haut niveau sur la modernisation des statistiques officielles (dans le cadre du projet de préservation de la confidentialité des données). Dans cette présentation, nous présenterons certaines de ces activités et nous échangerons nos points de vue sur les possibilités et les défis potentiels de l'adoption des TAC dans les statistiques officielles.

[12:00-12:30]

Jerome Reiter (Duke University) **Chengxin Yang** (Duke University)

Formally Private Verification of Statistical Analyses

Vérification formellement privée d'analyses statistiques

I present formally private algorithms for replication analyses. To provide context, suppose a researcher seeks to assess whether published results change substantially when certain data points (e.g., outliers) are included or excluded from the analysis, or when different models (e.g., with or without log transformations) are estimated on the data. Releasing the results of such sensitivity analyses could leak information about the confidential data. I present replication measures that allow researchers to bound this information loss. The replication measures rely on the sub-sample and aggregate method from the literature on differential privacy, in which we (1) split the data into disjoint subsets, (2) compute some measure summarizing the difference between the published and alternative analysis on each subset, (3) aggregate these subset estimates, and (4) add noise to the aggregated value to satisfy differential privacy. I illustrate the methods with empirical studies.

Je présente des algorithmes formellement privés pour les analyses de réplication. En guise de mise en contexte, supposons qu'un chercheur veuille évaluer si les résultats publiés varient considérablement lorsque certains points de données (p. ex., des valeurs aberrantes) sont inclus ou exclus de l'analyse, ou si différents modèles (p. ex., avec ou sans transformations logarithmiques) sont estimés à partir des données. La sortie de résultats provenant de telles analyses de sensibilité pourrait divulguer de l'information relative aux données confidentielles. Je présente des mesures de réplication qui permettent aux chercheurs de limiter cette perte d'information. Les mesures de réplication se basent sur le sous-échantillon et la méthode d'agrégation tirés de la documentation sur la confidentialité différentielle, dans laquelle nous (1) divisons premièrement les données en sous-ensembles disjoints (2) puis calculons certaines mesures résumant la différence entre les analyses publiées et parallèles pour chaque sous-ensemble (3) et agrégeons les estimations de sous-ensemble, et (4) finalement ajoutons le bruit à la valeur agrégée afin de satisfaire la confidentialité différentielle. J'illustre les méthodes par des études empiriques.

A Memorial Session for Hélène Massam
Séance commémorative pour Hélène Massam

Chair/Président: Christian Genest

Organizer/Responsable: Xin Gao, Christian Genest

Date: Wednesday June 1 / mercredi 1 juin

Time/Heure: 11:00-12:30

Abstract/Résumé

[11:06-11:34]

Laurent Briollais (University of Toronto/Lunenfeld-Tanenbaum Research Institute) **Nanwei Wang** (University of New Brunswick) **Xin Gao** (York University) **Helene Massam** (York University)

The Scalable Birth-death MCMC Algorithm for Mixed Graphical Model Learning with Application to Genomic Data Integration

L'Algorithme MCMC de naissance et de mort pour l'apprentissage de modèles graphiques mixtes appliqué à l'intégration de données génomiques

Recent advances in biological research have seen the emergence of high-throughput technologies with numerous applications. In cancer research, the challenge is now to perform integrative analyses of high-dimensional multi-omic data with the goal to better understand genomic processes that correlate with cancer outcomes. We propose here a novel mixed graphical model approach to analyze multi-omic data of different types (continuous, discrete and count) and perform model selection by extending the Birth-Death MCMC (BDMCMC) algorithm. We compare the performance of our method to the LASSO and the standard BDMCMC methods using simulations and found that our method is superior in terms of both computational efficiency and the accuracy of the model selection results. Finally, an application to the TCGA breast cancer data shows that integrating genomic information at different levels (mutation and expression data) leads to better subtyping of breast cancers.

Les récentes avancées en recherche biologique ont mené à l'émergence de technologies à haut débit ayant de nombreuses applications. Dans la recherche sur le cancer, le défi consiste à réaliser des analyses intégratives de données multiomiques de haute dimension dans le but de mieux comprendre les processus génomiques en corrélation avec les résultats cancéreux. Nous proposons ici une nouvelle approche de modèle graphique mixte pour analyser les données multiomiques de différents types (continus, discrets et de dénombrement) et sélectionner le modèle en élargissant l'algorithme MCMC de naissance et de mort (BDMCMC). Nous comparons la performance de notre méthode par rapport aux méthodes LASSO et BDMCMC standard à l'aide de simulations et avons établis la supériorité de notre méthode en termes de rendement en calcul et de précision relative aux résultats de sélection de modèles. Enfin, son application aux données du TCGA sur le cancer du sein démontre qu'intégrer les renseignements génomiques à différents niveaux (mutation et données d'expression) mène à un sous-typage supérieur des cancers du sein.

[11:34-12:02]

Yanyan Wu (University of Hawaii at Manoa)

Saddle Point Approximation for the Test of Equality of Covariance Matrices from Decomposable Graphical Gaussian Models

Méthode du point col pour le test de l'égalité des matrices de covariance découlant de modèles graphiques gaussiens décomposables

This paper considers the test of the equality of covariance matrices from graphical Gaussian models that are Markov with respect to a decomposable graph G. We first derived the modified likelihood ratio statistics, i.e., the Bartlett–Box M-statistic, which has a first-order ac-

Cet exposé s'intéresse au test de l'égalité des matrices de covariance découlant de modèles graphiques gaussiens décomposables qui sont des processus Markov quant au graphe décomposable G. Nous dérivons d'abord les statistiques du rapport de vraisemblance modifié, c.-à-d. la M-statistique de Bartlett–Box, dont l'exacti-

curacy. Next, we applied the saddle-point based approximation method for the cumulative distribution function of the Bartlett–Box M-statistic. The proposed saddle-point based method has a third-order accuracy. Simulation studies show that the proposed method has extremely good coverage properties even when the sample size is small.

tude est du premier ordre. Nous appliquons ensuite la méthode d'approximation de point col pour la fonction de distribution cumulative de la M-statistique de Bartlett–Box. L'exactitude de la méthode du point col proposée est de troisième ordre. Des études en simulation indiquent que la méthode proposé a d'excellentes propriétés de couverture lorsque l'échantillon est de petite taille.

[12:02-12:30]

Gerard Letac (Université de Toulouse)

Scale Mixtures of Gaussian Laws: the Quasi-Kolmogorov-Smirnov and Logistic Laws.

Mélanges de variance de lois gaussiennes : lois quasi Kolmogorov-Smirnov et quasi logistiques

One of our last papers with Helene Massam was devoted to the Gaussian approximation of the law of $M = Z\sqrt{V}$ where $V > 0$ is independent of $Z \sim N(0,1)$, (Kybernetika, 56 (6), 1063-1080, arXiv 1810.02036). For instance, if V has the Kolmogorov Smirnov law (K-S) then M has the logistic law. More generally we consider here the case where M has a quasi-logistic law, i.e. with density proportional to $(\cosh x+\theta)^{-1}$ with $\theta > -1$. We develop the elegant properties of the corresponding quasi K-S law of V related to the Brownian bridge.

Un de nos derniers articles avec Hélène Massam était consacré à l'approximation par une loi de Gauss de la loi de $M = Z\sqrt{V}$ où $V > 0$ est indépendante de $Z \sim N(0,1)$ (Kybernetika, 56 (6), 1063-1080, arXiv 1810.02036). Par exemple, si V suit la loi de Kolmogorov Smirnov (K-S) alors M est de loi logistique. Plus généralement, on considère ici le cas où M est de loi quasi-logistique, i.e. c'est a dire de loi de densité proportionnelle à $(\cosh x+\theta)^{-1}$ avec $\theta > -1$. On développe les propriétés élégantes de la loi de V correspondante, dite quasi K-S et ses liens avec le pont brownien.

The Business of Sports Analytics
Le commerce de l'analyse sportive

Chair/Président: Jean-Francois Plante

Organizer/Responsable: Shirley E. Mills

Date: Wednesday June 1 / mercredi 1 juin

Time/Heure: 11:00-12:30

Abstract/Résumé

[11:00-11:30]

Michael Jung (Maple Leaf Sports and Entertainment)
Creating Actionable Insights in the Business of Sports
Création de perspectives exploitables dans l'industrie du sport

In this session, we will provide an overview of the evolving role of analytics in the sports and entertainment industry across the different areas of the business and operations. We will highlight specific use cases that have been initiated by the Business Intelligence team @ MLSE that are creating actionable insight and data-based decision making throughout the organization. Through a data driven, fan experience design philosophy, MLSE Digital Labs enables dynamic reporting and visualization, real time data optimization, and development of models and forecasts to better anticipate a constantly evolving market. We will also identify business opportunities that the overall industry is looking to solution through advanced analytics.

Durant cet exposé, nous vous offrirons une vue d'ensemble du rôle grandissant de l'analytique dans l'industrie du sport et du divertissement dans les différentes spères de l'industrie et des opérations. Nous mettrons l'accent sur les cas d'utilisation précis ayant été mis en place par l'équipe de veille stratégique de MLSE pour produire une perspective exploitable et une prise de décision fondée sur des données dans l'ensemble de l'organisation. À partir d'une philosophie axée sur l'expérience des partisans et sur des données, MLSE Digital Labs offre des visualisations et des rapports dynamiques, une optimisation des données en temps réel et la conception de modèles et de prévisions afin de mieux anticiper un marché en constante évolution. Nous soulignerons aussi des occasions d'affaires que l'ensemble de l'industrie cherche à combler par l'analytique avancée.

[11:30-12:00]

Luke C. Bornn (Zelus Analytics) **Jacob Mortensen** (Zelus Analytics)
From Pixels to Points: Using Tracking Data to Measure Performance in Professional Sports
Des pixels aux points : l'utilisation de données de suivi pour mesurer la performance dans les sports professionnels

In this talk I will explore how players perform, both individually and as a team, on a basketball court. By blending advanced spatio-temporal models with geography-inspired mapping tools, we are able to understand player skill far better than either individual tool allows. Using optical tracking data consisting of hundreds of millions of observations, I will demonstrate these ideas by characterizing defensive skill and decision making in NBA players.

Notre exposé porte sur la performance sur le terrain de joueurs de basket-ball, à titre individuel et en équipe. En mélangeant des modèles spatio-temporels à des outils de mappage d'inspiration géographique, nous avons une bien meilleure compréhension des aptitudes d'un joueur qu'avec l'un ou l'autre séparément. À l'aide de données de suivi optiques consistant en centaines de millions d'observations, je démontre ces idées en caractérisant l'aptitude défensive et la prise de décision chez les joueurs de la NBA.

[12:00-12:30]

Shane Malloy (University of New Brunswick)
The Future of Statistics in NHL Hockey Operations
L'avenir de la statistique dans les opérations de hockey de la LNH

The presentation's focus will be a discussion on the value that statistics plays within the interdisciplinary and integration process when solving complex questions within National Hockey League operations departments. A primary purpose is to open the discussion towards the necessity of interdisciplinary research the integration of statistics to solve complex problems in NHL hockey operations departments. In addition, statistics is a valuable tool that links the qualitative and quantitative data together within a mixed methodology to tell a story. I will engage the audience to consider the alternative options available to utilize statistics in ways that require imagination, creativity, interdisciplinarity, mixed methodologies, and neo-generalism to solve complex problems within NHL Hockey operations. In addition, I will share aspects of previous studies and data I have collected to buttress the reasoning behind considering alternative uses of statistics for NHL hockey operations. In conclusion, I discuss the original question, the problems that face NHL hockey operations departments, and how statistics play a valuable role in finding solutions to complex issues. With the assistance of statistics practitioners, the National Hockey League and its membered clubs will be able to push the boundaries of their future efficiencies and success.

L'objectif de la présentation sera une discussion sur la valeur que la statistique joue dans le processus interdisciplinaire et d'intégration lors de la résolution de questions complexes au sein des départements des opérations de la Ligue nationale de hockey. L'objectif principal est d'ouvrir la discussion sur la nécessité d'une recherche interdisciplinaire et de l'intégration de la statistique pour résoudre des problèmes complexes dans ces départements. En outre, la statistique est un outil précieux qui relie données qualitatives et quantitatives au sein d'une méthodologie mixte pour raconter une histoire. J'inviterai l'auditoire à considérer les autres options disponibles pour utiliser la statistique d'une manière qui requiert de l'imagination, de la créativité, de l'interdisciplinarité, des méthodologies mixtes et du néo-généralisme pour résoudre des problèmes complexes au sein des opérations de hockey de la LNH. En outre, je partagerai les aspects d'études précédentes et les données que j'ai recueillies pour étayer le raisonnement derrière l'examen des utilisations alternatives de la statistique pour les opérations de hockey de la LNH. En conclusion, je discute de la question initiale, des problèmes auxquels sont confrontés ces départements et de la façon dont la statistique joue un rôle précieux dans la recherche de solutions à des problèmes complexes. Avec l'aide des praticiens de la statistique, la Ligue nationale de hockey et ses clubs membres seront en mesure de repousser les limites de leur efficacité et de leur succès futur.

Chair/Président: Golara Zafari

Date: Wednesday June 1 / mercredi 1 juin

Time/Heure: 11:00-12:30

Abstract/Résumé

[11:00-11:15]

Shi Zhang (University of New Brunswick) **Renjun Ma** (University of New Brunswick) **Guohua Yan** (University of New Brunswick)

Cox Survival Models with Partially Crossed Random Effects: an Application to Car Accident Data Cross-Classified by Location and Agent

Modèles de survie de Cox avec effets aléatoires partiellement croisés : application aux données d'accidents de voiture classées de manière croisée par lieu et par agent

In automobile insurance studies, car accident data are often partially cross-classified by location and agent. One research question of great interest is to link time to the occurrence of car accidents with various factors. An appropriate analysis of such data needs to account for location and agent effects. In this talk, we incorporate partially crossed random effects into Cox proportional hazards models for such data and propose a Poisson modeling approach to model estimation. We predict the random effects using the orthodox best linear unbiased predictor method, and obtain consistent estimators for the regression parameters. This estimating method relies on only the first and second moments of the random effects. Our approach is illustrated with a collection of large automobile insurance data. Another potential application of our approach is to study clinical data partially cross-classified by residential areas and medical service providers.

Dans les études sur l'assurance automobile, les données sur les accidents de voiture sont souvent partiellement classées de manière croisée par lieu et par agent. Une question de recherche de grand intérêt est de relier le temps avant occurrence des accidents de voiture et divers facteurs. Une analyse appropriée de ces données doit tenir compte des effets de lieu et d'agent. Dans cet exposé, nous incorporons des effets aléatoires partiellement croisés dans les modèles de risques proportionnels de Cox pour de telles données et proposons une approche de modélisation de Poisson pour estimer le modèle. Nous prédisons les effets aléatoires à l'aide de la méthode traditionnelle du meilleur prédicteur linéaire sans biais, et obtenons des estimateurs cohérents pour les paramètres de régression. Cette méthode d'estimation s'appuie uniquement sur les premiers et seconds moments des effets aléatoires. Nous illustrons notre approche à l'aide d'un ensemble de données d'assurance automobile. Une autre application potentielle de notre approche est l'étude de données cliniques partiellement classées de manière croisée par zones résidentielles et prestataires de services médicaux.

[11:15-11:30]

Ye Wang (University of Calgary) **Wenjun Jiang** (University of Calgary)

Optimal Reinsurance Under Vajda Condition and Range-Value-at-Risk

Réassurance optimale sous condition de Vajda et plage de valeur à risque

In this project we study an optimal reinsurance problem where the insurer's risk-adjusted liability gets minimized. To better reflect the spirit of reinsurance, we impose exogenously Vajda condition on indemnity functions which requires the reinsurer to pay an increasing proportion of loss. To consider both robustness and tail risk, the insurer is assumed to apply Range-Value-at-Risk (RVaR) to evaluate its risk. Under the expected

Cet exposé aborde un problème de réassurance optimale qui minimise la responsabilité de l'assureur ajustée au risque. Pour mieux refléter l'esprit de la réassurance, nous imposons de façon exogène une condition de Vajda sur les fonctions d'indemnisation qui exige que l'assureur paie une proportion accrue de la perte. Afin de prendre en compte la robustesse et le risque de queue, on suppose que l'assureur applique une plage de valeur à risque (RVaR) pour évaluer son risque. En vertu du principe de la prime de va-

value premium principle, we derive the closed-form solution to our problem, which includes the results in Chi and Weng (2013) as special cases. Some comparative studies and sensitivity analysis are also carried out through numerical examples.

leur attendue, nous dérivons une solution de forme fermée à notre problème, ce qui comprend les résultats de Chi et Weng (2013) à titre de cas particuliers. Des études comparatives et une analyse de sensibilité sont aussi menées à l'aide d'exemples numériques.

[11:30-11:45]

Louis Arsenault-Mahjoubi (Simon Fraser University) **Jean-François Bégin** (Simon Fraser University)

On the Bayesian Estimation of Jump-Diffusion Models in Finance

Sur l'estimation bayésienne des modèles de diffusion avec sauts en finance

The jump-diffusion framework encompasses most affine and nonaffine one-factor models used in finance. Due to the model complexity of this framework, particle filters and combinations of Gibbs and Metropolis-Hastings samplers have been the tools of choice for its estimation. However, recent research has shown that the discrete nonlinear filter (DNF) can also be used for fast and accurate maximum likelihood estimation of jump-diffusion models. We present a combination of the DNF with Markov chain Monte Carlo (MCMC) methods for Bayesian estimation in the spirit of the particle MCMC algorithm. In addition, we show that option prices can be easily included into the DNF's likelihood evaluations even in the nonaffine case, which allows for efficient joint Bayesian estimation. We finally present joint estimation results using affine and nonaffine models and S&P 500 data.

Les modèles de diffusion avec sauts englobent la majorité des modèles à un facteur affine et non-affine utilisés en finance. À cause de la complexité de ces modèles, les filtres particulaires et les combinaisons d'échantillonnage de Gibbs et de Metropolis-Hastings sont les principales techniques d'estimation choisies. Des études récentes démontrent cependant que le filtre discret non linéaire (DNF) peut également évaluer la vraisemblance de ces modèles avec rapidité et exactitude. Nous présentons une combinaison du DNF avec les chaînes de Markov Monte-Carlo (MCMC) similaires aux méthodes MCMC utilisant le filtre particulaire. Nous démontrons aussi que l'approche est particulièrement efficace pour inclure des produits dérivés dans la vraisemblance (par exemple, des options européennes), et ce même dans le cas non-affine. Nous présentons des résultats d'estimation avec des modèles affines et non-affines pour l'indice S&P 500.

[11:45-12:00]

Dechen Gao (Western University) **Jiandong Ren** (Western University)

Fuzzy credibility

Crédibilité floue

This paper studies the actuarial credibility theory when the information about the loss model or the prior distribution of its parameters is imprecise or vague. This problem has been studied by many authors. For example, Gómez-Déniz (2009) assumes that the parameters of the prior distribution belong to some interval and proposes to calculate credibility premium based on the posterior regret Γ-minimax principle. Hong & Martin (2021) derive interval estimators for Bühlmann credibility premium when only partial information about the loss distribution and prior distribution are available. In this paper, we propose to represent the imprecise/partial/vague information about model parameters as fuzzy numbers. Based on some basic results in fuzzy set theory, we derive formulas for "fuzzy credibility premiums". Our results extend those derived in

Dans cette étude, nous analysons la théorie de la crédibilité actuarielle lorsque l'information sur le modèle de pertes ou la distribution a priori des paramètres de celle-ci est imprécise ou vague. Ce problème a été étudié par de nombreux auteurs. Par exemple, dans l'étude de Gómez-Déniz (2009), on part du principe que les paramètres de la distribution a priori appartiennent à un certain intervalle et on propose de calculer la prime de crédibilité sur la base du principe de Γ-minimax du regret a posteriori. Hong et Martin (2021) obtiennent des estimateurs par intervalle pour la prime de crédibilité de Bühlmann seulement lorsque des informations partielles sur la distribution des pertes et la distribution a priori sont fournies. Dans cette présentation, nous proposons de représenter les informations imprécises, partielles et vagues sur les paramètres du modèle sous forme de nombres flous. À partir de certains résultats de base de la théorie des ensembles flous, nous obtenons des formules pour les ≪ primes de crédibilité floues ≫.

the above two papers and provide an alternative approach to set credibility premium when the information for model/prior distribution is vague.

Nos résultats étendent ceux qui ont été obtenus dans les deux articles précédents et donnent une autre approche pour fixer la prime de crédibilité lorsque l'information sur le modèle ou la distribution a priori est vague.

[12:00-12:15]

Tingting Chen **Peter Adamic** (Laurentian University) **Anthony F. Desmond** (University of Guelph)

Generalized Additive Modelling for the Accurate Estimation of Insurance Claims

Modélisation additive généralisée pour l'estimation précise des réclamations d'assurance

This paper examines the problem of accurately estimating the expected value and variance of the aggregate claims for each policyholder. To this end, the framework of generalized linear models (GLMs) for aggregate claims is extended to a structure of frequentist generalized additive models (GAMs) based on cubic penalized regression splines. The new structure could allow more flexible nonlinear and/or nonparametric trend terms for the marginal claim frequency and conditional claim severity models. This nonparametric approach is illustrated through simulation. The hypothesis tests' results, AIC values and graphical diagnostics all show that the GAMs give a better fit than the GLM approach.

Cet article examine le problème de l'estimation précise de la valeur attendue et de la variance des réclamations agrégées pour chaque titulaire de police. À cette fin, nous étendons le cadre des modèles linéaires généralisés (GLM) pour les réclamations agrégées à une structure de modèles additifs généralisés (GAM) fréquentistes basés sur des splines de régression à pénalité cubiques. La nouvelle structure pourrait permettre l'utilisation de termes de tendance non linéaires et/ou non paramétriques plus flexibles pour les modèles de fréquence marginale et de gravité conditionnelle des réclamations. Nous illustrons cette approche non paramétrique par une simulation. Les résultats des tests d'hypothèse, les valeurs AIC et les diagnostics graphiques montrent tous que les GAM offrent un meilleur ajustement que l'approche GLM.

[12:15-12:30]

Si Chen (Wilfrid Laurier University) **Zilin Wang** (Wilfrid Laurier University) **David Soave** (Wilfrid Laurier University) **Mary Kelly** (Wilfrid Laurier University)

Fitting Left Truncated Data using Aggregate Loss Model with Poisson-Tweedie Loss Frequency

Ajustement de données tronquées à gauche en utilisant le modèle de perte agrégée avec fréquence de perte Poisson-Tweedie

We extended the candidate pool for modelling the aggregate loss frequency to the three-parameter Poisson-Tweedie (PT) distribution family. With a reporting threshold, small losses will not be observed, thus causing a left-truncation phenomenon where the observed loss frequency is less than the real loss frequency. This raises a new challenge in parameter estimation. We prove that Poisson-Tweedie is closed under binomial thinning. This fact enables us to leverage the existing algorithm for untruncated data to estimate the parameters of the aggregate loss model with truncated data, thus, facilitating the application. With the estimated parameters, the value at risk of the aggregate loss model can be approximated by a Monte-Carlo method. We investigate its application through a simulation study and demonstrate our fitting approach using manual truncation of claims data from the Transportation Security Administration (TSA).

Nous avons élargi le bassin de candidats pour modéliser la fréquence de perte agrégée en incluant la famille de distribution Poisson-Tweedie (PT) à trois paramètres. Avec un seuil de déclaration, de faibles pertes ne seront pas observées, ce qui entraînera un phénomène de troncature à gauche dans lequel la fréquence de perte observée est inférieure à la fréquence de perte réelle. Par conséquent, un nouveau problème se pose pour l'estimation des paramètres. Nous prouvons la clôture de la fréquence de perte Poisson-Tweedie selon un amincissement binomial. Comme cela nous permet de tirer parti de l'algorithme existant pour les données non tronquées pour estimer les paramètres du modèle de perte agrégée avec données tronquées, l'application s'en trouve simplifiée. Avec des paramètres estimés, la méthode de Monte-Carlo peut servir à l'approximation de la valeur à risque du modèle de perte agrégée. Nous en étudions l'application en procédant à une étude en simulation et illustrons notre approche d'ajustement à l'aide d'une troncature manuelle de données de réclamations provenant de la Transportation Security Administra-

tion (TSA).

Chair/Président: Yifan Li

Date: Wednesday June 1 / mercredi 1 juin

Time/Heure: 11:00-12:30

Abstract/Résumé

[11:00-11:15]

Jack Davis (University of Waterloo)

Gambling and Games of Chance – A Course Proposal

Jeux d'argent et jeux de hasard – Une proposition de cours

This is a proposal for an applied, simulation-based, elective survey course on statistics that uses games of chance as its case bases. Students will develop an understanding about gambling to protect themselves from Canada's rapidly growing private gaming market and a foundation in probability with applications relevant for work in finance, insurance, or the gaming industry. Because gaming exists in many cultures, a gaming course provides an excellent opportunity to show representation of marginalized cultures in statistics by including respectful analyses of their games. Core topics include Monte Carlo simulation through bingo, coupon collecting, and Borel; combinatorics through video poker and Mahjong; game theory through Texas Hold'em poker; conditional probability through sporting event markets; time series through the roulette Martingale strategy. A substantial introductory chapter on gambling addiction is included.

Voici une proposition de cours facultatif de statistiques d'enquête appliquées fondé sur des simulations se servant de jeux de hasard en guise de scénarios de base. Les étudiants acquerront des connaissances sur les jeux d'argent afin de se prémunir contre l'industrie du jeu en grande expansion au Canada. Ils apprendront aussi les fondements en probabilité appliqués à des domaines pertinents comme la finance, l'assurance ou l'industrie du jeu. Les jeux existent dans un grand nombre de cultures, c'est pourquoi un cours sur les jeux offre une excellente occasion de présenter des cultures marginalisées en statistiques en intégrant des analyses respectueuses de leurs jeux. Les sujets principaux comprennent : des simulations par la méthode de Monte Carlo appliquées au bingo, à la collecte de coupons et au jeu ≪Borel≫ ; des combinatoires à partir du vidéopoker et du Mahjong ; la théorie du jeu à partir du poker Texas Hold'em ; la probabilité conditionnelle à partir des marchés d'événements sportifs ; des séries temporelles à partir de la stratégie de Martingale appliquée à la roulette. Un important chapitre d'introduction sur le jeu pathologique est inclus.

[11:15-11:30]

Suborna Shekhor Ahmed (University of British Columbia) **Michelle Zeng** (University of British Columbia) **Patrick Culbert** (University of British Columbia) **Yangqian Qi** (University of British Columbia)

Survey data analysis of engagement and self-efficacy in a concurrent hybrid modality

Analyses de données d'enquête sur l'engagement et l'autoefficacité dans un modèle concurrent hybride

A concurrent hybrid model was adopted to create an adequate environment for students in a computation course to learn collaboratively and engage with course materials in real-time when the teaching team members were present to support on-campus and remote students together. We gain insights into the pedagogical value of the concurrent hybrid model in education and better understand the factors affecting students' experiences through collecting surveys. Responses were analyzed using descriptive and inferential statistics to measure changes in confidence, engagement, self-efficacy using

Un modèle concurrent hybride a été adopté afin de créer un environnement adéquat pour que les étudiants dans un cours de calcul puissent apprendre de façon collaborative et aborder le matériel de cours en temps réel lorsque les membres de l'équipe d'enseignants peuvent soutenir les étudiants sur place et à distance. Par l'entremise d'enquêtes recueillies, nous apprenons rétroactivement la valeur pédagogique du modèle concurrent hybride et discernons mieux les facteurs qui influencent l'expérience des étudiants. Les réponses ont été analysées au moyen de statistiques inférentielles et descriptives pour mesurer les variations de confiance, d'engagement et d'autoefficacité à partir de tests d'échantillons par

paired samples tests, and trends were observed for the mastery of content knowledge. We found a significant increase in students' average confidence scores from midterm to the end of the term. However, there was no significant change on average for the level of engagement within individuals throughout the term. The study findings shed some light on the effectiveness of this flexible learning approach.

paires, et des schémas ont été observés relativement à la maîtrise du contenu d'apprentissage. Nous avons souligné une hausse considérable du niveau de confiance moyen des élèves entre la mi-session et la fin de session. Cependant, il n'y a pas eu de changement significatif en moyenne en ce qui concerne le niveau d'engagement parmi les individus durant la session. Les résultats de l'étude démontrent l'efficacité de cette méthode d'apprentissage polyvalente.

[11:30-11:45]

Samuel Perreault (University of Toronto)
Efficient Computation for Inference with Kendall's Tau
Calcul efficace pour l'inférence avec sur le tau de Kendall

The standard algorithm for computing Kendall's tau empirical correlation is modified so that it also returns a jackknife estimate of its variance. This is done efficiently in the sense that the log-linear time complexity of the original algorithm is preserved.

L'algorithme standard pour calculer le tau de Kendall empirique est modifié de telle sorte qu'il produit aussi un estimé de type jackknife de sa variance. Ceci est fait efficacement dans le sens où la complexité temporelle log-linéaire de l'algorithme d'origine est préservée.

[11:45-12:00]

Federico Severino (Université Laval) **Marzia Angela Cremona** (Université Laval) **Éric Dadié** (Université Laval)
COVID-19 effects on the Canadian Term Structure of Interest Rates
Effets de la COVID-19 sur la structure à terme des taux d'intérêt au Canada

In Canada, COVID-19 pandemic triggered exceptional monetary policy interventions by the central bank, which in March 2020 made multiple unscheduled cuts to its target rate. The aim of this paper is to assess the extent to which Bank of Canada interventions affected the determinants of the yield curve. By applying Functional Principal Component Analysis to the term structure of interest rates we find that, during the pandemic, the long-run dependence of level and slope components of the yield curve is unchanged with respect to previous months, although the shape of the mean yield curve completely changed after target rate cuts. Bank of Canada was effective in lowering the whole yield curve and correcting the inverted hump of previous months, but it was not able to reduce the exposure to already existing long-run risks.

Au Canada, la pandémie de COVID-19 a déclenché des interventions exceptionnelles de politique monétaire de la part de la banque centrale qui, en mars 2020, a procédé à de multiples réductions non programmées de son taux cible. L'objectif de cet article est d'évaluer dans quelle mesure les interventions de la Banque du Canada ont affecté les déterminants de la courbe des rendements. En appliquant l'Analyse en composantes principales fonctionnelle à la structure à terme des taux d'intérêt, nous constatons que, pendant la pandémie, la dépendance à long terme des composantes de niveau et de pente de la courbe de rendement est inchangée par rapport aux mois précédents, bien que la forme de la courbe de rendement moyenne ait complètement changé après les réductions du taux cible. La Banque du Canada a réussi à abaisser l'ensemble de la courbe de rendement et à corriger la bosse inversée des mois précédents, mais elle n'a pas été en mesure de réduire l'exposition aux risques à long terme déjà existants.

[12:00-12:15]

William Ruth (Simon Fraser University) **Richard Lockhart** (Simon Fraser University)
Simulated Epidemic Spread in University Classes
Simulation d'une propagation épidémique pendant les cours dans une université

We investigate transmission dynamics for SARS-CoV-2 on a real network of classes at Simon Fraser University, a medium-sized school in Western Canada. Outbreaks are simulated over the course of one semester

Nous étudions la dynamique de transmission du virus SARS-CoV-2 dans un réseau réel de cours à la Simon Fraser University, un établissement de taille moyenne dans l'Ouest canadien. Des éclosions sont simulées tout au long d'un semestre dans de nom-

across numerous parameter settings for a realistic compartment model, including asymptomatic and presymptomatic transmission. We investigate the control strategy of moving large classes online while small classes are allowed to meet in person. Regression trees are used to model the effect of disease parameters on simulation outputs; specifically, the total number of infections and the peak number of simultaneous cases.

breux contextes paramétriques pour obtenir un modèle à compartiments réaliste, y compris une transmission asymptomatique et présymptomatique. Notre enquête porte sur la stratégie de contrôle consistant à donner des cours à de vastes groupes en ligne, tandis que les cours en petits groupes peuvent se donner en présentiel. Des arbres de régression sont utilisés pour modéliser l'effet des paramètres de la maladie sur les résultats de la simulation ; plus précisément, le nombre total d'infections et le pic de cas simultanés.

[12:15-12:30]

Surani Matharaarachchi (University of Manitoba) **Mike Domaratzki** (University of Western Ontario) **Alan Katz** (University of Manitoba) **Saman Muthukumarana** (University of Manitoba)

Discovering Symptom Patterns of Long COVID Patients in Tweets using Association Rule Mining

Découverte de modèles de symptômes chez des patients atteints de COVID longue dans des tweets à l'aide de l'extraction de règles d'association

The COVID-19 pandemic is a significant public health crisis that negatively affects human health and well-being. In addition to being infected with the Coronavirus, patients can experience long-term health effects, called long COVID. This syndrome is characterized by multiple symptoms, and it is crucial to identify these symptoms as they might negatively impact patients' day-to-day lives. Breathlessness, fatigue, and brain fog are the three main continuing and debilitating symptoms that have been reported by long COVID patients, often months after the onset of the COVID-19 disease. Under such circumstances, understanding the patterns of long COVID symptoms and their behavior are vital to our understanding of long COVID. This study aimed to describe symptom patterns among long COVID patients using Twitter social media discussion forum data and using association rule mining techniques.

La pandémie de COVID-19 est une crise de santé publique importante qui frappe durement la santé et le bien-être des personnes. En plus de l'infection au coronavirus, les patients peuvent souffrir de séquelles de ce qu'il est convenu d'appeler la COVID longue. Ce syndrome se caractérise par de multiples symptômes qu'il est essentiel d'identifier en raison de leurs effets négatifs possibles sur la vie des patients au jour le jour. Essoufflement, fatigue et brouillard mental sont les trois principaux symptômes continus et débilitants rapportés par les patients atteints de la COVID longue, souvent des mois après l'apparition de la maladie. Dans de tels cas, la compréhension des modèles de symptômes de la COVID longue et de leur comportement est indispensable pour notre compréhension de cette forme longue de la maladie. Cette étude vise à décrire les modèles de symptômes chez des patients atteints de COVID longue en utilisant des données de forum de discussions du réseau social Twitter et des techniques d'extraction de règles d'association.

Memorial Session for Donald A. S. Fraser
Session commémorative pour Donald A. S. Fraser

Chair/Président: Mary E. Thompson

Organizer/Responsable: Mary E. Thompson

Date: Wednesday June 1 / mercredi 1 juin

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:45-14:25]

Nancy Reid (University of Toronto)
From Structural Inference to Asymptotic Theory
De l'inférence structurelle à la théorie asymptotique

Don Fraser's efforts to understand and extend Fisher's fiducial inference led him to what he called a structural approach to inference, which established a relationship between data and models that was, to him, quite tangible. This approach turned out also to be very helpful for advancing the theory of higher-order asymptotics for likelihood inference. I will try to describe this evolution, as I had the privilege of seeing it develop in real time.

Les efforts de Don Fraser pour comprendre et étendre l'inférence fiduciaire de Fisher l'ont mené à ce qu'il appelle une approche structurelle de l'inférence, qui établit un lien entre des données et des modèles qui étaient à ses yeux tout à fait tangibles. D'ailleurs, cette approche s'est avérée être très utile dans le développement de la théorie asymptotique d'ordres supérieurs pour les inférences de vraisemblance. Je tenterai de décrire cette évolution, considérant que j'ai eu le privilège de voir son développement en temps réel.

[14:25-15:00]

Mylène Bédard (Université de Montréal)
Recent Advances in Statistical Inference
Avancées récentes en inférence statistique

This talk presents some of the latest contributions of Professor D.A.S. Fraser; we go over the general ideas in our recent collaborations. Don's enthusiasm, energy, and ever-present originality were widely admired by all who worked with him, and have established him as a pillar of the statistical community. The statistical toolbox offers many methods for applied statistics, but reliability (reproducibility of frequency properties) is often unclear or even ignored. We study default Bayes methods and develop a prior that leads to full second-order inference for any regular scalar parameter of interest in presence of nuisance parameters; the new prior is Jeffreys based. In parallel, we use location model methodology to guide the least squares analysis in the traditional Lasso problem of variable selection and inference. The resulting Linear Lasso is one-dimensional rather than n-dimensional, removes ineffective variables by distributional shift, and is relatively easy to implement.

Nous présentons quelques contributions récentes du professeur D.A.S. Fraser ; L'enthousiasme, l'énergie et l'originalité omniprésente de Don ont été largement admirés par tous ceux qui ont travaillé avec lui et l'ont établi comme un pilier de la communauté statistique. Il existe de nombreuses méthodes statistiques, mais la fiabilité (reproductibilité des propriétés de fréquence) est souvent floue ou même ignorée. Nous étudions les méthodes de Bayes objectives et développons une loi à priori conduisant à une inférence de second ordre pour tout paramètre d'intérêt scalaire et régulier en présence de paramètres nuisibles ; celle-ci est basée sur la loi de Jeffreys. En parallèle, nous utilisons les modèles de position pour guider l'analyse des moindres carrés dans le problème traditionnel de Lasso de sélection de variables et d'inférence. Le Lasso linéaire résultant est unidimensionnel, supprime les variables inefficaces par glissement distributionnel et est facile à mettre en œuvre.

Chair/Président: Johanna G. Neslehova

Organizer/Responsable: Léo Belzile

Date: Wednesday June 1 / mercredi 1 juin

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-14:00]

Stanislav Volgushev (University of Toronto) **Sebastian Engelke** (University of Geneva) **Michaël Lalancette** (University of Toronto)

Structure Learning for Extremal Graphical Models

Apprentissage de structure pour des modèles graphiques extrêmes

Extremal graphical models are sparse statistical models for multivariate extreme events. The underlying graph encodes conditional independencies and enables a visual interpretation of the complex extremal dependence structure. For the important case of tree models, we provide a data-driven methodology for learning the graphical structure. We show that sample versions of the extremal correlation and a new summary statistic, which we call the extremal variogram, can be used as weights for a minimum spanning tree to consistently recover the true underlying tree. Remarkably, this implies that extremal tree models can be learned in a completely non-parametric fashion by using simple summary statistics and without the need to assume discrete distributions, existence of densities, or parametric models for marginal or bivariate distributions. Extensions to more general graphs are also discussed.

Les modèles graphiques extrêmes sont des modèles statistiques épars pour les événements extrêmes multivariés. Le graphique sous-jacent encode les indépendances conditionnelles et permet une interprétation visuelle de la structure complexe de dépendance extrême. Pour le cas important des modèles d'arbre, nous fournissons une méthodologie basée sur les données pour apprendre la structure graphique. Nous démontrons que les versions d'échantillon de la corrélation extrême et une nouvelle statistique sommaire, que l'on appelle ≪variogramme extrême≫, peuvent servir de poids pour produire un arbre de recouvrement minimal permettant de récupérer systématiquement l'arbre réel sous-jacent. Remarquablement, cela implique que les modèles d'arbre extrême peuvent être appris de façon non paramétrique à l'aide d'une statistique sommaire simple et sans devoir présumer que les distributions sont discrètes, de l'existence des densités ou de modèles paramétriques des distributions marginales ou bivariées. Des extensions à des graphes plus généraux seront aussi abordées.

[14:00-14:30]

Natalia Nolde (The University of British Columbia)

Linking Representations for Multivariate Extremes via a Limit Set

Liaison de représentations d'extrêmes multivariés par l'entremise d'un ensemble limite

The study of multivariate extremes is dominated by multivariate regular variation, although it is well known that this approach does not provide adequate distinction between random vectors whose components are not always simultaneously large. Various alternative dependence measures and representations have been proposed, with the most well-known being hidden regular variation and the conditional extreme value model. These varying depictions of extremal dependence arise through consideration of different parts of the multivari-

L'étude d'extrêmes multivariés est dominée par la variation régulière multivariée, même s'il est bien connu que cette approche ne fait pas une distinction adéquate entre les vecteurs aléatoires dont les composantes ne sont pas toutes simultanément grandes. Plusieurs options de mesures de dépendance et de représentations ont été proposées, les mieux connues étant la variation régulière cachée et le modèle de valeur extrême conditionnel. Ces représentations de dépendance extrême surviennent lorsqu'on considère différentes parties d'un domaine multivarié, et lorsqu'on découvre précisément ce qui se passe lorsque des

ate domain, and particularly exploring what happens when extremes of one variable may grow at different rates to other variables. Thus far, these alternative representations have come from distinct sources and links between them are limited. In this work we elucidate many of the relevant connections through a geometrical approach based on the shape of scaled sample clouds.

extrêmes d'une variable peuvent croître à différents taux par rapport à d'autres. Jusqu'à présent, ces différentes représentations proviennent de sources distinctes ayant peu de liens entre elles. Dans ce travail, nous élucidons un grand nombre de connexions pertinentes au moyen d'une approche géométrique basée sur la forme de nuages d'échantillon échelonnés.

Statistical Learning and Inference on Streaming and Online Data
Apprentissage statistique et inférence sur les données "streaming" et en ligne

Chair/Président: Dehan Kong

Organizer/Responsable: Linglong Kong

Date: Wednesday June 1 / mercredi 1 juin

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-13:52]

Yingqi Zhao

Constructing Stabilized Dynamic Surveillance Rules for Optimal Monitoring Schedules

Construction de règles de surveillance dynamique stabilisées pour programmes de surveillance optimaux

Dynamic surveillance rules (DSRs) are sequential surveillance decision rules informing monitoring schedules in clinical practice, which can adapt over time according to a patient's evolving characteristics. In many clinical applications, it is desirable to identify and implement optimal stabilized DSRs, where the parameters indexing the decision rules are shared across dierent decision points. We propose a new criterion for DSRs that accounts for benet-cost tradeo during the course of disease surveillance.We develop two methods to estimate the stabilized DSRs optimizing the proposed criterion, and establish asymptotic properties for the estimated parameters of biomarkers indexing the DSRs. The rst approach estimates the optimal decision rules for each individual at every stage via regression modeling, and then estimates the stabilized DSRs via a classication procedure with the estimated time-varying decision rules as the response. The second approach proceeds by optimizing a relaxation of the empirical objective, where a surrogate function is utilized to facilitate computation. Extensive simulation studies are conducted to demonstrate the superior performances of the proposed methods. The methods are further applied to the Canary Prostate Active Surveillance Study (PASS).

Les règles de surveillance dynamique (RSD) sont des règles de décision de surveillance séquentielle informant les programmes de surveillance en pratique clinique, qui peuvent s'adapter dans le temps en fonction des caractéristiques évolutives d'un patient. Dans de nombreuses applications cliniques, il est souhaitable d'identifier et de mettre en œuvre des RSD stabilisées optimales, où les paramètres d'indexation des règles de décision sont partagés entre différents points de décision. Nous proposons un nouveau critère pour les RSD qui tient compte du rapport coût/bénéfice au cours de la surveillance de la maladie. Nous développons deux méthodes pour estimer les RSD stabilisés qui optimisent le critère proposé et établissons des propriétés asymptotiques pour les paramètres estimés des biomarqueurs qui indexent les RSD. La première approche estime les règles de décision optimales pour chaque individu à chaque étape via un modèle de régression, puis estime les DSR stabilisés via une procédure de classification avec comme réponse les règles de décision variables dans le temps estimées. La deuxième approche procède par optimisation d'une relaxation de l'objectif empirique, où une fonction de substitution est utilisée pour faciliter le calcul. Nous menons des études de simulation approfondies pour démontrer les performances supérieures des méthodes proposées. Nous appliquons ensuite les méthodes à l'étude Canary Prostate Active Surveillance Study (PASS).

[13:52-14:14]

Peter X. Song (University of Michigan) **Emily Hector** (North Carolina State University) **Lan Luo** (University of Iowa)

Parallel-and-stream accelerator for computationally fast supervised learning with big data

Accélérateur parallèle et de diffusion pour l'apprentissage supervisé rapide sur le plan informatique avec des mégadonnées

Two dominant distributed computing strategies have emerged to overcome the computational bottleneck of supervised learning with big data: parallel data processing in the MapReduce paradigm and serial data process-

Deux stratégies prédominantes d'informatique distribuée ont émergé pour surmonter le goulot d'étranglement informatique de l'apprentissage supervisé avec des mégadonnées : le traitement parallèle des données dans le paradigme MapReduce et le traite-

ing in the online streaming paradigm. Although these two strategies are both common divide-and-combine approaches, they differ in how they aggregate information, leading to different trade-offs between statistical and computational performances. We propose a new hybrid paradigm, termed a Parallel-and-Stream Accelerator (PASA), that uses the strengths of both distributed strategies for computationally fast and statistically efficient supervised learning. PASA's architecture nests online streaming processing into each distributed and parallelized data process in a MapReduce framework. PASA leverages the advantages and mitigates the disadvantages of both the MapReduce and online streaming approaches to deliver a more flexible paradigm satisfying practical computing needs. We study the analytic properties and computational complexity of PASA and detail its implementation for two key statistical learning tasks. We illustrate its performance through simulations and a large-scale data example building a prediction model for online purchases from advertising data.

ment des données sérielles dans le paradigme de diffusion en ligne. Bien que ces deux stratégies soient des approches communes de division et de combinaison, elles diffèrent dans la manière dont elles agrègent les données, ce qui entraîne des options différentes entre les résultats statistiques et informatiques. Nous proposons un nouveau paradigme hybride, appelé « accélérateur parallèle et de diffusion », qui exploite les points forts des deux stratégies distribuées pour un apprentissage supervisé rapide sur le plan informatique et efficace sur le plan statistique. L'architecture de l'accélérateur parallèle et de diffusion intègre le traitement de la diffusion en continu dans chaque processus de données distribué et parallélisé dans un cadre de MapReduce. L'accélérateur parallèle et de diffusion exploite les avantages et atténue les inconvénients des approches de MapReduce et de diffusion en ligne afin d'offrir un paradigme plus souple répondant à des besoins informatiques concrets. Nous examinons les propriétés analytiques et la complexité informatique de l'accélérateur parallèle et de diffusion, puis nous décrivons sa mise en œuvre pour deux tâches essentielles d'apprentissage statistique. Nous illustrons les résultats de cet accélérateur par des simulations et un exemple de données à grande échelle, en créant un modèle de prédiction des achats en ligne à partir de données publicitaires.

[14:14-14:36]

Hengrui Cai (North Carolina State University) **Ye Shen** (North Carolina State University) **Rui Song** (North Carolina State University)

Doubly Robust Interval Estimation for Optimal Policy Evaluation in Online Learning

Estimation doublement robuste d'intervalles pour une évaluation optimale des politiques en matière d'apprentissage en ligne

Evaluating the performance of an ongoing policy plays a vital role in many areas to provide crucial instruction on the early-stop of the online experiment and timely feedback from the environment. Policy evaluation in online learning thus attracts increasing attention by inferring the mean outcome of the optimal policy (i.e., the value) in real-time. Yet, such a problem is particularly challenging due to the dependent data generated online, the unknown optimal policy, and the complex exploration and exploitation trade-off in the adaptive experiment. To overcome these difficulties, we explicitly derive the probability of exploration that quantifies the probability of exploring the non-optimal actions under commonly used bandit algorithms. We use this probability to conduct valid inference on the online conditional mean estimator under each action and develop the doubly robust interval estimation (DREAM) method to infer the value under the estimated optimal policy in online learning.

Une évaluation de la performance d'une politique courante est essentielle à plusieurs égards afin de fournir à la fois une instruction indispensable sur l'arrêt précoce de l'expérience en ligne et les commentaires de l'environnement en temps opportun. L'évaluation d'une politique en matière d'apprentissage en ligne soulève ainsi un intérêt croissant en inférant le résultat moyen d'une politique optimale (c.-à-d. la valeur) en temps réel. Un problème de cet ordre pose par contre un défi particulier en raison des données dépendantes générées en ligne, de la politique optimale inconnue ainsi que de la relation arbitraire complexe entre exploration et exploitation dans l'expérience adaptative. Pour surmonter ces difficultés, nous dérivons explicitement une probabilité d'exploration qui quantifie la probabilité d'explorer des actions non optimales sous des algorithmes de bandits d'utilisation courante. Nous nous servons de cette probabilité pour en arriver à une inférence valide sur l'estimateur de la moyenne conditionnelle en ligne selon chaque action et développons une méthode d'estimation doublement robuste des intervalles (DREAM) pour inférer la valeur sous une politique optimale estimée dans l'apprentissage

en ligne.

[14:36-14:58]

Linglong Kong (University of Alberta)

Damped Anderson Mixing for Deep Reinforcement Learning: Acceleration, Convergence, and Stabilization

Mélange d'Anderson amorti pour l'apprentissage par renforcement profond : accélération, convergence et stabilisation

Anderson mixing has been heuristically applied to reinforcement learning (RL) algorithms for accelerating convergence and improving the sampling efficiency of deep RL. In this paper, we provide deeper insights into a class of acceleration schemes built on Anderson mixing that improve the convergence of deep RL algorithms. Our main results establish a connection between Anderson mixing and quasi-Newton methods and prove that Anderson mixing increases the convergence radius of policy iteration schemes by an extra contraction factor. The key focus of the analysis roots in the fixed-point iteration nature of RL. We further propose a stabilization strategy by introducing a stable regularization term in Anderson mixing and a differentiable, non-expansive MellowMax operator that can allow both faster convergence and more stable behavior. Extensive experiments demonstrate that our proposed method enhances the convergence, stability, and performance of RL algorithms.

Le mélange d'Anderson a été appliqué de manière heuristique aux algorithmes d'apprentissage par renforcement pour accélérer la convergence et améliorer l'efficacité de l'échantillonnage de l'apprentissage par renforcement profond. Dans cette présentation, nous donnons un aperçu plus approfondi d'une classe de schémas d'accélération basés sur le mélange d'Anderson qui améliorent la convergence des algorithmes d'apprentissage par renforcement profond. Nos principaux résultats établissent un lien entre le mélange d'Anderson et les méthodes quasi-Newton. Ils montrent également que le mélange d'Anderson augmente le rayon de convergence des schémas d'itération des politiques par un facteur de contraction supplémentaire. Notre analyse se concentre sur la nature d'itération à point fixe de l'apprentissage par renforcement. De plus, nous proposons une stratégie de stabilisation en introduisant un terme de régularisation stable dans le mélange d'Anderson et un opérateur mellowmax différentiable et non-expansif qui permet à la fois une convergence plus rapide et un comportement plus stable. Des expériences approfondies démontrent que la méthode que nous proposons améliore la convergence, la stabilité et les résultats des algorithmes d'apprentissage par renforcement.

Active Learning in Statistics: Where Are We Now?
Apprentissage actif en statistique : où en sommes-nous ?

Chair/Président: Chelsea Uggenti

Organizer/Responsable: Douglas G. Woolford, Chelsea Uggenti

Date: Wednesday June 1 / mercredi 1 juin

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-15:00]

Alison L. Gibbs (University of Toronto) **Wesley Burr** (Trent University) **Sohee Kang** (University of Toronto Scarborough)

Active Learning in Statistics: Where Are We Now?

Apprentissage actif en statistique : où en sommes-nous ?

Active learning is an approach to instruction that requires students to thoughtfully engage with course material and often with one another in the classroom. This enhanced student engagement that comes from in-class active learning activities can lead to deeper learning. These ideas and techniques are not new – for the past few decades a plethora of research on active learning shows clear benefits. However, the design and guidance that instructors provide in the classroom is crucial for success. Join our session as we critically question and asses where we, as a community, are in terms of integrating active learning in our teaching of statistics and data science courses. Experts and educators spanning a wide range of career stages will each briefly share their successes, challenges, and overall lessons learned. This will be followed by a panel discussion of their thoughts on active learning within our discipline. Panelists will engage with the audience in a period of general discussion to address any questions they have.

L'apprentissage actif est une approche de l'enseignement qui exige des étudiants qu'ils s'engagent de manière réfléchie dans le matériel de cours et souvent les uns avec les autres dans la salle de classe. L'engagement accru des étudiants qui découle des activités d'apprentissage actif en classe peut mener à un apprentissage plus profond. Ces idées et techniques ne sont pas nouvelles – depuis quelques décennies, une pléthore de recherches sur l'apprentissage actif en montre les avantages évidents. Cependant, la conception et l'orientation que les instructeurs fournissent en classe sont cruciales pour en assurer le succès. Participez à notre session pour questionner et évaluer de manière critique où nous en sommes, en tant que communauté, dans l'intégration de l'apprentissage actif à notre enseignement de la statistique et de la science des données. Des experts et des éducateurs de toutes étapes de carrière partageront chacun brièvement leurs succès, leurs défis et les leçons générales apprises. Cette présentation sera suivie d'une table ronde sur l'apprentissage actif au sein de notre discipline. Les panélistes discuteront ensuite avec le public lors une période de discussion générale et de questions-réponses.

Collaborations and Consultations in an Academic World
Collaboration et consultations dans un monde académique

Chair/Président: Peijun Sang

Organizer/Responsable: Orla A Murphy

Date: Wednesday June 1 / mercredi 1 juin

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-15:00]

Mireille Schnitzer (Université de Montréal) **Thomas Loughin** (Simon Fraser University) **Dave Campbell** (Carleton University) **Gabriela Cohen Freue** (University of British Columbia)

Collaborations and Consultations in an Academic World

Collaborations et consultations dans le monde académique

The New Investigators Committee exists to provide junior academics with opportunities to receive advice on building successful careers. An important element of building a strong research program in statistics is the incorporation of applied work and real-world problems. However, for academics in the early stages of their careers, it can be difficult to start this type of work in a productive way that fosters the growth of their research programs. This session will provide junior academics the opportunity to learn successful techniques, and perhaps some tales of caution, from academics invited by the committee for their experience collaborating and consulting with other fields of academia and industry. This session is 90 minutes in duration and will consist of four panelists, each addressing the audience for 10-15 minutes with prepared points prior to a 30+ minute question-and-answer period facilitated by the session chair. The speakers' prepared points will include a brief self-introduction, as well as their biggest lessons, successes, and tips with regards to applied work. Examples of potential topics include: · How to select appropriate projects and incorporate them into your research program · Tips on conducting collaborations and consultations · How to incorporate graduate student training

Le comité des nouveaux chercheurs existe pour offrir aux jeunes universitaires des conseils pour une carrière réussie. Un élément important de la construction d'un programme de recherche solide en statistique est l'inclusion de travaux appliqués et de problèmes du monde réel. Cependant, pour les universitaires en début de carrière, il peut être difficile d'entamer productivement ce type de travail en favorisant la croissance de leur programme de recherche. Cette session proposera aux jeunes universitaires des techniques éprouvées et quelques conseils de prudence de la part d'universitaires invités par le comité pour leur expérience de collaboration et de consultation avec d'autres domaines du monde universitaire et de l'industrie. Cette session dure 90 minutes et se compose de quatre panélistes, chacun s'adressant au public pendant 10 à 15 minutes avec des points préparés, avant une période de questions et réponses de plus de 30 minutes animée par le président de la session. Les points préparés par les orateurs comprendront une brève présentation, ainsi que leurs plus grandes leçons, succès et conseils en matière de travail appliqué. Voici quelques exemples de sujets potentiels : - Comment sélectionner des projets appropriés et les intégrer à votre programme de recherche - Conseils pour mener des collaborations et des consultations - Comment intégrer la formation des étudiants diplômés.

Recent Advancement and Application of Bayesian Causal Inference Methods
Progrès récents et application des méthodes d'inférence causale bayésienne

Chair/Président: Kuan Liu

Organizer/Responsable: Kuan Liu

Date: Wednesday June 1 / mercredi 1 juin

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-14:00]

Olli Saarela (University of Toronto) **Thai-Son Tang** (University of Toronto) **Zhihui (Amy) Liu** (University Health Network)

Bayesian Non-Parametric Monotonic Regression for Radiotherapy Induced Normal Tissue Complications

Régression monotone non paramétrique bayésienne pour les complications aux tissus sains après radiothérapie

Normal tissue complication probability (NTCP) models are used to assess the dose-toxicity relationship in radiotherapy. Radiation exposure by organ volume is a functional covariate, and in principle its effect on dichotomous or ordinal toxicity outcomes can be modeled through functional generalized linear models, incorporating a monotonicity restriction which is biologically plausible for dose-toxicity relationships. In this talk we discuss the causal interpretation of monotonic functional regression and identifiability issues involved in such models. As an alternative to functional regression, we consider relating the toxicity outcomes marginally to bivariable dose-volume combinations. For this, we adapt a Bayesian non-parametric monotonic multivariable regression model which can also accommodate ordinal outcomes. The model can approximate arbitrary monotonic regression function shapes without common parametric modeling assumptions such as additivity, linearity or proportional odds.

Les modèles de probabilité de complications aux tissus sains servent à évaluer le rapport dose-toxicité en radiothérapie. L'exposition au rayonnement par volume d'un organe est une covariable fonctionnelle, et son effet sur les résultats de toxicité ordinale ou dichotomique peut en principe être modélisé par l'entremise de modèles linéaires généralisés fonctionnels, comprenant une restriction de monotonie qui est biologiquement plausible pour les rapports dose-toxicité. Lors de cet exposé, nous aborderons l'interprétation causale de la régression fonctionnelle monotone et les problèmes d'identification reliés à ce genre de modèles. En guise de deuxième choix à la régression fonctionnelle, nous examinons la possibilité de lier les résultats de toxicité marginalement aux combinaisons de bivariables dose-volume. Pour ce faire, nous adaptons une régression multivariée monotone non paramétrique bayésienne pouvant aussi adapter des résultats ordinaux. Le modèle peut estimer les formes d'une fonction de régression monotone arbitraire sans hypothèse de modélisation paramétrique commune telles que l'additivité, la linéarité et les probabilités proportionnelles.

[14:00-14:30]

Arman Oganisian (Brown University)

A Hierarchical Bayesian Bootstrap for Heterogenous Treatment Effect Estimation

Bootstrap bayésien hiérarchique pour l'estimation des effets de traitement hétérogènes

A major focus of causal inference is the estimation of heterogeneous average treatment effects (HTE) - average treatment effects within strata of another variable of interest such as levels of a biomarker, education, or age strata. Inference involves estimating a stratum-specific regression and integrating it over the distribution of confounders in that stratum - which itself must be estimated. Standard practice involves estimating these stratum-specific confounder distributions indepen-

L'une des priorités reliées à l'inférence causale est d'estimer les effets de traitement hétérogènes (ETH), c'est-à-dire les effets de traitement moyens à l'intérieur de strates de d'autres variables pertinentes comme les niveaux d'un biomarqueur, le niveau d'éducation ou les strates d'âge. L'inférence consiste entre autres à estimer une régression spécifique à des strates et de l'intégrer dans la distribution de confondants dans cette strate, qui doit aussi être estimée. Une pratique standard consiste à estimer indépendamment ces distributions de confondants spécifiques

dently (e.g. via the empirical distribution or Rubin's Bayesian bootstrap), which becomes problematic for sparsely populated strata with few observed confounder vectors. In this paper, we develop a nonparametric hierarchical Bayesian bootstrap (HBB) prior over the stratum-specific confounder distributions for HTE estimation. The HBB partially pools the stratum-specific distributions, thereby allowing principled borrowing of confounder information across strata when sparsity is a concern. We show that posterior inference under the HBB can yield efficiency gains over standard marginalization approaches while avoiding strong parametric assumptions about the confounder distribution. We use our approach to estimate the adverse event risk of proton versus photon chemoradiotherapy across various cancer types.

à des strates (p. ex. par distribution empirique ou par bootstrap bayésien de Rubin), qui deviennent problématiques pour les strates peuplées de façon éparse ayant peu de vecteurs de confondant observés. Dans cet article, nous concevons une priori bootstrap nonparamétrique bayésienne hiérarchique (BBH) à partir des distributions de confondants spécifiques à des strates pour estimer l'ETH. Le BBH regroupe partiellement les distributions spécifiques aux strates, ce qui permet en principe l'emprunt de renseignements relatifs aux confondants à travers les strates lorsque le caractère épars devient un problème. Nous démontrons que l'inférence a posteriori selon le BBH peut atteindre une efficacité supérieure à celle des approches de marginalisation standard, tout en évitant de fortes hypothèses paramétriques relatives à la distribution des confondants. Nous utilisons notre approche pour estimer le risque d'événements indésirables dans la chimioradiothérapie de proton contre photon parmi de nombreux types de cancers.

[14:30-15:00]

Paul Gustafson (University of British Columbia) **Daniel Daly-Grafstein** (University of British Columbia) **Conor Morrison** (University of British Columbia)

Bayesian Approaches to Causal Inference: The Present Position and the Path Ahead

Situation actuelle et perspectives d'avenir des approches bayésiennes d'inférence causale

It seems uncontroversial to note that solving causal inference problems demands principled management of complex uncertainty structures. Likewise, the Bayesian approach to statistical inference offers principled management of such structures. Thus it seems surprising that Bayesian approaches lack prominence in the causal inference realm. This talk offers some comments on why this is, and how the situation might change. One line of commentary addresses the foundational disconnect between Bayesian approaches and methods based on propensity scores. Another line addresses the level of parametric assumptions required in Bayesian tools for causal inference. Some highlights from two ongoing projects will be presented.

Il ne fait aucun doute que la résolution des problèmes d'inférence causale requiert une gestion raisonnée des structures d'incertitude complexes. De la même façon, l'approche bayésienne de l'inférence statistique offre une gestion raisonnée de telles structures. Il semble donc surprenant que les approches bayésiennes manquent de prédominance dans le domaine de l'inférence causale. Dans cette présentation, nous expliquerons pourquoi il en est ainsi et comment la situation pourrait changer. Le premier point porte sur la déconnexion fondamentale entre les approches bayésiennes et les méthodes basées sur les scores de propension. L'autre point concerne le niveau des hypothèses paramétriques nécessaires dans les outils bayésiens pour l'inférence causale. Enfin, nous présenterons les points forts de deux projets en cours.

Chair/Président: Devan G Becker

Date: Wednesday June 1 / mercredi 1 juin

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-13:45]

Matthew R.P. Parker (Simon Fraser University) **Jiguo Cao** (Simon Fraser University) **Laura L.E. Cowen** (University of Victoria) **Lloyd T. Elliott** (Simon Fraser University) **Junling Ma** (University of Victoria)

Estimating the Burden of COVID-19 in BC Using New Disease Analytic Multi-site Models

Estimer le fardeau de la COVID-19 en C.-B. à l'aide de nouveaux modèles multisites d'analyse de maladie

We provide a new multi-site model for disease analytics. This model uses publicly available disease counts data such as observed cases, recoveries among observed cases, and total deaths. These counts are used to estimate probability of recovery and probability of death among infected individuals, as well as several important population parameters over time including rate of spread, importation of external cases, and case detection probability. The model provides estimates of the total number of active COVID cases per region for each reporting interval. Simulation studies are used to validate the model, indicating that model parameters are identifiable. The multi-site model is applied to the five Health Authority regions of BC, Canada. We obtain simultaneous estimates for all five regions to produce an account of the pandemic over 30 weeks in the early pandemic. We compare multi-site model results to regional single-site models, showing improved model precision for the multi-site model.

Nous offrons un nouveau modèle multisite pour les analytiques de maladie. Ce modèle se sert de données de dénombrement de maladie publiques comme des cas observés, des rétablissements parmi les cas observés et le total de décès. Ces dénombrements sont utilisés pour estimer la probabilité de rétablissement et la probabilité de décès parmi les individus infectés, et aussi des paramètres de population importants au fil du temps comme le taux de propagation, l'importation de cas externes et la probabilité de détection de cas. Le modèle génère des estimations du nombre total de cas de COVID actif par région pour chaque intervalle de rapport. Nous adoptons des études en simulations pour valider le modèle et indiquer que les paramètres de modèle sont identifiables. Le modèle multisite est appliqué aux cinq régies de la santé de la C.-B., au Canada. Nous obtenons des estimations simultanées des cinq régions pour faire le compte rendu de la pandémie sur une période de 30 semaines au début de la pandémie. Nous comparons les résultats du modèle multisite à ceux des modèles régionaux à site unique, et démontrons la précision supérieure du modèle multisite.

[13:45-14:00]

Pingbo Hu (Western University)

Characterizing the COVID-19 Dynamics with a New Epidemic Model: Susceptible-Exposed-Symptomatic-Asymptomatic-Active-Removed

Caractériser les dynamiques de la COVID-19 à partir d'un nouveau modèle épidémique : susceptible, exposé, symptomatique, asymptomatique, actif et retiré

The coronavirus disease 2019 (COVID-19) has presented a tremendous threat to the public. It is important to investigate the transmission dynamics of COVID-19 to help understand the impact of the disease on public health and economy. In this talk, we develop a new epidemic model with unknown parameters to delineate the transmission process of COVID-19. The model accounts for asymptomatic infections as well as the lag

La maladie du coronavirus 2019 (COVID-19) a représenté une menace considérable pour le public. Il est important d'étudier les dynamiques de transmission de la COVID-19 afin de mieux comprendre son influence sur la santé publique et l'économie. Lors de cet exposé, nous développons un nouveau modèle épidémique avec paramètres inconnus pour définir le processus de transmission de la COVID-19. Le modèle tient compte des infections asymptomatiques ainsi que du décalage entre l'apparition de

between symptom onset and the confirmation date of infection. To reflect the transmission potential of an infected case, we derive the basic reproduction number from the proposed model. With the use of the reported number of confirmed cases, we adapt the iterated filter-ensemble adjustment Kalman filter algorithm to estimate the model parameters. To illustrate the use of the proposed model, we examine the COVID-19 data in Quebec for the period of April 2, 2020 to May 10, 2020 and further carry out sensitivity studies and simulation studies under a variety of settings.

symptôme et la date d'infection confirmée. Pour refléter le potentiel de transmission d'un cas infecté, nous dérivons le nombre de reproductions de bases à partir du modèle proposé. Au moyen du nombre rapporté de cas confirmés, nous adaptons l'algorithme de filtre Kalman ajusté à l'ensemble de filtres réitéré pour estimer les paramètres du modèle. Pour illustrer l'emploi du modèle proposé, nous observons les données de la COVID-19 au Québec pendant la période du 2 avril 2020 et 10 mai 2020, puis menons des études de sensibilité et des études en simulation selon une variété de contextes.

[14:00-14:15]

Leif Erik Lovblom (University of Toronto) **Laurent Briollais** (University of Toronto) **Bruce A. Perkins** (University of Toronto) **George Tomlinson** (University of Toronto)

A Joint Model for a Longitudinal Outcome and a Multistate Process Under Intermittent Observation, with Applications for Diabetes Complications

Un modèle conjoint pour un résultat longitudinal et un processus multi-états sous observation intermittente, avec des applications pour les complications du diabète

Uncertainties about the timing and co-development of diabetic microvascular complications could be addressed by joint models for longitudinal and event-time outcomes. However, such models are not fully-developed when the event-time follows a multistate process under intermittent observation. Our aim was to develop a joint model for this setting. Specifically, we formulated a shared random effects joint model with a linear mixed-effects submodel and a proportional intensities progressive 3-state Markov submodel, with interval censoring of entry into the second state and exact observation of entry into the absorbing state. Maximum likelihood estimation was used, with exploration of the functional forms for the baseline transition intensities, the linear mixed-effects submodel, and the association structure between the outcomes. An application of the model found predictive roles for albuminuria, glomerular filtration, and retinopathy on rates of progression through states of neuropathy.

Les incertitudes concernant le moment précis et le codéveloppement de complications microvasculaires diabétiques peuvent être résolues par des modèles conjoints pour les résultats longitudinaux et les évènements chronologiques. Cependant, ces modèles ne sont pas suffisamment développés pour les cas où l'évènement chronologique se base sur un processus multi-états selon une observation intermittente. Notre objectif est de concevoir un modèle conjoint adapté à cette situation. Plus précisément, nous avons formulé un modèle conjoint à effets aléatoires partagés avec un sous-modèle linéaire mixtes et un sous-modèle de Markov à trois états progressifs d'intensités proportionnelles, avec censure par intervalle dans le deuxième état et observation exacte de l'entrée dans l'état absorbant. Nous utilisons l'estimation du maximum de vraisemblance, l'exploration des formes fonctionnelles des intensités de transition de base, le sous-modèle linéaire mixtes et la structure d'association entre les résultats. L'application du modèle a trouvé des rôles prédictifs pour la protéinurie, la filtration glomérulaire et la rétinopathie relatives aux taux de progression à travers les états de neuropathie.

[14:15-14:30]

Mai Ghannam (University of Windsor) **Sévérien Nkurunziza** (University of Windsor)

Tensor Shrinkage Estimators in a Generalized Tensor Regression Model

Estimateurs à rétrécissement tensoriels dans un modèle de régression tensorielle généralisée

In this talk, we consider an estimation problem in a generalized tensor regression model with multi-mode covariates. We generalize the results in literature in five ways. First, we weaken assumptions underlying the results of the previous works. In particular, the depen-

Dans cet exposé, nous étudions un problème d'estimation dans un modèle de régression tensorielle généralisée à covariables multimodes. Nous généralisons de quatre façons les résultats récents. Premièrement, nous affaiblissons les présupposés des résultats pré-existants. Ainsi, la structure de dépendance du bruit

dence structure of the error and covariates are as weak as an L2-mixingale array, and the error term does not need to be uncorrelated with regressors. Second, we consider a more general constraint than the one in previous works. Third, we establish the asymptotic properties of the unrestricted tensor estimator (UE) and restricted tensor estimator (RE). Fourth, we propose a class of shrinkage estimators (SEs) in the case of tensor regression and we derive sufficient conditions for the SEs to dominate the UE. We also derive identities which are useful in studying the risk of SEs. Finally, we corroborate the results by some simulation studies of binary, Normal and Poisson data and we analyze a neuro-imaging dataset.

et des covariables est aussi faible que celle d'un L2-mixingale et le bruit n'est pas forcément non-corrélé avec les régresseurs. Deuxièmement, nous considérons une contrainte plus générale et établissons les propriétés asymptotiques des estimateurs sans restriction (UE) et avec restriction (RE). Troisièmement, nous proposons une classe d'estimateurs à rétrécissements (SEs) et établissons des conditions suffisantes pour que les SEs dominent l'UE. De plus, nous élaborons les identités qui permettent d'étudier le risk des SEs. Enfin, nous corroborons nos résultats par les simulations de données binaires, normales et de Poisson et analysons les données de neuro-imagerie.

[14:30-14:45]

Katherine Burak (University of Alberta) **Adam B. Kashlak** (University of Alberta)

Nonparametric confidence regions via the analytic wild bootstrap

Régions de confiance non paramétriques avec bootstrap sauvage analytique

The wild bootstrap is a nonparametric tool that can be used to estimate a sampling distribution in the presence of heteroscedastic errors, enabling us to compute confidence regions for regression parameters under non-i.i.d. models. While the wild bootstrap may perform well in these settings, its obvious drawback is a lack of computational efficiency. The wild bootstrap requires a large number of bootstrap replications, making the use of this tool impractical when dealing with big data. We introduce the analytic wild bootstrap (ANWB), which provides a nonparametric alternative way of constructing confidence regions for regression parameters. The ANWB is superior to the wild bootstrap from a computational standpoint, while exhibiting similar finite sample performance. We report simulation results and test the ANWB on a real dataset and compare its performance with that of other standard approaches. We also discuss the extension of the ANWB to the penalized regression setting.

Pour estimer une distribution d'échantillonnage en présence d'erreurs hétéroscédastiques, on recourt souvent à une méthode non paramétrique appelée bootstrap sauvage (≪ wild bootstrap ≫). Cette méthode permet, entre autres, de calculer des régions de confiance pour des paramètres de régression lorsque les variables ne sont pas indépendantes ni identiquement distribuées. Si le bootstrap sauvage peut donner de bons résultats dans ces contextes, il manque manifestement d'efficacité de calcul. En effet, comme sa mise en application requiert un grand nombre de réplications bootstrap, il devient difficile à utiliser en présence de mégadonnées. Nous proposons le bootstrap sauvage analytique, qui offre une solution de rechange non paramétrique à la conception de régions de confiance pour les paramètres de régression. Le bootstrap sauvage analytique est plus efficace que le bootstrap sauvage sur le plan des calculs, tout en présentant des résultats semblables pour les échantillons finis. Nous présentons des résultats de simulation et testons le bootstrap sauvage analytique sur un ensemble de données réelles et comparons ses résultats à ceux d'autres approches standard. Nous examinons également l'extension du bootstrap sauvage analytique au cadre de la régression pénalisée.

[14:45-15:00]

Meixi Chen (University of Waterloo) **Martin Lysy** (University of Waterloo) **Reza Ramezan** (University of Waterloo)

Decoding Multi-Neuronal Activities Through Latent Factor Models

Modèles de facteurs latents pour le décodage d'activités multineuronales

Nerve cells (a.k.a. neurons) communicate through spike trains which are sequences of consecutive electrochemical waves generated by each neuron. These spike trains code information in the brain. Recent quanti-

Les cellules nerveuses (aussi appelées neurones) communiquent par trains de pointes, soit des séquences d'ondes électrochimiques consécutives générées par chaque neurone. Ces trains de pointes codent l'information dans le cerveau. Des modèles quantitatifs

tative models have helped us understand neuroscientific phenomena through time-dependent interactions in neuronal ensembles; however, the scalability and interpretability of these models are still challenging. We present a novel hierarchical factor model to study interactions in small neuronal populations recorded simultaneously. We model the neuronal activities through correlated Wiener processes, which themselves depend on latent factors determining the neuronal clusters. Addressing the scalability problem, we demonstrate efficient ways to tackle the computational challenges imposed by high dimensional integration and inversions of large matrices. Through simulations and real data analyses, we show that our model is scalable and can accurately recover neuronal clusters.

récents aident à mieux comprendre le phénomène par le biais d'interactions dépendantes du temps dans des ensembles neuronaux ; cependant, la mise à l'échelle et l'interprétabilité de ces modèles posent toujours problème. Nous présentons un nouveau modèle hiérarchique de facteurs pour l'étude d'interactions enregistrées simultanément dans de petites populations neuronales. Nous modélisons les activités neuronales avec des processus de Wiener corrélés, qui dépendent eux-mêmes de facteurs latents déterminant les groupes neuronaux. En ce qui concerne la mise à l'échelle, nous faisons valoir des moyens efficaces de traiter les problèmes computationnels imposés par une intégration à haute dimension et des inversions de grandes matrices. Des simulations ainsi que des analyses de données réelles permettent de montrer que notre modèle est extensible et peut recouvrir des groupes neuronaux avec précision.

Chair/Président: Omidali Aghababaei Jazi

Date: Wednesday June 1 / mercredi 1 juin

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-13:45]

Marie-Pier Lemieux (Statistics Canada)

2021 Canadian Census: Using an Agile Non-Response Management Strategy to Obtain Quality Data during a Pandemic

Recensement Canadien de 2021 : Utilisation d'une stratégie agile pour la gestion de la non-réponse afin d'obtenir des résultats de qualité en temps de pandémie

Every five years, Statistics Canada conducts a census to obtain the count of the population and housing in Canada. The collection methodology of the Census has evolved considerably over the years to take advantage of new technologies, to reduce response burden, improve the efficiency of the program and increase data quality. The last census was carried out from May to August 2021 in the middle of a pandemic. In this unprecedented context, it was important to collect quality data in order to provide a socio-demographic portrait of the country. The tolerance process was developed to obtain quality data uniformly at detailed geographic levels and closely monitors response rates and other quality indicators, and flags areas across the country where the data collected meets the quality criteria. Nonresponse follow-up can then stop in these locations so that collection efforts can be redirected to other areas that need it most. This strategy allows for agile and flexible collection management that can adapt to various situations, including a pandemic. This presentation will focus on the Canadian experience of conducting a census of population during a pandemic, and the strategies used to manage collection to obtain high quality results uniformly across Canada.

Tous les cinq ans, Statistique Canada mène un recensement pour obtenir un compte de la population et des logements au Canada. La méthodologie de collecte du recensement a considérablement évolué au fil des années afin de prendre avantage de nouvelles technologies, de réduire le fardeau de réponse, d'améliorer l'efficacité du programme et d'accroître la qualité des données. Le dernier recensement a été réalisé de mai à août 2021 en plein milieu d'une pandémie. Dans ce contexte sans précédent, il était primordial de collecter des données de qualité afin d'obtenir un portrait socio-démographique du pays. Le processus de la tolérance a été développé afin d'obtenir des données de qualité, et ce de façon uniforme à des niveaux géographiques détaillés, d'effectuer un suivi serré des taux de réponse ainsi que de d'autres indicateurs de qualité et finalement d'identifier les régions à travers le pays où les données collectées atteignent les critères de qualité. Le suivi de non-réponse peut alors se terminer dans ces endroits afin que les efforts de collecte soient redirigés à d'autres régions qui en ont le plus besoin. Cette stratégie permet une gestion de collecte agile et flexible permettant de s'adapter à diverses situations, dont celle d'une pandémie. Cette présentation portera sur l'expérience canadienne d'un recensement de la population en temps de pandémie et des stratégies utilisées pour gérer la collecte en vue d'obtenir des résultats de qualité et uniforme à travers le Canada.

[13:45-14:00]

Audrey Béliveau (University of Waterloo)

Design-Unbiased Trapezoid Area-Under-the-Curve Estimators for Estimating Salmon Escapement

Estimateurs de l'aire sous la courbe par la méthode des trapèzes qui soient sans biais par rapport au plan afin d'estimer l'échappée de saumons

The trapezoid area-under-the-curve (TAUC) method to estimating salmon escapement involves linearly interpolating periodic counts of live salmon and calculating the area under the interpolated curve. Currently, there is not

La méthode d'évaluation de l'aire sous la courbe par trapèzes employée pour l'estimation de l'échappée de saumons consiste en l'interpolation linéaire d'un dénombrement périodique de saumons vivants et au calcul de l'aire sous la courbe interpolée.

a statistically founded recommended practice for the selection of the sampling days. In practice, few days are usually sampled due to budget constraints which can result in consequential bias when sampling days are chosen deterministically. An important perspective that has yet to be explored is to eliminate this bias by considering probabilistic sampling mechanisms for the sampling days. In this work, we show that systematic sampling, simple random sampling or Bernoulli sampling, combined with a judicious choice of end-adjustments for the TAUC estimator, allow unbiased estimation of the AUC. In addition, a variance estimator is proposed. The theoretical results are supported by a simulation study and illustrated on salmon counts collected in the Pacific Northwest.

Présentement, il n'existe aucune pratique statistiquement fondée et recommandée pour la sélection des jours d'échantillonnage. En pratique, peu de jours sont échantillonnés en raison des contraintes budgétaires, ce qui mène à des biais corrélatifs lorsque les jours d'échantillonnage sont choisis de façon déterministe. Une perspective importante qui reste encore à être étudiée est l'élimination de ces biais en considérant les mécanismes d'échantillonnage probabiliste pour les jours d'échantillonnage. Dans le cadre de ce travail, nous démontrons que l'échantillonnage systématique, aléatoire simple ou de Bernoulli, combinés à un choix judicieux de rajustements de fin pour l'estimateur par méthode des trapèzes, peut permettre une estimation sans biais de l'aire sous la courbe. De plus, nous proposons un estimateur de variance. Les résultats théoriques sont soutenus par des études en simulation et illustrés à partir de dénombrements de saumons provenant du Nord-Ouest du Pacifique.

[14:00-14:15]

Thomas Yoon (Statistics Canada)

Modernization of the Canadian Census: An Administrative Data-Driven Approach to Defining Households

Modernisation du recensement canadien : Une approche axée sur les données administratives pour définir les ménages

Many national statistical offices are conducting research to better utilize administrative records, defined as data collected as part of administering a program or service. Administrative records offer the possibility to complement the traditional survey enumeration approach and potentially improve quality and efficiency in estimation. A combined census is currently under research at Statistics Canada whereby administrative data and traditional data collection are used jointly to enumerate the population. One part of ongoing census research is the household model, which aims to group administrative individuals into ''households" using statistical models, and to evaluate their quality as compared to traditional census outputs. The talk will host the methodology and the evaluation of the key quality indicators of the household model approach.

De nombreux bureaux nationaux de statistiques mènent des recherches pour mieux utiliser les données administratives, définis comme des données recueillies dans le cadre de l'administration d'un programme ou d'un service. Les données administratives offrent la possibilité de compléter l'approche traditionnelle de collecte par enquête et d'améliorer potentiellement la qualité et la précision de l'estimation. Un recensement combiné fait actuellement l'objet de recherches à Statistique Canada dans le cadre duquel les données administratives et la collecte de données traditionnelles sont utilisées conjointement pour dénombrer la population. Une partie de la recherche en cours sur le recensement est le modèle des ménages, qui vise à regrouper les individus administratifs en « ménages » à l'aide de modèles statistiques, et à évaluer leur qualité par rapport aux résultats traditionnels du recensement. L'exposé présentera la méthodologie et l'évaluation des indicateurs clés de qualité de l'approche modèle ménage.

[14:15-14:30]

Abel C. Dasylva (Statistics Canada) **Arthur Goussanou** (Statistics Canada)

A new model for the automated identification of duplicate records

Nouveau modèle d'identification automatique des enregistrements en double

Duplicate records are records from the same unit in a given data source, regardless of whether they are identical. Their identification is required when the source is used to produce official statistics, such as a sampling frame or a census. To date, many Bayesian models have been described to perform this task in an au-

Les enregistrements en double proviennent de la même unité dans une source de données précise, qu'ils soient identiques ou non. Il est nécessaire de les identifier lorsque la source est utilisée pour produire des statistiques officielles (cadre d'échantillonnage ou recensement). À ce jour, on a décrit de nombreux modèles bayésiens afin de réaliser cette tâche de manière automatisée. Cependant,

tomated manner. Yet, they involve computer-intensive procedures and tend to assume that the linkage variables are conditionally independent, when this is seldom the case in practice. To overcome these limitations, a new model is described for applications, where one can reasonably assume that each unit is associated with at most two records because duplication is rare, as for persons living in private dwellings, in the census of population. The duplication is modeled through the number of links adjacent to a record from a given unit, as in a recent model of linkage errors, while extending the latter to account for the multiplicity of false positives from some other unit.

ils nécessitent des procédures informatiques intensives et partent généralement du principe que les variables de liaison sont conditionnellement indépendantes, alors que c'est rarement le cas dans la pratique. Pour pallier ces contraintes, nous décrivons un nouveau modèle d'applications, dans lequel on peut raisonnablement supposer que chaque unité est associée au maximum à deux enregistrements, car les doublons sont rares, comme c'est le cas pour les personnes vivant dans des logements privés, dans le cadre du recensement de la population. Nous modélisons la duplication par le nombre de liens adjacents à un enregistrement d'une unité donnée, comme dans un modèle récent d'erreurs de liaison tout en étendant ce dernier afin de tenir compte de la multiplicité des faux positifs provenant d'une autre unité.

[14:30-14:45]

Yiran Wang (University of Waterloo) **Martin Lysy** (University of Waterloo) **Audrey Béliveau** (University of Waterloo)

Genetic Mark-Recapture Methods for Estimating Seasonal River Run Size of Stock Populations

Méthodes de marquage et de recapture génétiques pour l'estimation de la taille saisonnière de l'effectif à la montaison en rivière de stocks

Genetic mark-recapture (GMR) is a statistical technique used in estimating population size in ecology. By combining genetic data on the relative abundance of species from a sample with population counts obtained for some of the species, GMR allows the estimation of the total population size and the contributions of each species. The current method is based on the Lincoln-Petersen estimator and provides an asymptotically unbiased estimate for the total population size. However, the variance estimator does not account for the uncertainty in the sampling process of the genetics data. As a result, this approach can suffer from a significantly underestimated variance. In this work, we propose a novel Bayesian GMR framework to address this issue. The Bayesian framework can explicitly incorporate the sampling error in the genetic sample and lends itself nicely to combining additional sources of data into a single model. Simulation studies and real data analysis are conducted in the study.

La méthode de marquage et de recapture génétique est une technique statistique qui sert à estimer la taille des populations en écologie. Grâce à la combinaison des données génétiques sur l'abondance relative des espèces d'un échantillon avec les recensements de population obtenus pour certaines des espèces, la méthode de marquage et de recapture génétique permet d'estimer la taille totale de la population et les contributions de chaque espèce. La méthode actuelle repose sur l'estimateur de Lincoln-Petersen et fournit une estimation asymptotiquement sans biais de la taille totale de la population. Par conséquent, cette approche peut se traduire par une sous-estimation significante de la variance. Dans cette étude, nous proposons un nouveau cadre de marquage et de recapture génétique bayésien pour résoudre ce problème. Le cadre bayésien peut intégrer explicitement l'erreur d'échantillonnage de l'échantillon génétique, et se prête bien à la combinaison de sources de données supplémentaires dans un modèle unique. Dans le cadre de ces travaux, nous menons des études de simulation et des analyses de données réelles.

[14:45-15:00]

Inesh Prabuddha Munaweera Arachchilage (University of Manitoba) **Saman Muthukumarana** (University of Manitoba) **Darren Gillis** (University of Manitoba) **Les N. Harris** (Fisheries and Oceans Canada)

Bayesian Multi-state Capture-recapture Modelling for Estimating Survival Probabilities of Arctic Char using Acoustic Telemetry Data

Modélisation bayésienne multi-états de capture-recapture pour l'estimation des probabilités de survie de l'omble chevalier à l'aide de données de télémétrie acoustique

Recent advances in animal tracking technologies such as acoustic telemetry (AT) have enabled researchers to

Les progrès récents des technologies de suivi des animaux, telles que la télémétrie acoustique (TA), ont permis aux chercheurs de

collect enormous amounts of data on animal movement and habitat use over large geographic scales. More recently, AT data have been used to estimate demographic parameters such as survival probability and population size, with comparable or better precision than conventional capture-mark-recapture studies. The most popular method for estimating survival probabilities with AT data has been the Cormack-Jolly-Seber (CJS) model. However, the estimated survival probabilities with CJS models suffer from low precision when the recapture rate is low in certain regions. In this context, multi-state mark-recapture models can be used to deal with sparseness in data by borrowing information across regions. In this study, we use Bayesian multi-state mark-recapture models combined with AT data to study the survival of Arctic Char in different habitats of the Cambridge Bay region.

collecter d'énormes quantités de données sur les mouvements des animaux et l'utilisation de leur habitat à de grandes échelles géographiques. Plus récemment, les données TA ont été utilisées pour estimer des paramètres démographiques tels que la probabilité de survie et la taille des populations, avec une précision comparable ou supérieure à celle des études conventionnelles de capture-marquage-recapture. La méthode la plus populaire pour estimer les probabilités de survie avec les données TA a été le modèle Cormack-Jolly-Seber (CJS). Cependant, les probabilités de survie estimées avec les modèles CJS sont peu précises lorsque le taux de recapture est faible dans certaines régions. Dans ce contexte, on peut utiliser des modèles de marquage-recapture multi-états pour gérer la rareté des données en empruntant des informations d'autres régions. Dans cette étude, nous utilisons des modèles bayésiens de marquage-recapture multi-états combinés à des données TA pour étudier la survie de l'omble chevalier dans différents habitats de la région de Cambridge Bay.

Missing Data, Causal Inference, and New Algorithms for Differential Equations
Données manquantes, inférence causale et nouveaux algorithmes pour les équations différentielles

Chair/Président: Nikola Surjanovic

Date: Wednesday June 1 / mercredi 1 juin

Time/Heure: 13:30-14:45

Abstract/Résumé

[13:30-13:45]

Abdoulaye Dioni (Université Laval) **Alexandre Bureau** (Université Laval) **Lynne Moore** (Université Laval) **Aida Eslami** (Université Laval)

Development of a method for missing not at random

Développement d'une méthode pour les données manquantes non aléatoirement

In several domains, in particular health, we are interested to reduce the bias due to missing data. Among the mechanisms, missing not at random (MNAR) is the most problematic. The objective is to develop a simple, valid, and accessible method to analyze data under MNAR, and to implement it in an R Package. First, we build an imputation method under MAR hypothesis. Then, the imputed data are transformed by a Bayesian approach to incorporate a priori information. The final analysis is performed according to Rubin's rule for combinations of variables. The entire approach aims to assess the robustness of the MAR. Application is on a simulated dataset and data from the Canada trauma system. The approach is simple and will serve as a tool for non-statisticians but the choice of prior parameters, the presentation of results, and their interpretation remain a challenge.

Dans plusieurs domaines en particulier la santé, des efforts sont faits pour réduire le biais dû aux données manquantes. Parmi les mécanismes, les données manquantes non aléatoirement (MNAR) est le plus problématique. Notre objectif est de développer une méthode simple, valide et accessible pour analyser les données sous MNAR puis l'implanter dans un module de R. Comme méthode, nous construisons un modèle d'imputation sous l'hypothèse MAR. Ensuite, les données imputées sont transformées par approche bayésienne pour incorporer une information a priori. L'analyse finale est faite selon la règle de Rubin pour des combinaisons de variables. L'ensemble de l'approche vise à évaluer la robustesse de MAR. L'approche sera appliquée sur un jeu de données simulées et sur des données de traumatisme au Canada. Certes elle est simple et servira d'outil aux non-statisticiens mais le choix des paramètres a priori, la présentation des résultats et leur interprétation restent un défi.

[13:45-14:00]

Renny Doig (Simon Fraser University) **Liangliang Wang** (Simon Fraser University)

Probabilistic Numerical Solution of Differential Equations as a Remedy for Discretization-Induced Bias

Solution numérique probabiliste d'équations différentielles comme remède au biais induit par la discrétisation

Ordinary differential equations (ODEs) are often used in the applied sciences to describe the evolution of a variable(s). In many practical settings the solution to these systems must be approximated by numerical techniques, e.g. Runge-Kutta (RK) methods. The error of a single step of these methods is the local truncation error (LTE). In simple ODEs the LTE at each step is negligible. Often statistical models of ODE systems rely on this by ignoring the bias induced by this LTE. However, in systems where small changes to the state of the system can have a drastic impact on future values, the LTE and the bias induced by it, can no longer be dismissed. Here we

Les équations différentielles ordinaires (EDO) sont souvent utilisées dans les sciences appliquées pour décrire l'évolution d'une ou plusieurs variables. Dans de nombreux contextes pratiques, la solution de ces systèmes doit être approchée par des techniques numériques, par exemple les méthodes de Runge-Kutta (RK). L'erreur d'une seule étape de ces méthodes est l'erreur de troncature locale (ETL). Dans les EDO simples, l'ETL à chaque étape est négligeable. Souvent, les modèles statistiques des systèmes EDO s'appuient sur ce fait en ignorant le biais induit par cette ETL. Cependant, dans les systèmes où de petits changements de l'état du système peuvent avoir un impact drastique sur les valeurs futures, l'ETL et le biais induit par celle-ci ne peuvent plus être écartés.

present a probabilistic ODE solver which incorporates variability designed to reflect the LTE into a deterministic RK method. By combining this probabilistic solver with SMC, we demonstrate how accounting for LTE can provide estimates of ODE trajectories that are more robust to these errors than estimates that neglect it.

Nous présentons ici un solveur EDO probabiliste qui intègre la variabilité conçue pour refléter la LTE dans une méthode de RK déterministe. En combinant ce solveur probabiliste avec la SMC, nous démontrons comment la prise en compte de l'ETL peut fournir des estimations des trajectoires d'EDO qui sont plus robustes à ces erreurs que les estimations qui la négligent.

[14:00-14:15]

Jonathan Ramkissoon (University of Waterloo) **Martin Lysy** (University of Waterloo)

Smoothly Differentiable Particle Filters for Stochastic Differential Equations

Filtres à particules facilement différentiables pour des équations différentielles stochastiques (EDS)

Parameter inference for stochastic differential equations is challenging due to intractable likelihood functions that integrate over the entire latent space. Particle filters offer a principled solution by providing a consistent estimate of SDE log-likelihood. However, the multinomial resampling step traditionally used in particle filters is not smoothly differentiable with respect to SDE parameters, which can be problematic in many likelihood-based inference methods. In this work we propose a smoothly differentiable particle filter by replacing the resampling step with a multivariate Normal approximation and utilizing the reparameterization trick. This enables a host of gradient based methods for parameter inference, which we compare on different SDE examples.

L'inférence de paramètres pour les équations différentielles stochastiques représente un défi en raison des fonctions de vraisemblance intraitables qui intègrent l'espace latent entier. Les filtres à particules offrent une solution de principe en fournissant une estimation convergente de la log-vraisemblance des EDS. Cependant, l'étape de rééchantillonnage multinomial traditionnellement utilisé dans les filtres à particules n'est pas facilement différentiable à l'égard aux paramètres des EDS, ce qui peut poser problème dans bon nombre de méthodes d'inférence basées sur la vraisemblance. Dans le cadre de ce travail, nous proposons un filtre à particules facilement différentiable, en remplaçant l'étape de rééchantillonnage par une approximation de la Normale multivariée et en utilisant l'astuce de reparamétrisation, ce qui permet de comparer une série de méthodes des gradients pour l'inférence paramétrique à l'aide de divers exemples d'EDS.

[14:15-14:30]

Mohan Wu (University of Waterloo) **Martin Lysy** (University of Waterloo)

Parameter Inference for Differential Equations using Bridge Proposal

Inférence de paramètres pour des équations différentielles à l'aide d'une proposition de pont

Parameter inference for ordinary differential equations (ODEs) involves the evaluation of the likelihood function for each ODE solution. While this solution is typically approximated by deterministic algorithms, new research indicates that probabilistic solvers produce more reliable estimates by better considerations of numerical errors. A particularly simple and effective probabilistic method uses Kalman filtering to obtain the ODE solution. However, the solver does not condition on the observed data, which can lead to extreme sensitivity of the likelihood function to model parameters. Here we propose the bridge proposal which accounts for these data in the filtering algorithm in a computationally efficient manner. Several examples are used to demonstrate the effectiveness of this approach.

L'inférence de paramètres pour des équations différentielles ordinaires (ODE) fait appel à l'évaluation de la fonction de vraisemblance pour chaque solution aux ODE. Même si cette solution est généralement une approximation par algorithmes déterministes, une nouvelle recherche indique que les solveurs probabilistes produisent des estimations plus fiables par une meilleure prise en compte des erreurs numériques. Une méthode probabiliste particulièrement simple et efficace utilise le filtre de Kalman pour obtenir une solution aux ODE. Le solveur n'est toutefois pas conditionné aux données d'observation, ce qui peut entraîner une sensibilité extrême de la fonction de vraisemblance pour modéliser les paramètres. Nous proposons ici une structure de pont qui rend compte de ces données dans l'algorithme de filtre d'une façon computationnelle efficace. Plusieurs exemples illustrent l'efficacité de cette approche.

[14:30-14:45]

Pranav Subramani (University of Waterloo) **Jonathan Ramkissoon** (University Of Waterloo) **Mohan Wu** (University Of

Missing Data, Causal Inference, and New Algorithms for Differential Equations
Données manquantes, inférence causale et nouveaux algorithmes pour les équations différentielles

Waterloo) **Martin Lysy** (University Of Waterloo)

A Method for Parameter Inference for Stochastic Differential Equations

Méthode d'inférence des paramètres pour équations différentielles stochastiques

Inference for Stochastic Differential Equations has seen a lot of improvement in the recent past with modern hardware and better algorithms. In this talk, we focus on developing a computationally scalable approach to stochastic differential equations using Particle Filters. Particle filters allow us to obtain an estimate of the log-likelihood and using tools from automatic differentiation, we are able to obtain partial derivatives of the log-likelihood with respect to the parameters. We then apply a stochastic optimization algorithm to the parameters to converge to a local optima since most of the parameter trajectories are non-convex. We implement this in a modern high-performance computing framework, JAX which utilizes automatic differentiation and Just-in Time compilation which provides tremendous speed-ups. Finally, we evaluate the performance of this approach on multiple problems and evaluate the results.

L'inférence pour équations différentielles stochastiques a connu beaucoup d'améliorations récemment grâce au matériel moderne et à de meilleurs algorithmes. Dans cet exposé, nous nous concentrons sur le développement d'une approche computationnellement extensible pour les équations différentielles stochastiques qui utilise des filtres de particules. Les filtres de particules nous permettent d'obtenir une estimation de la log-vraisemblance et, en utilisant des outils de différenciation automatique, d'obtenir des dérivées partielles de la log-vraisemblance relativement aux paramètres. Nous appliquons ensuite un algorithme d'optimisation stochastique aux paramètres pour converger vers un optimum local puisque la plupart des trajectoires des paramètres sont non convexes. Nous implémentons cette méthode dans un cadre moderne de calcul à haute performance, JAX, qui utilise la différenciation automatique et la compilation Just-in-Time, ce qui permet des accélérations considérables. Enfin, nous évaluons les performances de cette approche sur plusieurs problèmes et en évaluons les résultats.

Recent Developments in Methodology and Applications of Mixture Models
Développements récents en méthodes et applications des modèles de mélange

Chair/Président: Abbas Khalili

Organizer/Responsable: Abbas Khalili

Date: Wednesday June 1 / mercredi 1 juin

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-16:00]

Nhat Ho (University of Texas, Austin)

Bayesian Sieves and Excess Mass Behavior in Infinite Mixtures

Cribles bayésiens et comportement de masse excédante dans des mélanges infinis

Dirichlet Process mixture models (DPMM) have been an important modeling toolbox for numerous practical domains. Despite their popularity, there are three important inferential questions to ask. (I) How do you choose between heavy or light-tailed kernels for appropriate inference? (II) Suppose we allow the number of components to grow with the sample size, can we efficiently estimate the parameters corresponding to components in an arbitrary region of the parameter space? (III) Gaussian kernels are the most popular choices for model fitting, however, the parameter estimation rates for Gaussian kernels tend to be very low. There is, therefore, an inconsistency in practice and theory. In this talk, we aim to address these questions via a novel notion: Orlicz-Wasserstein distance, a generalization of the Wasserstein distance corresponding to Orlicz norms. We establish lower bounds of Hellinger distance between the mixing densities based on the Orlicz-Wasserstein distance between the mixing measures. Then, we use these results to establish the posterior convergence rates of mixing measures in DPMM under both the misspecified and well-specified parameter space. Our rates under Orlicz-Wasserstein distance provide faster local rates of contraction of parameters in comparison to the uniform global contraction rate of parameters under Wasserstein.

Les modèles de mélange de processus de Dirichlet (DPMM) sont des boîtes à outils importantes en modélisation pour de nombreux domaines pratiques. Malgré leur popularité, il y a trois questions à poser relatives à l'inférence : (1) comment choisir entre des noyaux à queue longue ou à queue lourde pour une inférence appropriée ? (2) En supposant que le nombre de composantes croît avec la taille d'échantillon, pouvons-nous estimer efficacement les paramètres correspondants aux composés dans une région arbitraire de l'espace de paramètre ? (3) Les noyaux gaussiens sont les choix les plus employés pour l'ajustement de modèle, pourtant le taux d'estimation de paramètre de ceux-ci a tendance à être très bas. Conséquemment, il y a incohérence entre la théorie et la pratique. Lors de cet exposé, nous cherchons à aborder ces questions à l'aide d'une nouvelle notion : la distance Orlicz-Wasserstein, une généralisation de la distance de Wasserstein correspondant aux normes d'Orlicz. Nous établissons des limites inférieures de la distance d'Hellinger entre les densités de mélanges basées sur la distance Orlicz-Wasserstein entre les mesures de mélanges. Ensuite, nous utilisons ces résultats pour établir les taux de convergence postérieurs des mesures de mélange dans les DPMM à la fois pour un espace de paramètre bien spécifié et mal spécifié. Nos taux selon la distance Orlicz-Wasserstein procurent des taux locaux de contraction de paramètres plus rapides par rapport au taux global de contraction de paramètres selon Wasserstein.

[16:00-16:30]

Tudor A. Manole (Carnegie Mellon University) **Cody Mazza-Anthony** (Shopify) **Nhat Ho** (University of Texas, Austin) **Abbas Khalili** (McGill University)

Order Selection in Finite Mixture of Regression Models

Sélection de l'ordre dans un mélange fini de modèles de régression

We propose a new penalized likelihood approach for estimating the number of components in finite mixture of

Nous proposons une nouvelle approche de vraisemblance pénalisée pour l'estimation du nombre de composantes dans un mélange

regression models, called the Group-Sort-Fuse (GSF) procedure. Unlike methods which fit and compare models with varying orders using criteria involving model complexity, our method directly penalizes a continuous function of the model parameters. Specifically, given a conservative upper bound on the true order, the GSF groups and sorts regression parameters across the various mixture components, in order to merge those which are redundant. For a broad class of finite mixture of regression models, we show that the GSF is consistent in estimating the true number of components, and nearly achieves the pointwise parametric convergence rate for parameter estimation, as measured by a suitable Wasserstein distance over the space of finite mixing measures.

fini de modèles de régression, appelée procédure Group-Sort-Fuse (GSF). Contrairement aux méthodes qui ajustent et comparent des modèles de divers ordres en utilisant des critères faisant intervenir la complexité du modèle, notre méthode pénalise directement une fonction continue des paramètres du modèle. Plus précisément, étant donné une limite supérieure prudente de l'ordre véritable, la procédure GSF regroupe et trie les paramètres de régression dans l'ensemble des diverses composantes du mélange, afin de fusionner celles qui sont redondantes. Dans une grande classe de modèles de mélanges finis de régression, nous montrons que la procédure GSF est convergente pour l'estimation du nombre réel de composantes et atteint presque le taux de convergence paramétrique ponctuelle pour l'estimation des paramètres, tel qu'il est mesuré par une distance de Wasserstein appropriée pour l'espace mesuré de mélanges finis.

[16:30-17:00]

Alejandro Murua (Université de Montréal)

A Bayesian Semi-parametric Mixture of Survival Regression Model for Survival Prediction

Un mélange bayésien semi-paramétrique de modèles de survie pour la prédiction de survie

We consider a Bayesian semi-parametric survival regression model with latent partitions. Our goal is to predict survival. We propose the Potts clustering model as a prior on the covariates space in order to drive cluster formation on individuals. For any given partition, our model assumes an interval-wise Exponential distribution for the baseline hazard rate. The number of intervals is unknown. It can be estimated with a fused-lasso type penalty given by a sequential double exponential prior. Estimation and inference are done with the aid of Markov chain Monte Carlo. To simplify the computations, we use the Laplace's integral approximation method to estimate some constants, and to propose parameter updates within Markov chain Monte Carlo. We illustrate the methodology with an application to cancer survival. This is joint work with Danae Martinez-Vargas (MSc).

Nous considérons un modèle bayésien semi-paramétrique de régression de survie avec des partitions latentes. Notre but est de prédire la survie. Nous proposons le modèle de groupement de Potts comme a priori sur l'espace des covariables afin de piloter la formation de grappes sur les individus. Pour chaque partition donnée, nous supposons une distribution exponentielle par intervalle pour le taux de risque de base. Le nombre d'intervalles est inconnu, mais peut être estimée avec une pénalité de type lasso fusionné donnée par une a priori exponntielle double séquentielle. L'estimation et l'inférence sont faites à l'aide de méthodes Monte Carlo par chaîne de Markov. Pour simplifier les calculs, nous utilisons la méthode d'approximation de Laplace pour estimer certaines constantes, et proposer des mises à jour de paramètres à l'intérieur des méthodes Monte Carlo par chaîne de Markov. Nous illustrons la méthodologie par une application à la survie au cancer. Il s'agit d'un travail conjoint avec Danae Martinez-Vargas (MSc).

Recent Advances on Model Assessment in Recurrent Event Analysis
Progrès récents en évaluation des modèles pour l'analyse des événements récurrents

Chair/Président: Hua Shen

Organizer/Responsable: Hua Shen

Date: Wednesday June 1 / mercredi 1 juin

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-16:00]

Eleanor M. Pullenayegum (Hospital for Sick Children)

The Role of Recurrent Event Models in Handling Longitudinal Data Subject to Irregular Observation: Determining the Assessment Mechanism and Undertaking Sensitivity Analysis.

Rôle des modèles d'événements récurrents dans le traitement des données longitudinales observées de façon irrégulière : détermination du mécanisme d'évaluation et analyse de sensibilité

Irregular observation in longitudinal data is common, with assessment times forming a recurrent event process. The analytic approach to such longitudinal data depends on whether the timing of assessments is independent of the outcomes (assessment completely at random, or ACAR), conditionally independent of outcomes given previously observed data (assessment at random, or AAR), or dependent on outcomes given previously observed data (assessment not at random, or ANAR). This talk will discuss the role of recurrent event models in discerning the assessment mechanism, as well as their use in sensitivity analysis when assessment is not at random.

L'observation irrégulière des données longitudinales est courante, les moments d'évaluation formant un processus d'événements récurrents. L'approche analytique de ces données longitudinales varie selon que le moment des évaluations est indépendant des résultats (évaluation complètement aléatoire), conditionnellement indépendant des résultats au vu des données observées précédemment (évaluation aléatoire), ou dépendant des résultats au vu des données observées précédemment (évaluation non aléatoire). Dans cette présentation, nous nous nous pencherons sur le rôle des modèles d'événements récurrents pour discerner le mécanisme d'évaluation, ainsi que leur utilisation dans l'analyse de sensibilité lorsque l'évaluation n'est pas aléatoire.

[16:00-16:30]

Candemir Cigsar (Memorial University of Newfoundland)

Model Assessment for Dynamic Recurrent Event Processes with Dependent Gap Times

Évaluation de modèles pour des processus dynamiques d'événements récurrents avec des laps de temps dépendants

There is an increasing interest in the analysis of recurrent event data through dynamic models. Most work in this area has focused on formulating and fitting models. Comparatively little work has been done on the assessment of models. Because of the lack of robust methods, model assessment is especially important for dynamic models. In this talk, I will introduce two model assessment methods for dynamic recurrent event models, in which copulas are used to model the dependency between gap times. The first method includes a two-stage procedure. The first stage checks the adequacy of the assumed dependence structure between gap times. The second stage tests the fit of marginal distributions

L'analyse d'événements récurrents à l'aide de modèles dynamiques soulève un intérêt croissant. La plupart des travaux dans ce domaine sont centrés sur la formulation et l'ajustement de modèles. Par comparaison, peu d'études sont menées sur l'évaluation des modèles. En raison d'un manque de méthodes robustes, cette évaluation est particulièrement importante pour les modèles dynamiques. Notre exposé présente deux méthodes d'évaluation pour les modèles dynamiques d'événements récurrents dans lesquelles les copules sont utilisées pour modéliser la dépendance entre les laps de temps. La première méthode est une procédure à deux étapes : l'étape initiale sert à vérifier si la structure de dépendance hypothétique entre les laps de temps est adéquate, tandis que la deuxième évalue l'ajustement des distri-

of gap times by using supremum or quadratic test statistics. The second method is based on comparison of conditional marginal distributions of gap times with their nonparametric counterparts. An illustration based on a study of asthma attacks in children will be given.

butions marginales des laps de temps en utilisant des statistiques de test de supremum ou de formes quadratiques. La deuxième méthode est basée sur une comparaison des distributions marginales conditionnelles des laps de temps avec leurs contreparties non paramétriques. Une étude sur les crises d'asthme chez les enfants servira à illustrer nos propos.

Environmental Data Science: Growth and Opportunities
Science des données environnementales : Croissance et opportunités

Chair/Président: Wesley S. Burr

Organizer/Responsable: Wesley S. Burr

Date: Wednesday June 1 / mercredi 1 juin

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-16:00]

Allison Horst (University of California, Santa Barbara) **Samantha Csik** (National Center for Ecological Analysis & Synthesis) **Jamie Montgomery** (University of California, Santa Barbara)

Filling a Training Gap in Environmental Workplaces: The Emergence of Environmental Data Science Degree Programs, and Lessons Learned from Running One

Combler l'écart en formation dans les milieux de travail en environnement : l'émergence de programmes d'études en science des données environnementales et leçons tirées de la direction d'un tel programme

The increasing volume and complexity of environmental data, paired with the urgency and scale of current environmental challenges, has created demand across sectors (academia, industry, government, non-profits, etc.) for environmental scientists with both domain expertise and computational skills to responsibly work with data. Limited quantitative training in Environmental Science (and related) degree programs, however, can create a gap between academic preparation and computational skills needed - and increasingly expected - in many environmental workplaces. Environmental Data Science programs are emerging to fill that gap. In this talk, we highlight the need for, and emergence of, environmental data science training in academia, then share what we have learned during the inaugural year of our new Master of Environmental Data Science program at the University of California, Santa Barbara.

Le volume et la complexité croissants des données environnementales, couplés avec l'urgence et l'ampleur des défis environnementaux actuels, ont créé une demande à travers divers secteurs (universités, industries, gouvernements, organisations à but non lucratif, etc.) pour des scientifiques environnementaux ayant à la fois une expertise du domaine et des compétences informatiques pour travailler de manière responsable avec les données. Une formation quantitative limitée dans les programmes diplômant en sciences environnementales (et autres programmes connexes) peut toutefois créer un écart entre la préparation universitaire et les compétences informatiques requises – et de plus en plus attendues – dans plusieurs milieux de travail en environnement. Des programmes de science des données environnementales voient le jour afin de combler cet écart. Dans cette présentation, nous soulignons le besoin et l'émergence de programmes universitaires en sciences des données environnementales, puis partageons les apprentissages que nous avons acquis pendant l'année inaugurale de notre nouveau programme de maîtrise en sciences des données environnementales à University of California, Santa Barbara.

[16:00-16:30]

Holly N. Steeves (University of Western Ontario) **Sofiia Romanovska** (University of Victoria) **Laura L.E. Cowen** (University of Victoria)

Exploring the Robustness of Citizen Science Golden Eagle Data

Exploration de la robustesse des données sur l'aigle royal issues de la science citoyenne

Citizen science data, that is data provided by citizen volunteers, is becoming increasingly common yet is still often not accepted by the scientific community. Concerns about the robustness of such datasets lead to analyses being rejected or criticized. In order to validate and ver-

Les données issues de la science citoyenne, c'est-à-dire fournies par des citoyens bénévoles, sont de plus en plus courantes, mais ne sont toujours pas souvent acceptées par la communauté scientifique. Les inquiétudes concernant la robustesse de ces ensembles de données conduisent au rejet ou à la critique des analyses. Afin

ify data quality standards, many such studies have introduced expert validation where scientists in the field examine samples such as photographs or soundclips to validate random or unusual observations. However, in studies where no such samples are available, there is a need for validation techniques and data quality assessments. In order to assess robustness of the data to volunteer skill level, a proxy based on recorded experience was calculated. With this and through the use of standard statistical methods such as clustering, chi-square and Kruskall-Wallis tests, and generalized linear models, we explore the validity and rigidity of Golden Eagle data from the Rocky Mountain Eagle Research Foundation.

de valider et de vérifier les normes de qualité des données, de nombreuses études de ce type ont introduit la validation par des experts où des scientifiques du domaine examinent des échantillons tels que photographies ou clips sonores pour valider des observations aléatoires ou inhabituelles. Cependant, dans les études où de tels échantillons ne sont pas disponibles, il est nécessaire de recourir à des techniques de validation et d'évaluation de la qualité des données. Afin d'évaluer la robustesse des données par rapport au niveau de compétence des volontaires, un proxy basé sur l'expérience enregistrée a été calculé. Grâce à celui-ci et à des méthodes statistiques standard telles que la classification, les tests de chi carré et de Kruskall-Wallis et les modèles linéaires généralisés, nous explorons la validité et la rigidité des données sur l'aigle royal provenant de la Rocky Mountain Eagle Research Foundation.

[16:30-17:00]

Susan Simmons (North Carolina State University)

Best Practices for Virtual Research Groups

Les meilleures pratiques pour les groupes de recherche virtuels

With the advent of better virtual platforms, we are seeing an increase in international research and working groups. However, there are numerous challenges and issues that must be overcome to have successful outcomes. In this talk, we focus on good strategies in mastering some of the difficulties in managing these groups. References will be drawn to three international working groups through The International Environmetrics Society from 2021.

L'avènement de plateformes virtuelles supérieures a mené à une hausse de recherches internationales et de groupes de travail. Toutefois, il y a de nombreux défis et problèmes à surmonter afin d'aboutir à de bons résultats. Lors de cet exposé, nous présenterons de bonnes stratégies afin de gérer certaines difficultés de la gestion de groupes. Des références seront tirées de trois groupes de travail internationaux de The International Environmetrics Society datant de 2021.

Chair/Président: Mélina Mailhot

Organizer/Responsable: Mélina Mailhot

Date: Wednesday June 1 / mercredi 1 juin

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-16:00]

Silvana Manuela Pesenti (University of Toronto) **Mélina Mailhot** (Concordia University) **Emily Wright** (Concordia University)

Renyi Divergence for Extreme Value Distributions

Divergence de Rényi pour les distributions des valeurs extrêmes

We propose a sensitivity testing framework suitable for insurance losses that follow heavy tailed distributions, including the family of generalised Pareto distributions. Starting from a baseline probability measure, e.g., a parametric model estimated from data, we solve the problem of finding the perturbed probability measure that exceeds a given risk tolerance and has smallest Rényi divergence to the baseline measure. We study the optimisation problem with risk tolerances given by the Value-at-Risk, Expected Shortfall, and expectiles, and prove that the perturbed probability measures exist and are uniqueness. We provide semi-analytical solutions and characterise the perturbed measures via the so-called lambda-exponential functions. Our findings are illustrated on a Canadian insurance loss dataset stemming from natural catastrophes.

Nous proposons un cadre de test de sensibilité adapté aux pertes d'assurance qui suivent des distributions à queue lourde, y compris la famille des distributions de Pareto généralisées. En prenant comme point de départ une mesure de probabilité de base, comme un modèle paramétrique estimé à partir de données, nous résolvons le problème consistant à trouver la mesure de probabilité altérée qui excède une tolérance au risque définie et qui présente la plus petite divergence de Rényi par rapport à la mesure de base. Nous analysons le problème d'optimisation avec des tolérances au risque définies par la valeur à risque, le déficit attendu et les expectiles, puis nous démontrons que les mesures de probabilité altérées existent et sont uniques. Nous fournissons des solutions semi-analytiques et caractérisons les mesures altérées à l'aide des fonctions dites « lambda-exponentielles ». Nous illustrons nos résultats par un ensemble de données canadiennes de sinistres provenant de catastrophes naturelles.

[16:00-16:30]

Tobias Fissler (Vienna University of Economics and Business) **Michael Merz** (University of Hamburg) **Mario V. Wüthrich** (ETH Zurich)

Deep Quantile and Deep Composite Model Regression

Régression quantile profonde et modèle de régression composite profond

A main difficulty in actuarial claim size modeling is that there is no simple off-the-shelf distribution that simultaneously provides a good distributional model for the main body and the tail of the data. In particular, covariates may have different effects for small and for large claim sizes. To cope with this problem, we introduce a deep composite regression model whose splicing point is given in terms of a quantile of the conditional claim size distribution rather than a constant. To facilitate M-estimation for such models, we introduce and charac-

L'une des principales difficultés de la modélisation actuarielle de la taille du sinistre est qu'il n'existe pas de distribution simple standard qui fournisse simultanément un bon modèle de distribution pour le corps principal et la queue des données. Plus particulièrement, les covariables peuvent avoir des effets différents selon que la taille du sinistre est petite ou grande. Pour résoudre ce problème, nous utilisons un modèle de régression composite profond dont le point de jonction est défini en fonction d'un quantile de la distribution conditionnelle de la taille du sinistre plutôt que d'une constante. Afin de faciliter l'estimation M de

terize the class of strictly consistent scoring functions for the triplet consisting a quantile, as well as the lower and upper expected shortfall beyond that quantile. In a second step, this elicitability result is applied to fit deep neural network regression models. We demonstrate the applicability of our approach and its superiority over classical approaches on a real accident insurance data set.

ces modèles, nous créons et caractérisons la classe des fonctions de pointage strictement cohérentes pour le triplet constitué d'un quantile, ainsi que du déficit inférieur et supérieur estimé au-delà de ce quantile. Dans un deuxième temps, nous appliquons ce résultat d'élicitabilité pour ajuster des modèles de régression à réseaux de neurones profonds. Nous démontrons l'applicabilité de notre approche et sa supériorité par rapport aux approches classiques sur un ensemble de données réelles d'assurance contre les accidents corporels.

[16:30-17:00]

Klaus Herrmann (Université de Sherbrooke) **Marius Hofert** (University of Waterloo) **Johanna G. Nešlehová** (McGill University)

Copula Diagonals, Distortions and the Asymptotic Distribution of Maxima

Diagonales des copules, distorsions et distribution asymptotique des maxima

In its most common form, extreme value theory is concerned with the limiting distribution of location-scale transformed block-maxima of a sequence of identically distributed random variables. In case the members of the sequence are independent, the weak limiting behavior of the maximum is adequately described by the classical Fisher–Tippett–Gnedenko theorem. In this presentation we are interested in the case of dependent random variables, while retaining a common marginal distribution function for all members of the sequence. This approach is facilitated by highlighting the connection between block-maxima and copula diagonals in an asymptotic context. The main goal of this presentation is to discuss a generalization of the Fisher–Tippett–Gnedenko theorem in this setting, leading to limiting distributions that are not in the class of generalized extreme value distributions.

Dans sa forme plus connue, la théorie des valeurs extrêmes traite de la loi limite du maximum standardisé d'une série de variables aléatoires identiquement distribuées. Si les variables sont indépendantes, la loi limite du maximum est décrite par le théorème classique de Fisher—Tippett—Gnedenko. Dans cet exposé, je m'intéresse au cas d'une série de variables qui sont dépendantes, mais je conserve l'hypothèse selon laquelle les variables sont identiquement distribuées. Dans cette approche j'accentue la connexion entre les maximums et la structure de dépendance dans un cadre asymptotique. Un point capital de mon exposé est la généralisation du théorème classique de Fisher–Tippett–Gnedenko dans ce contexte et, surprenamment, en général les lois asymptotiques ne font pas partie de la classe des lois d'extremums généralisées.

Chair/Président: Yanbo Tang

Date: Wednesday June 1 / mercredi 1 juin

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-15:45]

Di Meng (Wilfrid Laurier University) **Mark Reesor** (Wilfrid Laurier University) **Adam Metzler** (Wilfrid Laurier University)

Calibration and Pricing of Contingent Convertible Securities

Calibrage et fixation du prix des titres convertibles conditionnels

This paper develops a structural model for pricing contingent convertible securities (CoCo) and calibrates the CoCo conversion trigger level. Two asset value processes are compared: (1) geometric Brownian motion, and (2) jump-diffusion, where the firm's total payout is linked to its cost of capital, leading to a total payout proportional to asset value. The proposed model can be used to gauge the relative merits of different variations of CoCo. An empirical example using data from the Canadian banking sector is provided to illustrate the insights of the market participants' anticipation on regulation.

Cet article développe un modèle structurel pour l'évaluation des titres convertibles conditionnels (CoCo) et calibre le niveau de déclenchement de la conversion des CoCo. Nous comparons deux processus de valeur de l'actif : (1) le mouvement brownien géométrique, et (2) le saut-diffusion, où le paiement total de l'entreprise est lié à son coût du capital, ce qui conduit à un paiement total proportionnel à la valeur de l'actif. Le modèle proposé peut être utilisé pour évaluer les mérites relatifs de différentes variantes de CoCo. Nous fournissons un exemple empirique utilisant des données du secteur bancaire canadien pour illustrer les perspectives d'anticipation des participants du marché sur la réglementation.

[15:45-16:00]

Félix Locas (Université du Québec à Montréal)

De Finetti's Control Problem with Absolutely Continuous Strategies in a Diffusion Model

Problème de contrôle stochastique de De Finetti avec stratégies absolument continues dans un modèle de diffusion

De Finetti's optimal control problem has been thoroughly studied for dividend payouts, for natural resources extraction and for population dynamics/harvesting. We present a new version of de Finetti's control problem in a diffusion model, in which the admissible strategies have control rates taking values in a path-dependent interval. Under some assumptions on the model parameters, we find an optimal strategy of bang-bang type, consisting of using the maximal path-dependent rate as long as the controlled process is above a barrier.

Les problèmes de contrôle stochastique de De Finetti ont été étudiés en lien avec des modèles de paiement de dividendes, d'extraction de ressources naturelles, ou pour des dynamiques de populations/récoltes. Nous présentons une nouvelle version du problème de contrôle stochastique dans un modèle de diffusion dans lequel les stratégies admissibles ont des taux de contrôle compris dans un intervalle dépendant de la valeur actuelle du contrôle. Sous certaines hypothèses, nous trouvons une stratégie optimale de type bang-bang, consistant à utiliser le taux de contrôle maximal tant et aussi longtemps que le processus se retrouve au-dessus d'une barrière.

[16:00-16:15]

Yifan Li (University of Western Ontario) **Reg Kulperger** (University of Western Ontario) **Hao Yu** (University of Western Ontario)

Semi-G-Structure: A Flexible Framework to Deal with Model Uncertainty

Semi-structure G : un cadre souple pour traiter l'incertitude du modèle

The G-expectation framework is a generalization of classical probability system based on subadditive proba-

Le cadre de G-espérance est une généralisation du système de probabilité classique basé sur des probabilités sous-additives et conçu

bilities, designed to handle situations with model uncertainty that cannot be described by a single probabilistic model. However, the distributions and independence in the G-expectation world are quite different from the classical setup. It requires some care in general practice and interpretation. Hence, a fundamental and unavoidable problem is how to better understand the G-version distributions and independence from a statistical perspective. To explore this problem, we develop the so-called semi-G-structure. The semi-G-structure plays a hybrid role connecting the classical and G-framework, so that it allows more computational and statistical advantages. We use several examples to show its flexibility and how it can be used in different areas in statistics, interval data and finance.

pour traiter des cas avec incertitude du modèle qui ne peuvent pas être décrits par un seul modèle probabiliste. Les distributions et l'indépendance dans l'univers de la G-espérance sont cependant très différentes de celles du système classique. La pratique et l'interprétation générales requièrent une certaine attention. Ainsi, un problème fondamental inévitable est de savoir comment mieux comprendre les distributions et l'indépendance en version G dans une perspective statistique. Pour traiter ce problème, nous développons ce que nous appelons une semi-structure G. En connectant les cadres classique et G elle tient un rôle hybride qui offre plus d'avantages computationnels et statistiques. Plusieurs exemples servent à illustrer sa souplesse et ses utilisations possibles à diverses fins en statistique, données d'intervalles et finance.

[16:15-16:30]

Yunhong Lyu (University of Windsor) **Sévérien Nkurunziza** (University of Windsor)

Estimation and Testing in Generalized Cox–Ingersoll–Ross Processes

Estimation et test dans les processus de Cox–Ingersoll–Ross Généralisés

In this talk, we propose a generalized Cox–Ingersoll–Ross (GCIR) process which is suitable for modeling some periodic financial data. We also consider an inference problem, about the drift parameters of the introduced GCIR, in the case where the target parameters may satisfy some restrictions. We derive the Unrestricted maximum likelihood estimator (UMLE) and the restricted maximum likelihood estimator (RMLE) and establish their asymptotic properties. We also construct a test for testing the restriction and propose shrinkage estimators (SEs). Further, we derive the asymptotic power and the asymptotic distributional risk of the proposed estimators as well as their relative efficiency. Finally, we present simulation results which corroborate our theoretical findings, and we apply the proposed method to the 10-year U.S. treasury bond yield. The additional novelty consists in the fact that we overcome the difficulty due to the fact that the GCIR does not have an explicit solution.

Dans cet exposé, nous proposons un processus de Cox-Ingersoll-Ross généralisé (CIRG) qui convient à certaines données financières périodiques. Nous considérons aussi un problème d'inférence concernant le paramètre de dérive du CIRG, dans le cas où celui-ci peut satisfaire à une restriction. Nous établissons les estimateurs du maximum de vraisemblance sans restriction (UMLE) et avec restriction (RMLE) ainsi que leurs propriétés asymptotiques. De plus, nous construisons un test de restriction et des estimateurs à rétrécissement (SEs). En outre, nous élaborons la puissance asymptotique et le risque distributionnel asymptotique des estimateurs proposés ainsi que leur efficacité relatives. Enfin, nous présentons les résultats de simulations qui corroborent nos résultats théoriques, et appliquons la méthode proposée aux obligations du Trésor américain de 10 ans. La nouveauté de plus réside dans le fait que nous surmontons la difficulté due au fait que le CIRG n'a pas de solution explicite.

[16:30-16:45]

Elham Soufiani (University of Regina)

Generalization of Hoeffding's inequality for Extended Acceptable Random Variables

Généralisation de l'inégalité d'Hoeffding pour les variables aléatoires acceptables étendues

The large deviation inequalities are significant results in probability and statistics applications. The formalization of the theory of large deviations started with Insurance Mathematics, namely 'Ruin Theory'. This theory

Les grandes inégalités de déviation sont d'importants résultats appliqués en statistiques et probabilités. La formalisation de la théorie des grandes déviations a commencé avec les mathématiques d'assurance, aussi connue comme la ≪théorie de la ruine≫. Cette

is mostly concerned about the exponential decline of the probability of tail events, like the Ruin probability in Insurance businesses. For independent random variables with uniform bounds, several exponential inequalities are available in the literature, such as Bernstein, Bennett, Hoeffding, and so on. However, the treatment of dependency in the sequence of random variables has attracted probabilists and statisticians in recent years. In this work, we are going to concentrate on one of the most important large deviation inequalities, known as Hoeffding's, which has many applications in Ruin Theory and Insurance. We will generalize this inequality for Extended Acceptable random variables which contain a large group of dependency structures.

théorie se concentre principalement sur le déclin exponentiel de la probabilité des événements extrêmes, comme la probabilité de ruine dans l'industrie de l'assurance. Pour les variables aléatoires indépendantes à bornes uniformes, un certain nombre d'inégalités exponentielles sont offertes, comme celle de Bernstein, Bennett, Hoeffding, etc. Cependant, le traitement de la dépendance dans la suite des variables aléatoires a attiré l'attention des probabilistes et des statisticiens ces dernières années. Dans cet article, nous nous concentrons sur l'une des plus importantes inégalités de grande déviation : celle de Hoeffding, couramment employée dans la théorie de la ruine et en assurance. Nous généraliserons cette inégalité pour des variables aléatoires acceptables étendues contenant un grand groupe de structures de dépendances.

[16:45-17:00]

Sharandeep Singh Pandher (University of Regina) **Shakhawat Hossain** (University of Winnipeg) **Andrei Volodin** (University of Regina)

Generalized Autoregressive Moving Average (GARMA) Models: An Efficient Estimation Approach

Modèles de moyenne mobile autorégressive généralisée (GARMA) : une approche d'estimation efficace

In this talk, we consider the efficient estimation approach, so-called pretest and shrinkage approach in estimating the regression parameters of the generalized autoregressive moving average (GARMA) model, which are pervasive for modeling binary and count time series data. This model accommodates a set of covariates in addition to the ARMA parameters. We want to estimate the regression and ARMA parameters when some of the regression parameters may be belong to a subspace. We apply the maximum partial likelihood method to obtain the unrestricted maximum partial likelihood estimator(UMPLE) and also the restricted maximum partial likelihood estimator (RPMLE) and then present the improved pretest and shrinkage estimators. We establish the asymptotic distributional biases and risks of the proposed estimators and evaluate their relative performance with respect to the UMPLE. The methodology is investigated using simulation studies and then demonstrated on using a real data example.

Dans cet exposé, nous considérons l'approche d'estimation efficace, dite approche de prétest et de retrécissement pour estimer les paramètres de régression du modèle de moyenne mobile autorégressive généralisée (GARMA), qui sont omniprésents pour la modélisation de données de séries temporelles binaires et de comptage. Ce modèle intègre un ensemble de covariables en plus des paramètres ARMA. Nous voulons estimer les paramètres de la régression et ceux du ARMA lorsque certains des paramètres de régression appartiennent à un sous-espace. Nous appliquons la méthode du maximum de vraisemblance partielle pour obtenir l'estimateur du maximum de vraisemblance partielle sans restriction (UMPLE) ainsi que l'estimateur du maximum de vraisemblance partielle restreint (RPMLE), puis présentons les estimateurs améliorés de prétest et de rétrécissement. Nous établissons les biais et les risques des distributions asymptotiques des estimateurs proposés et nous évaluons leur performance relative par rapport à l' UMPLE. La méthodologie est étudiée à l'aide d'études de simulation, et ensuite démontrée à l'aide d'un exemple de données réelles.

Statistical Methods for Handling Ordinal Data, Missing data, and Data with Measurement Error
Méthodes statistiques pour le traitement des données ordinales, des données manquantes et des données comportant des erreurs de mesure

Chair/Président: Joan X. Hu

Date: Wednesday June 1 / mercredi 1 juin

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-15:45]

Lyubov Doroshenko (Université Laval) **Brunero Liseo** (La Sapienza University of Rome)

Generalized Linear Mixed Model with Bayesian Rank Likelihood

Modèle linéaire généralisé mixte à l'aide de la probabilité de rang bayésienne

We consider situations where a model for an ordered categorical response variable is deemed. The interest of the analysis lies in the marginal probability effects. Therefore, the questions to be addressed are focused not on the scale of each variable, but rather on the association between variables themselves. Standard ordered response models may not be very suited to perform this analysis, being these effects to a large extent predetermined by the rigid parametric structure of the model. We propose to use a rank likelihood approach in a non-Gaussian framework and show how additional flexibility can be gained by modeling individual heterogeneity in terms of latent structure. In particular, the rank likelihood approach is extended to Generalized Linear Mixed Effects models' framework. A Bayesian approach using Markov Chain Monte Carlo is adopted. The performance of the model is illustrated in the context of Sovereign Credit Ratings modeling and forecasting.

Nous examinons les situations dans lesquelles on estime qu'il existe un modèle pour une variable de réponse catégorielle ordonnée. L'intérêt de notre analyse réside dans les effets de probabilité marginale. Par conséquent, les questions à traiter ne portent pas sur l'échelle de chaque variable, mais plutôt sur l'association entre les variables elles-mêmes. Les modèles standards de réponse ordonnée ne sont peut-être pas très adaptés pour effectuer cette analyse, ces effets étant dans une large mesure prédéterminés par la structure paramétrique rigide du modèle. Nous proposons une approche de probabilité de rang dans un cadre non gaussien et montrons la manière dont on obtient une plus grande souplesse par la modélisation de l'hétérogénéité individuelle en termes de structure latente. Plus particulièrement, nous étendons l'approche de la probabilité de rang au cadre des modèles linéaires généralisés mixtes. Nous adoptons une approche bayésienne au moyen de la méthode de Monte-Carlo par chaînes de Markov. Nous illustrons l'efficacité du modèle dans le contexte de la modélisation et de la prévision des cotes de crédit souveraines.

[15:45-16:00]

Aya A. Mitani (University of Toronto) **Osvaldo Espin-Garcia** (University Health Network, University of Toronto) **Victoria Landsman** (Institute of Work and Health, University of Toronto)

Using Stereotype Regression for Unbiased Inference from Ordinal Outcome-Dependent Samples

Emploi de régression de stéréotype pour une inférence non biaisée à partir d'échantillons dépendant des résultats ordinaux

Outcome-dependent sampling (ODS) is commonly used when the disease prevalence is low or when study resources are limited. The case-control study is an example of ODS, and importantly, the odds ratio estimated via logistic regression from the case-control sample is a consistent estimate of the population. When the outcome has more than two categories and the sampling is dependent on the ordinal outcome, the stereotype regression (SR) model produces consistent estimates. However, the more common regression models for ordinal outcomes, such as the proportional odds (PO) and

L'échantillonnage dépendant des résultats (ODS) est couramment adopté lorsque la prévalence d'une maladie est basse ou lorsque les ressources de l'étude sont limitées. L'étude cas-témoins est un exemple d'ODS, et tout particulièrement, le rapport des cotes estimé par l'entremise d'une régression logistique tirée d'un échantillon de cas-témoins est une estimation convergente de la population. Lorsqu'un résultat possède plus de deux catégories et que l'échantillonnage est dépendant des résultats ordinaux, le modèle de régression de stéréotype (RS) produit des estimations convergentes. Cependant, les modèles courants de régression pour des résultats ordinaux, comme la probabilité proportionnelle

continuation ratio (CR), produce biased estimates when the underlying disease prevalence is unequal across the categories. Through simulation, we quantify the inconsistencies of various ordinal regression estimates from ODS under a wide range of disease prevalence and highlight the robustness of the SR model even under the most extreme scenarios. Finally, we apply each model to data from a knee osteoarthritis study.

(PP) et le rapport de continuation (RC), génèrent des estimations biaisées lorsque la prévalence de maladie sous-jacente est inégale dans l'ensemble des catégories. À partir de simulation, nous quantifions les incohérences de plusieurs estimations de régression ordinales à partir d'ODS selon une grande variété de prévalences de maladie et soulignons la robustesse du modèle de RS même dans les scénarios très extrêmes. Enfin, nous appliquons chaque modèle à des données tirées d'une étude sur l'ostéo-arthrite du genou.

[16:00-16:15]

Gurbakhshash Singh (Central Connecticut State University) **Gordon H. Fick** (University of Calgary)

Ordinal Outcomes: Considerations for the Generalized Linear Model with the Identity Link

Résultats ordinaux : Considérations au sujet du modèle linéaire généralisé avec lien d'identité

There are many options available for the analysis of ordinal outcomes. The Proportional Odds Model, based on the logit link, is, perhaps, still the most often seen. Recently, a comparable model has been suggested based on the log link. We now extend to our work to the identity link which introduces further constraints on the parameter space. We present results 1) on conditions for the uniqueness of the Maximum Likelihood Estimate (MLE) based on the rank of certain matrices, 2) about using constrOptim in R to determine the MLE, 3) and some other fundamental results. We also offer some closed form expressions for the MLE and we offer some discussion about hypothesis tests in constrained spaces.

De nombreuses options s'offrent pour l'analyse de résultats ordinaux. Le modèle de cotes proportionnelles basé sur la fonction de lien logit est sans doute encore celui que l'on voit le plus souvent. Un modèle comparable basé sur la fonction de lien log a récemment été suggéré. Notre travail s'étend maintenant au lien d'identité qui introduit d'autres contraintes à l'espace paramétrique. Nous présentons 1) des résultats sur les conditions de l'unicité de l'estimation du maximum de vraisemblance (EMV) basée sur le rang de certaines matrices, 2) des résultats sur l'utilisation de constrOptim dans R pour déterminer l'EMV, 3) et certains autre résultats fondamentaux. Nous proposons aussi certaines expressions analytique pour l'EMV et discutons des tests d'hypothèses dans des espaces contraints.

[16:15-16:30]

Hon-Yiu So (University of Waterloo) **Parminder Raina** (McMaster University) **Jinhui Ma** (McMaster University)

Application of Machine Learning in Imputing Heterogeneous Co-missing Data

Application de l'apprentissage automatique dans l'imputation de données co-manquantes hétérogènes

In epidemiological studies, paired data, such as systolic and diastolic blood pressure, forced expiratory volume in one second (FEV1) and forced vital capacity (FVC) from the spirometry test, may be collected. Paired data are usually missing simultaneously (co-missingness) and their relationship varies across various latent groups in the population (heterogeneity). Ordinary imputation methods may not be appropriate to handle this type of missing data, let alone they are problematic in adapting to interactions, nonlinearity, and mixed variable types. This work reviewed the literature and identified distinct statistical and machine learning approaches for handling the missing data. We compared the performance of ordinary statistical imputation strategies with the machine learning approaches through a simulation study in terms of accuracy and computational efficiency.

En études épidémiologiques, on peut recueillir des données en paires comme la pression sanguine diastolique et systolique, le volume expiratoire maximal en 1 seconde (FEV1) et la capacité vitale forcée (CVF) tirée d'un examen spirométrique. Les données en paires sont généralement manquantes simultanément (co-manquantes) et leur lien varie dans différents groupes latents dans la population (hétérogénéité). Conséquemment, les méthodes d'imputation ordinaires pourraient ne pas suffire pour traiter ce genre de données manquantes, sans compter que ces dernières posent un problème pour l'adaptation aux interactions, à la nonlinéarité et aux types de variables mixtes. Ce travail a passé en revue la documentation et a repéré différentes approches statistiques et d'apprentissage automatique servant à gérer les données manquantes. Au moyen d'une étude en simulation, nous avons comparé la performance relative à la précision et au rendement en calcul entre des stratégies d'imputation statistique ordinaire par et les approches d'apprentissage automatique.

Statistical Methods for Handling Ordinal Data, Missing data, and Data with Measurement Error
Méthodes statistiques pour le traitement des données ordinales, des données manquantes et des données comportant des erreurs de mesure

[16:30-16:45]

Yifan Sun (University of Western Ontario)

Estimation and Variable Selection for Function-on-scalar Linear Model with Covariate Measurement Error

Estimation et sélection de variables pour modèle linéaire à fonctions scalaires avec erreur de mesure de covariables

This paper concerns estimation and variable selection problems for function-on-scalar linear regression model with error-prone covariates. We propose a debiased loss function combined with the group smoothly clipped absolute deviation penalty to simultaneously estimate functional coefficients and select relevant predictors. An efficient computing algorithm and a data-driven tuning parameter selection method are developed. The estimation and selection consistency are established under regularity conditions. We investigate the finite sample performance of the proposed method through simulation studies and a real data application.

Cet article s'intéresse aux problèmes d'estimation et de sélection de variables pour un modèle de régression linéaire à fonctions-scalaires avec covariables sujettes à erreur. Nous proposons une fonction de perte non biaisée combinée à la pénalité de groupe à écart absolu avec coupure lisse afin d'estimer simultanément les coefficients fonctionnels et de choisir les prédicteurs pertinents. Nous développons aussi un algorithme de calcul efficace et une méthode de sélection de paramètres de réglage orientée par les données. La sélection et la convergence des estimateurs est établie selon des conditions de régularité. Nous examinons la performance à échantillon fini de la méthode proposée grâce à des études en simulation et une application à partir de données réelles.

[16:45-17:00]

Max Turgeon (University of Manitoba)

Generalized Soft Impute for Matrix Completion

Imputation douce généralisée pour la complétion matricielle

Missing data is a common challenge in data science. As the number of measurements increases, so does the likelihood that at least one of them is missing for a given observation, leading to inefficient complete-case analyses. Matrix completion algorithms have gained popularity recently for their simplicity and computational efficiency. In this talk, we present a matrix completion algorithm based on generalized Singular Value Decomposition (SVD), which unlike classical SVD imposes constraints on the rows and columns of the data matrix. This framework is particularly suitable for multivariate methods like Weighted Principal Component Analysis and Correspondence Analysis. We obtain good performance by regularizing the nuclear norm of the completed matrix, and we achieve computational efficiency by using proximal gradient descent. Finally, we discuss applications of our algorithm to the field of statistical genetics.

Les données manquantes sont un défi commun en science des données. La probabilité qu'une observation manque une mesure augmente conjointement au nombre de variables, ce qui mène à des analyses de cas complètes inefficaces. Les algorithmes de complétion matricielle représentent une approche populaire en raison de leur simplicité et de leur efficacité computationnelle. Ainsi, nous présentons un algorithme de complétion matricielle basé sur la Décomposition en Valeurs Singulières (SVD) généralisée, ce qui permet d'imposer des contraintes sur les lignes et les colonnes de la matrice de données. Ce cadre est idéal pour des méthodes multivariées telles que l'Analyse en Composantes Principales pondérée et l'Analyse Factorielle des Correspondances. Nous obtenons une bonne performance en régularisant la norme nucléaire de la matrice complétée, ainsi qu'une bonne efficacité computationnelle en utilisant un algorithme gradient proximal. Enfin, nous discutons les applications de notre algorithme pour les statistiques génétiques.

Dynamic Treatment Regime Analysis and Dynamic Modelling
Analyse du régime de traitement dynamique et modélisation dynamique

Chair/Président: Andrea Benedetti

Date: Wednesday June 1 / mercredi 1 juin

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-15:45]

Marzieh Mussavi Rizi (University of Waterloo) **Joel A. Dubin** (University of Waterloo) **Michael Wallace** (University of Waterloo)

Dynamic Treatment Regimes in Dyadic Networks

Régimes de traitement dynamiques dans les réseaux dyadiques

Precision medicine is a paradigm which prioritizes tailoring patient treatments according to their relevant traits. In this framework, a Dynamic Treatment Regime (DTR) is a sequence of decision rules that generate treatment recommendations based on patient-level data. A goal of DTR is estimating the optimal DTR: that which maximizes the population's expected outcome. Typically, this relies on a no-interference assumption: individuals' treatments do not impact others' outcomes. However, this assumption is questionable in practice, such as in the treatment of infectious diseases. We explore interference as dyadic (pairwise) networks in DTR estimation and show that ignoring this interference can lead to non-optimal DTRs. We extend the estimation method of dynamic weighted ordinary least squares, an easily implemented method which is robust to model misspecification, to account for interference of this form. We demonstrate the performance of our proposed method through simulation.

La médecine de précision est un paradigme qui privilégie l'adaptation des traitements des patients en fonction de leurs caractéristiques pertinentes. Dans ce cadre, un régime de traitement dynamique (RTD) est une séquence de règles de décision qui génère des recommandations de traitement basées sur les données du patient. Un objectif du RTD est d'estimer le RTD optimal : celui qui maximise le résultat attendu de la population. Généralement, cela repose sur une hypothèse de non-interférence : les traitements des individus n'ont pas d'impact sur les résultats des autres. Cependant, cette hypothèse est discutable en pratique, notamment dans le traitement des maladies infectieuses. Nous explorons l'interférence en tant que réseaux dyadiques (par paires) dans l'estimation des RTD et montrons que l'ignorance de cette interférence peut conduire à des RTD non optimaux. Nous étendons la méthode d'estimation des moindres carrés ordinaires dynamiques pondérés, une méthode facile à mettre en œuvre et robuste à la mauvaise spécification du modèle, pour tenir compte de l'interférence de cette forme. Nous démontrons la performance de notre méthode proposée par le biais d'une simulation.

[15:45-16:00]

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

Doubly-Robust Dynamic Treatment Regimen Estimation for Binary Outcomes

Estimation dynamique du régime de traitement à double robustesse pour les résultats binaires

In precision medicine, Dynamic Treatment Regimes (DTRs) are sets of treatment rules that take an individual patient's information as input and output actions to be taken. Our aim is to identify the DTR that optimizes expected patient outcomes. Many methods proposed for optimal DTR use estimation with continuous outcomes, but ones with binary outcomes have received little attention. We propose a new method for DTR estimation, dynamic weighted generalized linear models,

Dans le cadre de la médecine de précision, les régimes de traitement dynamiques (RDT) sont des ensembles de règles de traitement qui prennent en compte les informations d'un patient individuel et produisent des actions à entreprendre. Notre objectif est d'identifier le RDT qui optimise les résultats espérés à propos du patient. De nombreuses méthodes proposées pour un RDT optimal utilisent une estimation avec des résultats continus, mais celles avec des résultats binaires ont reçu peu d'attention. Nous proposons une nouvelle méthode d'estimation du

which accommodates binary outcomes while offering relatively straightforward implementation and robustness to model misspecification. We will introduce the method and its underlying theory, and illustrate both in an analysis of e-cigarette usage and smoking cessation, using observed data from the Population Assessment of Tobacco and Health study.

RDT, les modèles linéaires généralisés dynamiques pondérés, qui prennent en compte les résultats binaires tout en offrant une mise en œuvre relativement simple et une robustesse quant à la mauvaise spécification du modèle. Nous présenterons la méthode et la théorie qui la sous-tend, et nous les illustrerons par une analyse de l'utilisation des e-cigarettes et du sevrage tabagique, à l'aide des données observées de l'étude Population Assessment of Tobacco and Health.

[16:00-16:15]

Dan Liu (Western University) **Wenqing He** (Western University)

Q-learning with Misclassified Response in Binary Regression

Apprentissage Q avec réponses mal classées dans une régression binaire

The study of precision medicine involves dynamic treatment regimes (DTRs). DTRs recommend sequences of treatment decision rules based on the history of patient-level information and previous treatments. Many statistical methods have been developed in recent years to estimate DTRs including Q-learning, a regression-based method in DTRs. However, the existing methods may break down in the presence of noisy data, such as misclassified response. In this talk, we consider Q-learning with misclassified response in binary regression. We examine the effect of misclassification in the response on the estimation of optimal treatment decision rules in Q-learning and propose effective error correction methods to accommodate the misclassification effect. Numerical studies are conducted to evaluate the performance of proposed methods.

L'étude de la médecine de précision fait appel à des régimes de traitement dynamiques. Ces régimes recommandent des séquences de règles de décision de traitement reposant sur l'historique des informations relatives au patient et des traitements précédents. Ces dernières années, de nombreuses méthodes statistiques ont été créées pour estimer les régimes de traitement dynamiques, notamment l'apprentissage Q, une méthode basée sur la régression des régimes de traitement dynamiques. Cependant, les méthodes actuelles risquent de mal performer en présence de données bruitées, telles que des réponses mal classées. Dans cette présentation, nous nous penchons sur l'apprentissage Q avec des réponses mal classées dans une régression binaire. Nous examinons l'effet d'une mauvaise classification de la réponse sur l'estimation des règles optimales de décision de traitement dans l'apprentissage Q et nous proposons des méthodes efficaces de correction des erreurs pour tenir compte de l'effet de la mauvaise classification. Nous réalisons des études numériques pour évaluer les résultats de ces méthodes.

[16:15-16:30]

Nathaniel David Osgood (University of Saskatchewan) **Jeremy Eng** (Saskatchewan Polytechnic)

Effective Use of PMCMC for Daily Epidemiological Monitoring and Reporting: Methodological Lessons

Utilisation efficace du PMCMC pour la surveillance et le rapport épidémiologiques quotidiens : leçons méthodologiques

Throughout the pandemic, CEPHIL has used Particle Filtering to support effective cross-Canada COVID-19 decision making by regularly reporting estimates of the current epidemiological and health care utilization situation based on model-based analysis integrating incoming health system and wastewater data, and by performing short-term projections of epidemiology and healthcare capacity utilization so as to support triggering of public health measures and surge capacity mobilization. Particle Markov Chain Monte Carlo-leveraged dynamic models offer greater power and versatility in supporting reporting, but PMCMC requires careful configura-

Tout au long de la pandémie, le CEPHIL a utilisé le filtrage de particules pour soutenir la prise de décisions efficaces concernant la COVID-19 à l'échelle nationale, en rapportant régulièrement des estimations de la situation actuelle en matière d'épidémiologie et d'utilisation des soins de santé basées sur une analyse modélisée intégrant les données entrantes du système de santé et des eaux usées, et en effectuant des projections à court terme de l'épidémiologie et de l'utilisation des capacités de soins de santé, afin de soutenir le déclenchement des mesures de santé publique et la mobilisation des capacités de pointe. Les modèles dynamiques à levier de chaîne de Markov de Monte Carlo à particules (PMCMC) offrent une plus grande puissance et une plus grande polyvalence

tion to be practical. We discuss lessons learned from applying PMCMC and its integration into our infrastructure for automated daily reporting. We particularly characterize the acceptance and effective sampling rate impacts of particle count, MCMC step size, timepoint count, form and dispersion of the likelihood function, as well as GPU acceleration.

dans le soutien des rapports, mais le PMCMC nécessite une configuration minutieuse pour être pratique. Nous discutons des leçons tirées de l'application du PMCMC et de son intégration dans notre infrastructure pour la production automatisée de rapports quotidiens. Nous caractérisons particulièrement l'impact sur l'acceptation et le taux d'échantillonnage effectif du nombre de particules, de la taille de l'étape MCMC, du nombre de points de temps, de la forme et de la dispersion de la fonction de vraisemblance, ainsi que de l'accélération GPU.

2022 CRM-SSC Prize in Statistics Invited Address
Allocution du récipiendaire du prix CRM-SSC en statistique 2022

Chair/Président: David Haziza

Date: Thursday June 2 / jeudi 2 juin

Time/Heure: 11:00-12:30

Abstract/Résumé

[11:00-12:00]

Pengfei Li (University of Waterloo)

Density ratio model and its new applications

Modèle du rapport de densité et nouvelles applications

The semiparametric density ratio model (DRM) is a flexible platform for combining information from multiple sources, and it permits elegant inference solutions through the empirical likelihood (EL). This talk discusses our recent contributions to the DRM. In the first part, we consider statistical inference under two-sample DRMs with additional parameters defined through and/or additional auxiliary information expressed as estimating equations. We examine the asymptotic properties of the maximum empirical likelihood estimators (MELEs) of the unknown parameters in the DRMs and/or defined through estimating equations and establish the chi-square limiting distributions for the empirical likelihood ratio (ELR) statistics. We also propose an ELR test for the validity and usefulness of the auxiliary information. In the second part, we discussed the applications of the DRM in unordered homologous chromosome pairs problem, quantile inference for clustered data, and inference of the Youden index and the optimal cut-off point.

Le modèle du rapport de densité (MRD) semi-paramétrique est une plateforme flexible qui permet de combiner des informations provenant de sources multiples et de produire des solutions d'inférence élégantes grâce à la vraisemblance empirique (VE). Cet exposé discute de nos récentes contributions au MRD. Dans la première partie, nous considérons l'inférence statistique sous les MRD à deux échantillons avec des paramètres supplémentaires définis par et/ou des informations auxiliaires supplémentaires exprimées sous forme d'équations d'estimation. Nous examinons les propriétés asymptotiques des estimateurs du maximum de vraisemblance empirique (EMVE) des paramètres inconnus dans les DRM et/ou définis par des équations d'estimation, puis établissons les distributions chi carré limites pour les statistiques du rapport de vraisemblance empirique (RVE). Nous proposons également un test RVE pour la validité et l'utilité des informations auxiliaires. Dans la deuxième partie, nous discutons des applications du MRD dans le problème des paires de chromosomes homologues non ordonnés, de l'inférence quantile pour les données groupées, et de l'inférence de l'indice de Youden et du point de coupure optimal.

DataFest in Canada
DataFest au Canada

Chair/Président: Samantha-Jo Caetano

Organizer/Responsable: Samantha-Jo Caetano

Date: Thursday June 2 / jeudi 2 juin

Time/Heure: 11:00-12:30

Abstract/Résumé

[11:00-11:30]

Nathan A. Taback (University of Toronto)
ASA DataFest@UofT

ASA DataFest@UofT

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.

L'ASA DataFest est en quelque sorte un programmathon destiné aux étudiants de premier cycle, sauf que le problème en est un de science des données plutôtque de programmation. On remet un ensemble de données à des équipes d'étudiants le premier jour et elles s'occupent d'un problème jusqu'au lendemain où elles présentent leurs résultats. Après deux jours intenses d'argumentation, d'analyse de données et de conception de présentation, chaque équipe doit impressionner un jury en seulement quelques minutes et avec deux diapos tout au plus. Des prix sont attribués par catégories. L'ASA DataFest rassemble la communauté de la science des données. Des étudiants du premier cycle effectuent le travail, aidés par des consultants volants qui sont des étudiants de cycles supérieurs ou des professionnels du milieu universitaire ou de l'industrie. L'événement est une occasion d'apprentissage expérientiel pour de nombreux étudiants. J'aborderai mes expériences à titre de professeur d'université qui a organisé une ASA DataFest@UofT et offrirai des conseils pratiques aux professeurs qui songent à organiser une compétition en science des données.

[11:30-12:00]

Karen Buro (MacEwan University) **Jordan A. Slessor** (MacEwan University)
DataFests in Edmonton, 2019 and 2022, Two Perspectives

DataFests à Edmonton, 2019 et 2022, deux points de vue

MacEwan University and the University of Alberta joined to host DataFests in 2019 and 2022. In 2019, five teams from each university analyzed data from the Canadian National Women's Rugby team spending an entire weekend on MacEwan's campus. This year, the University of Alberta will host a virtual DataFest allowing teams from both universities to engage again in a friendly competition and learn from data. Participants have the chance to meet and exchange ideas with indus-

L'Université MacEwan et l'Université de l'Alberta se sont jointes à l'organisation de DataFests en 2019 et 2022. En 2019, cinq équipes de chaque université ont analysé les données de l'équipe nationale féminine canadienne de rugby qui a passé une fin de semaine entière sur le campus de MacEwan. Cette année, l'Université de l'Alberta organisera un DataFest virtuel qui permettra aux équipes des deux universités de participer à nouveau à une compétition amicale et d'apprendre des données. Les participants ont la chance de rencontrer et d'échanger des idées avec des

try experts, university professors, and their peers. We will share our experiences and the impact of these fun events from two perspectives, organizer and participant.

experts de l'industrie, des professeurs d'université et leurs pairs. Nous partagerons nos expériences et l'impact de ces événements amusants de deux points de vue, organisateur et participant.

[12:00-12:30]

Shojaeddin Chenouri (University of Waterloo)

ASA DataFest: The Waterloo Chapter

ASA DataFest : Le chapitre de Waterloo

The American Statistical Association (ASA) DataFest is a weekend-long undergraduate data analysis competition. DataFest was founded at UCLA in 2011, and the Department of Statistics and Actuarial Science at the University of Waterloo joined in 2017. I have been actively involved in organizing the ASA DataFest at Waterloo since the beginning. In this talk, I will share our experience at Waterloo in organizing this event.

Le Datafest de la société américaine de statistique (ASA) est une compétition d'analyse de données pour les étudiants de premier cycle qui se déroule pendant une fin de semaine. Le DataFest a été fondé à L'UCLA en 2011, puis le département de statistiques et de sciences actuarielles de l'université de Waterloo s'est joint à l'événement en 2017. Je me suis grandement impliqué dans l'organisation du ASA DataFest à Waterloo depuis ses débuts. Lors de cet exposé, je partagerai notre expérience concernant l'organisation de cet événement à Waterloo.

Chair/Président: Tim B. Swartz

Organizer/Responsable: Shirley E. Mills

Date: Thursday June 2 / jeudi 2 juin

Time/Heure: 11:00-12:30

Abstract/Résumé

[11:00-11:30]

Brian Macdonald (Yale University)

Age, Experience, and Player Performance

Âge, expérience et performance des joueurs

Athletes improve over time with experience and training, while injuries, general wear-and-tear on the body, and age tend to have the opposite effect. As a result, players typically improve early in their career, reach a peak, and decline later in their career as the impacts of injuries and age overpower any improvements in experience and skills. Performance can improve or erode more slowly or quickly depending on the physical and mental requirements of a particular role a player is responsible for, and understanding these trends is important for team management and media analysts as they attempt to project a player's future performance. We discuss methods of estimating so-called "age curves" and assess their accuracy, apply these methods to multiple statistics in multiple sports, and show how age curves can vary greatly depending on the statistic and the sport.

Les athlètes s'améliorent avec le temps grâce à l'expérience et à l'entraînement, tandis que les blessures, l'usure générale du corps et l'âge ont tendance à avoir l'effet inverse. Par conséquent, les joueurs s'améliorent généralement au début de leur carrière, atteignent un sommet, puis déclinent plus tard dans leur carrière, car l'impact des blessures et de l'âge l'emporte sur toute amélioration de l'expérience et des compétences. Les performances peuvent s'améliorer ou s'éroder plus lentement ou plus rapidement selon les exigences physiques et mentales du rôle particulier dont un joueur est responsable, et la compréhension de ces tendances est importante pour la direction des équipes et les analystes des médias lorsqu'ils tentent de projeter les performances futures d'un joueur. Nous discutons des méthodes d'estimation de ce que l'on appelle les « courbes d'âge » et évaluons leur précision, nous appliquons ces méthodes à plusieurs statistiques dans plusieurs sports et nous montrons comment les courbes d'âge peuvent varier considérablement en fonction de la statistique et du sport.

[11:30-12:00]

Alexander Hinton (Vancouver Whitecaps Football Club)

Data Science at the Vancouver Whitecaps

La science des données chez les Whitecaps de Vancouver

An overview of the role of Data Science at the Vancouver Whitecaps, a professional soccer team competing in the MLS. I will discuss aims of Data Science within the overall performance strategy, and how this may contrast with other analytics departments across professional sporting organizations. The main data sources (event, physical, tracking, broadcast tracking, etc.) available to data scientists in the soccer analytics space will be discussed and compared, with a brief discussion and overview of typical use cases and modeling approaches derived from each one.

Je présente un aperçu du rôle de la science des données chez les Whitecaps de Vancouver, une équipe professionnelle de soccer de la ligue majeure de soccer (MLS). J'aborde les objectifs de la science des données dans la stratégie globale de performance et comment elle se distingue d'autres services d'analyse dans les organisations sportives professionnelles. Les principales sources de données (événements, données physiques, suivi, suivi de diffusion, etc.) à la disposition des scientifiques des données dans l'espace analytique du soccer seront discutées et comparées, avec un bref exposé et un aperçu général de cas d'utilisation typiques et d'approches de modélisation dérivées de chacun.

[12:00-12:30]

Lucas Friesen (Canadian Tire Bank)

Owning The Podium: Supporting Team Canada Through Analytics

À nous le podium : soutenir l'équipe canadienne par l'analytique

Through partnership with Own the Podium (OTP), the Canadian Tire Sports Analytics team provides support to Canada's National Sport Organizations (NSOs) by leveraging performance data and analytics to achieve Olympic and Paralympic objectives. In consultation with each NSO, the Sports Analytics team provides a full suite of services to key sport stakeholders, including data collection & engineering, modeling & analysis, and visualization & reporting. Analysis varies across different sport environments and is heavily influenced by the availability of appropriate data and the individual sport landscape. Each NSO has individualized priorities which usually consist of team/athlete performance evaluation, junior & youth talent identification, career progression assessment, and podium probability estimation. This talk will outline the relationship described above and will present examples of analyses and work being done to meet the objectives of specific Canadian NSOs.

Par notre partenariat avec ≪À nous le podium≫ (≪Own the Podium≫ ou OTP), l'équipe d'analytique sportive de Canadian Tire procure un soutien aux organismes nationaux de sport (ONS) en tirant avantage des données de performances et d'analyses pour atteindre des objectifs relatifs aux jeux olympiques et paralympiques. En consultant les ONS, l'équipe d'analytique sportive offre un éventail complet de services aux intervenants sportifs clés, comme la collecte de données, l'ingénierie, la modélisation, l'analyse, la visualisation et la production de rapports. L'analyse varie selon les différents environnements sportifs et est grandement influencée par la disponibilité des données valides et le contexte spécifique à un sport. Chaque ONS a ses propres priorités concernant généralement l'évaluation de la performance d'un athlète ou d'une équipe, le repérage de jeunes talents, l'évaluation d'une progression de carrière et l'estimation d'atteinte à un podium. Cet exposé décrira la relation mentionnée plus haut et présentera des exemples d'analyses et de travaux réalisés pour atteindre des objectifs spécifiques aux ONS canadiens.

Statistical Applications in P&C Insurance
Applications statistiques dans les assurances IARD

Chair/Président: Mathieu Pigeon

Organizer/Responsable: Mathieu Pigeon

Date: Thursday June 2 / jeudi 2 juin

Time/Heure: 11:00-12:30

Abstract/Résumé

[11:00-11:30]

Anas Abdallah (McMaster University)

An Aggregate Trend Renewal Micro Model for Loss Reserving, with Trend, Inflation and Discount.

Un micro-modèle de renouvellement pour le provisionnement des pertes, avec tendance, inflation et escompte.

The provisions for payment obligations from losses that have occurred but have not yet been settled usually comprise most of the liabilities of an insurance company. Therefore, the determination and evaluation of loss reserving techniques is vital and one of the most critical problems in the insurance industry. This paper addresses new issues in the loss reserving literature by introducing an adaptive micro model in the context of loss reserves. Our model incorporates trend effects, inflation, discount factors and a possible dependence between payments, expenses, and delays. These factors are important for pricing and reserving considerations, as well as for decision making from a risk capital perspective. We first propose to use an aggregate trend renewal model as an individual claim generating process. Some important theoretical results will be then derived, including closed-form expressions for the reserves calculations. A sensitivity analysis of the model will be conducted, in regard to the different assumptions. Specifically, we aim to identify the most influencing factors on the reserves and risk capital estimation. A case study will be performed on a real database, from medical malpractice. Model calibration, heterogeneity, parameters uncertainty, distributions approximations, and risk capital analyses will also be examined.

Les provisions pour les obligations de paiement des sinistres qui se sont produits mais qui n'ont pas encore été réglés constituent généralement la majeure partie du passif d'une compagnie d'assurance. Par conséquent, la détermination et l'évaluation des techniques de provisionnement des pertes sont vitales et constituent l'un des problèmes les plus critiques du secteur de l'assurance. Cet article aborde de nouvelles questions dans la littérature sur le provisionnement des pertes en introduisant un micro-modèle adaptatif dans le contexte des réserves. Notre modèle intègre des effets de tendance, l'inflation, des facteurs d'actualisation et une dépendance possible entre les paiements, les dépenses et les délais paiement. Ces facteurs sont importants pour la tarification et le provisionnement, ainsi que pour la prise de décision dans une perspective de capital du risque. Nous proposons d'abord d'utiliser un modèle agrégé de renouvellement des tendances comme processus de génération de sinistres individuels. Certains résultats théoriques importants seront ensuite dérivés, y compris des expressions à forme fermée pour le calcul des réserves. Une analyse de sensibilité du modèle sera menée, en ce qui concerne les différentes hypothèses. Plus précisément, nous cherchons à identifier les facteurs les plus influents sur l'estimation des réserves et du capital-risque. Une étude de cas sera réalisée sur une base de données réelle, issue de la responsabilité civile médicale. La calibration du modèle, l'hétérogénéité, l'incertitude des paramètres, les approximations des distributions et les analyses du risque de capital seront également examinées.

[11:30-12:00]

Andrei L. Badescu (University of Toronto) **Tsz Chai Fung** (Georgia State University) **X. Sheldon Lin** (University of Toronto)

Fitting censored and truncated regression data using the Mixture of Experts models

Ajustement de données de régression censurées et tronquées à l'aide de modèles de mélange d'experts

The logit-weighted reduced mixture of experts model (LRMoE) is a flexible yet analytically tractable non-

Le modèle de mélange d'experts réduit pondéré par un logit est un modèle de régression non linéaire souple et facile

linear regression model. In this presentation, we extend the Expectation-Conditional-Maximization (ECM) algorithm that efficiently fits the LRMoE to random censored and random truncated regression data. Using real automobile insurance data sets, the usefulness and importance of the proposed algorithm are demonstrated through an actuarial application on individual claim reserving.

à analyser. Dans cette présentation, nous étendons l'algorithme d'espérance-maximisation conditionnelle (qui ajuste efficacement le mélange d'experts réduit pondéré par un logit) aux données de régression aléatoires censurées et tronquées. À l'aide d'ensembles de données d'assurance automobile réels, nous démontrons l'utilité et l'importance de l'algorithme proposé au moyen d'une application actuarielle sur le provisionnement des sinistres individuels.

[12:00-12:30]

Juan-Sebastian Yanez (Université du Québec à Montréal)

Parametric Outstanding Claim Payment Count Modelling Through a Dynamic Claim Score

Modélisation paramétrique du nombre de paiements de sinistres en suspens grâce à un score de sinistres dynamique

By modelling reserves with micro-level models, individual claims information is better preserved and can be more easily handled in the fitting process. Some of the claim information is available immediately at the report date and remains known until the closure of the claim. However, other useful information changes unpredictably as claims develop, for example, the previously observed number of payments. In this paper, we seek to model payment counts in a discrete manner based on past information both in terms of claim characteristics and previous payment counts. We use a dynamic score that weighs the risk of the claim based on previous claim behavior and that gets updated at the end of each discrete interval. In this paper's model we will also distinguish between the different types of payments. We evaluate our model by fitting it into a data set from a major Canadian insurance company.

La modélisation des réserves à l'aide de modèles à micro-niveau permet de mieux préserver les informations sur les sinistres individuels et de les traiter plus facilement dans le processus d'ajustement. Certaines informations sur les sinistres sont immédiatement accessibles à la date du rapport et restent connues jusqu'à la clôture du sinistre. Cependant, d'autres informations utiles changent de manière imprévisible au fur et à mesure que les sinistres évoluent, comme, le nombre de paiements précédemment observés. Dans cet article, nous cherchons à modéliser le nombre de paiements de manière discrète sur la base d'informations antérieures, à la fois en ce qui concerne les caractéristiques du sinistre et le nombre de paiements précédents. Nous utilisons un score dynamique qui pondère le risque de sinistre en fonction du comportement antérieur en matière de sinistres, qui est mis à jour à la fin de chaque intervalle discret. Dans ce modèle, nous faisons également la distinction entre les différents types de paiements. Nous évaluons notre modèle en l'adaptant à un ensemble de données provenant d'une grande compagnie d'assurance canadienne.

Chair/Président: Mireille Schnitzer

Date: Thursday June 2 / jeudi 2 juin

Time/Heure: 11:00-12:30

Abstract/Résumé

[11:00-11:15]

Mariia Samoilenko (Université du Québec à Montréal) **Geneviève Lefebvre** (Université du Québec à Montréal)

On the Power to Detect a Natural Indirect Effect in Causal Mediation Analysis with a Categorical Mediator and a Binary Outcome

Sur la puissance à détecter un effet naturel indirect dans l'analyse de médiation causale avec un médiateur catégoriel et une réponse binaire

Investigations of statistical power in causal mediation analysis are scant, especially for binary outcomes. In parallel, categorical mediators are not rare in epidemiologic practice. In the present work, we aim to generalize our exact regression-based approach for a binary outcome and a binary mediator to mediation settings with a categorical mediator. Namely, we will introduce exact natural effect estimators based on binary outcome and multinomial mediator logistic models. Formulas for the delta method standard errors will be also provided. Since the flexibility of a mediation analysis with a categorical mediator comes at a cost in terms of power, we will present results of a simulation study designed to evaluate capacity to detect natural indirect effect under different scenarios, including ones where the outcome is affected by certain categories of the mediator but not necessarily by others.

Les études sur la puissance statistique de l'analyse de médiation causale sont peu nombreuses, en particulier pour les réponses binaires. En parallèle, les médiateurs catégoriels ne sont pas rares dans la pratique épidémiologique. Dans le présent travail, nous visons la généralisation notre approche exacte pour une réponse et un médiateur binaires basée sur la régression au cas d'un médiateur catégoriel. Plus précisément, nous introduirons des estimateurs exacts des effets naturels basés sur le modèle logistique binaire pour la réponse et le modèle multinomial logistique pour le médiateur. Les formules pour les erreurs standards par la méthode delta seront également présentées. Puisque la flexibilité d'une analyse de médiation avec un médiateur catégoriel a un coût en termes de puissance, nous présenterons les résultats d'une étude de simulation conçue pour évaluer la capacité à détecter un effet naturel indirect sous différents scénarios, y compris ceux où la réponse est affectée par certaines catégories du médiateur mais pas nécessairement par d'autres.

[11:15-11:30]

Md Rashedul Hoque (Simon Fraser University) **Yi Qian** (University of British Columbia) **Lawrence McCandless** (Simon Fraser University) **J. Antonio Aviña-Zubieta** (University of British Columbia) **Mary A. De Vera** (University of British Columbia) **Hui Xie** (Simon Fraser University)

An Index of Sensitivity to Non-Exchangeability

Indice de sensibilité à la non-échangeabilité

Standard statistical methods assuming the exchangeability of units between treatment groups can yield biased treatment effect estimates if the assumption does not hold. Existing methods evaluate the sensitivity of treatment effect estimates to non-exchangeability due to unobserved confounders only. We propose an index of sensitivity to non-exchangeability (ISENSE). Unlike many existing methods, it does not require any assumptions regarding the distribution or number of unmea-

Les méthodes statistiques standards qui reposent sur l'hypothèse de l'interchangeabilité des unités entre les groupes de traitement peuvent donner des estimations biaisées de l'effet de traitement si cette hypothèse n'est pas vérifiée. Les méthodes actuelles évaluent la sensibilité des estimations de l'effet de traitement à la non-échangeabilité causée uniquement par des facteurs de confusion non observés. Nous proposons un indice de sensibilité à la non-échangeabilité. Contrairement aux nombreuses méthodes connues, cet indice ne nécessite aucune hypothèse quant

sured confounders, and it can handle both unmeasured confounders and reverse causality. ISENSE is a local sensitivity method based on a Taylor-series approximation to the non-exchangeability likelihood, evaluated at the parameter estimates under exchangeability. One can interpret ISENSE intuitively through the MinNE statistic, which captures the minimum non-exchangeability to change the results. We evaluate ISENSE using simulation studies and illustrate its use with an example using administrative data.

à la distribution ou au nombre de facteurs de confusion non mesurés, et il peut traiter à la fois les facteurs de confusion non mesurés et la causalité inverse. L'indice de sensibilité à la non-échangeabilité est une méthode de sensibilité locale qui repose sur une approximation en série de Taylor de la vraisemblance de la non-échangeabilité, évaluée aux estimations des paramètres en cas d'échangeabilité. On peut interpréter l'indice de sensibilité à la non-échangeabilité de manière intuitive au moyen de la statistique MinNE qui reflète la non-échangeabilité minimale pour modifier les résultats. Nous analysons l'indice de sensibilité à la non-échangeabilité à l'aide d'études de simulation et illustrons son utilisation par un exemple avec des données administratives.

[11:30-11:45]

Eric Morenz (University of Washington)
Statistical Anatomy of Autopsy Studies
Anatomie statistique des études d'autopsie

Drawing causal inference from observational studies is challenging because the exposure-outcome relationship is likely confounded. Additional hurdles present themselves when there is a random time element to the outcome. While many biomarkers can be recorded in vivo, others can only be measured by analyzing tissues collected during autopsy. This is the case, for example, of many neurological biomarkers that require sampling brain tissues; outcome data collection can only occur when a participant dies. Comparing observed biomarker values across exposure groups can be highly misleading when the exposure under consideration affects survival. This occurs because such comparison ignores that observation times then tend to be different across exposure groups, and that biomarker values typically vary temporally across time. In this work, we present a causal inference framework for studying the effect of a point-exposure on a time-varying biomarker process that can only be sampled at death.

Il est difficile d'établir une inférence causale à partir d'études d'observation, car la relation entre l'exposition et le résultat est probablement confondue. Des obstacles supplémentaires se présentent lorsque le résultat comporte un élément de temps aléatoire. Si de nombreux biomarqueurs peuvent être mesurés in vivo, d'autres ne peuvent l'être que par l'analyse de tissus prélevés lors d'une autopsie. C'est le cas, par exemple, de nombreux biomarqueurs neurologiques qui nécessitent le prélèvement de tissus cérébraux. Ainsi, la collecte de données sur les résultats ne peut avoir lieu qu'au moment du décès du participant. La comparaison des valeurs observées des biomarqueurs entre les groupes d'exposition peut être très trompeuse lorsque l'exposition en question a une incidence sur la survie. En effet, une telle comparaison ne tient pas compte du fait que les temps d'observation sont souvent différents entre les groupes d'exposition, et que les valeurs des biomarqueurs varient en général dans le temps. Dans le cadre de ces travaux, nous présentons un cadre d'inférence causale pour étudier l'effet d'une exposition ponctuelle sur un processus de biomarqueur variant dans le temps qui ne peut être échantillonné qu'au moment du décès.

[11:45-12:00]

Blair Bilodeau (University of Toronto) **Linbo Wang** (University of Toronto) **Daniel M. Roy** (University of Toronto)
Adaptively Exploiting d-Separators with Causal Bandits
Exploitation adaptative des séparateurs d avec des bandits causaux

Multi-armed bandit problems provide a framework to identify the optimal intervention over a sequence of repeated experiments. Without additional assumptions, minimax optimal performance (measured by cumulative regret) is well-understood. When observed variables d-separate the intervention from the outcome, causal

Les problèmes de bandit à bras multiples fournissent un cadre pour déterminer l'intervention optimale sur une séquence d'expériences répétées. En l'absence d'hypothèses supplémentaires, on comprend bien le rendement optimal du minimax (mesuré par le regret cumulatif). Lorsque les variables d observées séparent l'intervention du résultat, les algorithmes de bandit causal engendrent

bandit algorithms provably incur less regret. However, an ideal algorithm should be adaptive; that is, perform nearly as well as an algorithm with oracle knowledge of whether a d-separator is observed, without requiring this knowledge. We formalize this notion of adaptivity, and provide a new algorithm that achieves (a) optimal regret when a d-separator is observed, improving on classical algorithms, and (b) significantly less regret than causal bandit algorithms when no d-separator is observed. Crucially, our algorithm does not require any oracle knowledge of the presence of a d-separator. We also generalize this adaptivity to other conditions, e.g. the front-door criterion.

manifestement moins de regret. Cependant, un algorithme idéal devrait être adaptatif, c'est-à-dire qu'il devrait être presque aussi efficace qu'un algorithme avec une connaissance d'oracle pour savoir si un séparateur d est observé, sans avoir besoin de cette connaissance. Nous définissons cette notion d'adaptabilité et fournissons un nouvel algorithme qui atteint : a) un regret optimal lorsqu'un séparateur d est observé, améliorant ainsi les algorithmes classiques, et b) un regret significativement moindre que les algorithmes de bandit causal lorsqu'aucun séparateur d n'est pas observé. En outre, notre algorithme ne nécessite aucune connaissance d'oracle quant à la présence d'un séparateur d. Nous étendons également cette adaptabilité à d'autres conditions, par exemple au critère de la porte d'entrée.

[12:00-12:15]

Lijia Wang (University of Waterloo) **Yeying Zhu** (University of Waterloo) **Richard J. Cook** (University of Waterloo)

A Doubly Robust Joint Modelling Approach of Multiple Uncausally Correlated Mediators

Approche de modélisation conjointe doublement robuste de multiples médiateurs corrélés de manière non causale

Causal mediation analysis has been of great interest, due to its strong ability for disentangling the effects of a treatment on an outcome via a variety of paths through either the mediator(s) or the treatment. Recently, mediation analysis on multiple mediators is attracting much attention, where the relationship between the multiple mediators play an important role. In this paper, we review and extend the concept of multiple mediators uncausally related, which depicts the phenomenon that the multiple mediators are related given the baseline covariates but their correlation structure cannot be causally ordered or identified. We further provide a copula-based approach jointly modelling the mediators. A doubly robust approach is also proposed to tackle model misspecification. Theoretical properties and simulation studies are also presented, with the theoretical standard error derived based on the sandwich formula. We finally apply the proposed method on a genetic psychiatric study dataset.

L'analyse de médiation causale a suscité un grand intérêt en raison de sa forte capacité à démêler les effets d'un traitement sur un résultat via une variété de chemins passant par le ou les médiateurs ou le traitement. Récemment, l'analyse de médiation sur des médiateurs multiples attire beaucoup d'attention, la relation entre les médiateurs multiples jouant un rôle important. Dans cet article, nous examinons et étendons le concept de médiateurs multiples corrélés de manière non causale, qui décrit le phénomène selon lequel les médiateurs multiples sont liés sachant les covariables de base, mais sans que leur structure de corrélation ne puisse être ordonnée ou identifiée de manière causale. Nous proposons en outre une approche basée sur les copules pour modéliser conjointement les médiateurs. Nous proposons également une approche doublement robuste pour lutter contre la mauvaise spécification du modèle. Nous présentons les propriétés théoriques et des études de simulation, l'erreur type théorique étant dérivée sur la base de la formule sandwich. Nous appliquons enfin la méthode proposée à un ensemble de données d'une étude psychiatrique génétique.

Chair/Président: Olli Saarela

Date: Thursday June 2 / jeudi 2 juin

Time/Heure: 11:00-12:15

Abstract/Résumé

[11:00-11:15]

Changchang Xu (University of Toronto) **Changchang Xu** (University of Toronto; Lunenfeld-Tanenbaum Research Institute, Sinai Health) **Shelley B. Bull** (University of Toronto; Lunenfeld-Tanenbaum Research Institute, Sinai Health)

Improving Mixture Cure Modelling of Molecular Genetic Biomarkers in Cancer Prognosis by Penalized Likelihood with Profile Likelihood Interval Estimation

Améliorer la modélisation du mélange avec taux de guérison des biomarqueurs génétiques moléculaires dans le pronostic du cancer par vraisemblance pénalisée avec une estimation des intervalles de vraisemblance du profil

In analysis of time-to-event data, the mixture cure (MC) model is more appropriate than conventional Cox proportional hazards model when the study sample includes long-term survivors. In samples with few events, where standard maximum likelihood (ML) can be biased, Firth-type penalized likelihood (FT-PL) effectively reduces bias and improves efficiency of parameter estimation under MC models. However, as the Wald-type confidence interval (CI) may not be valid, we developed profile likelihood confidence interval (PLCI) and confidence region (PLCR) methods for parameter inference. Via data-based simulation studies, we found that, for the true value coverage rate: 1) FT-PL exceeds ML, 2) PLCR is more resilient to low event rates than PLCI, 3) profile likelihood-based methods surpass Wald-type ones. We also illustrate the practicality and strength of FT-PL and PLCI/PLCR for MC analysis in a cohort study of breast cancer prognosis with long-term followup for disease free survival.

Dans l'analyse de données de temps d'événement, le modèle de mélange avec taux de guérison est plus approprié que le modèle à risques proportionnels de Cox conventionnel lorsque l'échantillonnage de l'étude comprend des survivants à long terme. Dans les échantillons avec peu d'événements dans lesquels le maximum de vraisemblance standard peut être biaisé, la vraisemblance pénalisée de Firth (FT-PL) réduit effectivement le biais et améliore l'efficacité de l'estimation des paramètres sous les modèles de Monte-Carlo (MC). Cependant, comme l'intervalle de confiance (CI) de Wald peut ne pas être valide, nous avons développé des méthodes d'intervalles de confiance par vraisemblance du profil (PLCI) et par régions de confiance (PLCR) pour l'inférence des paramètres. À l'aide d'études en simulation avec données, nous avons trouvé pour le taux de couverture de la valeur réelle que : 1) FT-PL dépasse ML, 2) PLCR est plus résilient que PLCI à faible taux d'événements, 3) les méthodes fondées sur la vraisemblance du profil sont supérieures à celles de type Wald. Nous illustrons également la praticité et la force de FT-PL et de PLCI/PLCR pour une analyse MC dans une étude de cohorte de pronostics de cancer du sein avec un suivi à long terme de survie sans la maladie.

[11:15-11:30]

Shenita Pramij (Memorial University of Newfoundland) **Candemir Cigsar** (Memorial University of Newfoundland)

A Dynamic Model for the Analysis of Recurrent Events with Application to Epidemic Data

Un modèle dynamique pour l'analyse d'événements récurrents avec application aux données épidémiques

Models and methods for the statistical analysis of recurrent events can be useful to make inferences on epidemic processes. In this talk, we introduce a new dynamic model for recurrent event processes, which can include internal and external covariates, and dynamically adapts to change points. We discuss the estimation

Les modèles et méthodes pour l'analyse statistique d'événements récurrents peuvent être utiles afin de faire des inférences concernant les processus épidémiques. Dans cette présentation, nous introduisons un nouveau modèle paramétrique dynamique pour les processus d'événements récurrents, qui peut inclure des covariables internes et externes, et s'adapte dynamiquement à cer-

of model parameters and asymptotic properties of the estimators under different settings. We present the results of a simulation study conducted to investigate the finite sample properties of the estimators and their robustness against model misspecifications. Finally, we discuss possible applications of our model to epidemic processes and illustrate our methods with an infectious disease dataset.

tains points de changement. Nous discutons de l'estimation des paramètres du modèle et des propriétés asymptotiques des estimateurs dans diverses situations. Nous présentons les résultats d'une étude de simulation menée pour étudier les propriétés d'échantillon de taille finie des estimateurs, ainsi que leur robustesse face aux erreurs de spécification du modèle. Enfin, nous discutons des possibilités d'application de nos méthodes aux processus épidémiques, et illustrons nos méthodes en utilisant un jeu de données provenant d'une étude épidémiologique de maladies infectieuses.

[11:30-11:45]

Shakhawat Hossain (University of Winnipeg) **Jody Krahn** (Statistics Canada) **Shahedul Khan** (University of Saskatchewan)

An Efficient Estimation Approach to Joint Modelling of Longitudinal and Survival Data

Une approche d'estimation efficace pour la modélisation conjointe de données longitudinales et de survie

Joint models typically combine linear mixed effects models for repeated measurement data and Cox models for survival time. When we are jointly modelling the longitudinal and survival data, variable selection and efficient estimation of parameters are especially important for performing reliable statistical analyses. In this talk, we discuss the pretest and shrinkage estimation methods for jointly modelling longitudinal and survival time data when some of the covariates in the model may not be relevant for predicting survival times. In this situation, we fit two models: the full model that contains all the covariates and the subset model that contains a reduced number of covariates. We combine the full model estimators and the estimators that are restricted by a linear hypothesis to define pretest and shrinkage estimators. We provide their numerical mean squared errors (MSE) and relative MSE. Our proposed methods are illustrated by extensive simulation studies and a real-data example.

Les modèles conjoints combinent généralement les modèles linéaires à effets mixtes pour les données de mesures répétées et les modèles de Cox pour le temps de survie. Lorsqu'on modélise conjointement les données longitudinales et de survie, la sélection de variables et l'estimation efficace des paramètres sont particulièrement importantes afin de réaliser des analyses statistiques fiables. Lors de cet exposé, nous aborderons les méthodes d'estimation de rétrécissement et de prétest pour la modélisation conjointe de données longitudinales et de survie lorsque certaines covariables dans le modèle ne sont peut-être pas pertinentes pour prédire les temps de survie. Dans cette situation, nous ajustons deux modèles : le modèle complet contenant toutes les covariables et le modèle de sous-ensemble qui contient un nombre réduit de covariables. Puis nous combinons les estimateurs du modèle complet avec les estimateurs restreints par une hypothèse linéaire afin de définir les estimateurs de rétrécissement et de prétest. Enfin, nous obtenons leurs erreurs quadratiques moyennes numériques (MSE) et MSE relative. Les méthodes que nous proposons sont illustrées à l'aide d'études en simulation approfondies et d'exemples tirés de données réelles.

[11:45-12:00]

Awa Diop (Université Laval) **Denis Talbot** (Université Laval) **Caroline Sirois** (Université Laval)

History-Restricted Marginal Structural Model and Latent Class Growth Modeling of Treatment Trajectories for a Time-Dependent Outcome

Modèles structurels marginaux à historique restreint et modèles d'analyse de trajectoires groupées pour une issue qui varie dans le temps

Combining latent class growth models (LCGM) and marginal structural models (MSMs) is useful to summarize numerous time-varying treatment patterns into a few interpretable trajectory groups and to give a direct population-level causal interpretation. However, the

Combiner les modèles d'analyse de trajectoires (LCGM) et les modèles marginaux structurels (MSM) permet de résumer en quelques groupes un traitement qui varie dans le temps et de donner une interprétation causale au niveau populationnel. Toutefois, le LCGM-MSM n'est pas approprié en présence d'une issue qui

LCGM-MSM framework is not suitable when the outcome is time-dependent. Instead, we propose combining a nonparametric history-restricted marginal structural model (HRMSM) with LCGM. HRMSMs consist of defining a shorter history of exposure and are seen as a generalization of standard MSMs. To the best of our knowledge, we present the first application of HRMSMs with a time-to-event outcome. It was previously noted that HRMSMs could pose interpretation problems in survival analysis. The causal parameter we propose circumvents this caveat. We consider three different estimators of the parameters: inverse probability of treatment weighting, g-computation, and a pooled longitudinal targeted maximum likelihood estimator.

varie dans le temps. On propose plutôt de combiner un MSM à historique restreint (HRMSM) avec le LCGM. Les HRMSM permettent de définir plusieurs historiques réduits de l'exposition et sont présentés comme une généralisation des MSM. Au mieux de nos connaissances, nous proposons la première application des HRMSM pour une issue de survie. Aussi, il est relevé que les HRMSM peuvent poser un problème d'interprétation en analyse de survie. Notre définition du paramètre causal permet de contourner ce problème. On considère trois estimateurs différents des paramètres : l'inverse de la probabilité de traitement, la formule-g et l'estimateur ciblé par le maximum de vraisemblance groupé.

[12:00-12:15]

Denis Larocque (HEC Montréal) **Weichi Yao** (New York University) **Halina Frydman** (New York University) **Jeffrey S. Simonoff** (New York University)

Ensemble Methods for Survival Function Estimation with Time-Varying Covariates

Méthodes d'ensemble pour l'estimation de la fonction de survie avec covariables qui varient dans le temps

Survival data with time-varying covariates are common in practice. If relevant, they can improve on the estimation of survival function. However, the traditional survival forests - conditional inference forest, relative risk forest and random survival forest - have accommodated only time-invariant covariates. We generalize the conditional inference and relative risk forests to allow time-varying covariates. We also propose a general framework for estimation of a survival function in the presence of time-varying covariates. We compare their performance with that of the Cox model and transformation forest, adapted here to accommodate time-varying covariates, through a comprehensive simulation study in which the Kaplan-Meier estimate serves as a benchmark. Taking into account all other factors, under the proportional hazard (PH) setting, the best method is always one of the two proposed forests, while under the non-PH setting, it is the adapted transformation forest.

La présence de covariables qui varient dans le temps est courante dans les analyses de survie. Le cas échéant, ils peuvent améliorer l'estimation de la fonction de survie. Cependant, les forêts de survie traditionnelles - forêt d'inférence conditionnelle, forêt à risque relatif et forêt aléatoire de survie – peuvent seulement utiliser des covariables invariantes dans le temps. Nous généralisons les forêts d'inférence conditionnelle et de risque relatif pour permettre d'intégrer des covariables qui varient dans le temps. Nous proposons également un cadre général pour l'estimation d'une fonction de survie en présence de covariables qui varient dans le temps. Nous comparons leurs performances avec celles du modèle de Cox et de la forêt de transformation, adaptée ici pour tenir compte des covariables qui varient dans le temps, grâce à une étude de simulation complète dans laquelle l'estimation de Kaplan-Meier sert de référence. Globalement, la meilleure méthode est toujours l'une des deux forêts proposées dans les cas avec risque proportionnel (PH), tandis que la forêt de transformation adaptée domine dans les cas non-PH.

Information on NSERC Competition Results and Discovery Grant Preparation
Information sur les résultats du concours du CRSNG et la préparation des subventions à la découverte

Chair/Président: Joanna Elizabeth Mills Flemming

Organizer/Responsable: Henrik Stryhn

Date: Thursday June 2 / jeudi 2 juin

Time/Heure: 12:30-13:30

Abstract/Résumé

[12:30-13:30]

Adele Ngi-Song (NSERC) **Caroline Bicker** (NSERC) **Aurélie Labbe** (HEC Montreal)

Information on NSERC Competition Results and Discovery Grant Preparation

Information sur les résultats du concours du CRSNG et la préparation des subventions à la découverte

This session will commence with a short presentation of results from the 2022 Competition for EG1508. NSERC staff will then discuss the NOI (Notification of Intent to Apply) and Full Application process, the Discovery Grant evaluation process principles (criteria and ratings), and tips for preparing a Discovery Grant application. There will be an opportunity for participants to ask questions.

Cette session débutera par une brève présentation des résultats du concours 2022 pour le GE1508. Le personnel du CRSNG discutera ensuite de l'Avis d'intention (Avis d'intention de présenter une demande) et du processus de demande complète, des principes du processus d'évaluation des subventions à la découverte (critères et cotes) et des conseils pour préparer une demande de subvention à la découverte. Les participants auront l'occasion de poser des questions.

Chair/Président: Linbo Wang

Organizer/Responsable: Linbo Wang

Date: Thursday June 2 / jeudi 2 juin

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-14:00]

Yuexia Zhang (University of Toronto) **Jianhua Hu** (Columbia University)

High dimensional mediation analysis for microbiome data

Analyse de médiation en haute dimension pour données du microbiome

Mediation analysis is an important tool to study casual associations in biomedical and other scientific areas and has recently gained attention in microbiome studies. With a microbiome study of acute myeloid leukemia (AML) patients, we investigate whether the effect of induction chemotherapy intensity levels on the infection status is mediated by the microbial taxa abundance. The unique characteristics of the microbial mediators—high-dimensionality, zero-inflation, and dependence—call for new methodological developments in mediation analysis. The presence of an exposure-induced mediator-outcome confounder, antibiotics usage, further requires a delicate treatment in the analysis. To address these unique challenges brought by our motivating microbiome study, we propose a novel nonparametric identification formula for the interventional indirect effect (IIE), a measure recently developed for studying mediation effects, and develop the corresponding estimation algorithm and test. Both simulation studies and the AML microbiome study demonstrate the promise of the proposed method.

L'analyse de médiation est un outil important pour l'étude des associations occasionnelles dans le domaine biomédical et dans autres domaines scientifiques. Elle a récemment attiré l'attention dans les études sur le microbiome. Sur la base d'une étude du microbiome de patients atteints de leucémie myéloïde aiguë (LMA), nous cherchons à savoir si l'effet des niveaux d'intensité de la chimiothérapie d'induction sur le statut infectieux est médié par l'abondance des taxons microbiens. Les caractéristiques uniques des médiateurs microbiens - haute dimensionnalité, surreprésentation de zéros et dépendance - exigent de nouvelles méthodes d'analyse de médiation. La présence d'un facteur de confusion médiateur-résultat induit par l'exposition, l'utilisation d'antibiotiques, exige en outre un traitement délicat dans l'analyse. Pour relever ces défis uniques posés par notre étude motivante sur le microbiome, nous proposons une nouvelle formule d'identification non paramétrique pour l'effet d'intervention indirect, mesure récemment développée pour étudier les effets de médiation, et nous développons l'algorithme d'estimation et le test correspondants. Les deux études de simulation et l'étude du microbiome de la LMA démontrent le potentiel de la méthode proposée.

[14:00-14:30]

Geneviève Lefebvre (Université du Québec à Montréal) **Miguel Caubet Fernandez** (Université du Québec à Montréal) **Mariia Samoilenko** (Université du Québec à Montréal)

Investigating the Performance of the Exact Estimator for Causal Mediation Analysis of Binary Outcomes and Binary Mediators in Case-control Designs

Étude de la performance de l'estimateur exact pour l'analyse de médiation causale pour les réponses et médiateurs binaires dans les devis cas-témoins

In the causal mediation analysis framework, regression-based approaches have been introduced in past years for decomposing the total effect of an exposure on a binary outcome into a direct effect and an indirect ef-

En analyse de médiation causale, plusieurs approches basées sur la régression ont été introduites ces dernières années pour décomposer l'effet total d'une exposition sur une réponse binaire en un effet direct et un effet indirect à travers un médiateur

fect through a target mediator. In this context, a well-known strategy involves specifying a logistic model for the outcome and invoking the rare outcome assumption (ROA) to simplify estimation. Recently, an exact estimator for natural direct and indirect effects has been introduced to circumvent the challenges prompted by the ROA, but the approach cannot be used as is on case-control data. Considering a binary mediator, we propose adapting the exact estimator for outcome-selected samples using inverse-probability-weighting. Contrary to estimators relying on the ROA, the exact approach may be more delicate to use in this setting. We investigate this hypothesis and examine the performance of the exact estimator based on case-control data with varied outcome prevalence.

d'intérêt. Dans ce contexte, une stratégie bien connue consiste à spécifier un modèle logistique pour la réponse et à invoquer l'hypothèse de la réponse rare (HRR) pour simplifier l'estimation. Récemment, un estimateur exact pour les effets direct et indirect naturels a été introduit pour contourner les problèmes induits par la HRR, mais l'approche ne peut pas être utilisée telle quelle sur des données cas-témoins. En considérant un médiateur binaire, nous proposons une adaptation de l'estimateur exact via pondération par probabilité inverse. Contrairement aux estimateurs d'effets naturels reposant sur la HRR, l'approche exacte pourrait être plus délicate à utiliser dans les devis cas-témoins. Dans le présent travail, nous étudions cette hypothèse et examinons la performance de l'estimateur exact sur des données cas-témoins avec différentes prévalences de la réponse.

[14:30-15:00]

Dehan Kong (University of Toronto) **Zhenhua Lin** (National University of Singapore) **Linbo Wang** (University of Toronto)

Causal Inference on Distribution Functions

Inférence causale sur des fonctions de distribution

Understanding causal relationships is one of the most important goals of modern science. So far, the causal inference literature has focused almost exclusively on outcomes coming from the Euclidean space $\mathbb{R}^p$. However, it is increasingly common that complex datasets collected through electronic sources, such as wearable devices, cannot be represented as data points from $\mathbb{R}^p$. In this paper, we present a novel framework of causal effects for outcomes from the Wasserstein space of cumulative distribution functions, which in contrast to the Euclidean space, is non-linear. We develop doubly robust estimators and associated asymptotic theory for these causal effects. As an illustration, we use our framework to quantify the causal effect of marriage on physical activity patterns using wearable device data collected through the National Health and Nutrition Examination Survey.

L'un des objectifs principaux de la science moderne est de comprendre les liens de cause à effet. Jusqu'à présent, la documentation portant sur l'inférence causale se concentre presque exclusivement sur les résultats tirés de l'espace euclidien $\mathbb{R}^p$. Cependant, il est très fréquent que des jeux de données complexes provenant de sources électroniques (comme des appareils portables) ne puissent pas être représentés en point de données à partir de $\mathbb{R}^p$. Dans cet article, nous présentons un nouveau cadre d'effets causaux pour des résultats tirés de l'espace Wasserstein des fonctions de distribution cumulatives, qui est non linéaire contrairement à l'espace euclidien. Nous élaborons des estimateurs doublement robustes et une théorie asymptotique associée pour ces effets causaux. En guise d'exemple, nous employons notre méthode pour quantifier l'effet causal du mariage sur des facteurs d'activité physique à l'aide de données tirées d'appareils portables provenant de la National Health and Nutrition Examination Survey.

Chair/Président: Xuekui Zhang

Date: Thursday June 2 / jeudi 2 juin

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-14:00]

Lynn Lin (Duke University)

Multi-source Single-cell Data Integration by MAW Barycenter for Gaussian Mixture Models

Intégration de données à cellule unique et sources multiples par barycentre MAW pour les modèles de mélanges gaussiens

One key challenge encountered in single-cell-data clustering is to combine clustering results of datasets acquired from multiple sources. We propose to represent the clustering result of each dataset by a Gaussian mixture model (GMM) and produce an integrated result based on the notion of Wasserstein barycenter. However, the precise barycenter of GMMs, a distribution on the same sample space, is computationally infeasible to solve. Importantly, the barycenter of GMMs may not be a GMM containing a reasonable number of components. We thus propose to use the Minimized Aggregated Wasserstein (MAW) distance to approximate the Wasserstein metric and develop a new algorithm for computing the barycenter of GMMs under MAW. Recent theoretical advances further justify using the MAW distance as an approximation for the Wasserstein metric between GMMs. Our proposed algorithm for clustering integration scales well with the data dimension and the number of mixture components, with complexity independent of data size. We demonstrate that the new method achieves better clustering results on several single-cell RNA-seq datasets than some other popular methods.

L'un des défis principaux que l'on rencontre dans le regroupement des données à cellule unique est la combinaison de résultats de regroupement de jeux de données tirés de sources multiples. Nous proposons de représenter le résultat de regroupement de chaque ensemble de données par un modèle de mélanges gaussiens (GMM) et de produire un résultat intégré à partir de la notion de barycentre de Wasserstein. Cependant, le barycentre exact des GMM (une distribution sur le même espace d'échantillon) est informatiquement impossible à résoudre. De plus, le barycentre des GMM n'est pas nécessairement un GMM contenant un nombre raisonnable de composés. Pour cette raison, nous proposons d'utiliser la distance de Wasserstein agrégée et minimisée (MAW) pour estimer la métrique de Wasserstein et développer un nouvel algorithme servant à calculer le barycentre des GMM selon MAW. Les avancés théoriques récentes justifient davantage l'emploi de la distance de MAW en guise d'approximation de la métrique de Wasserstein entre les GMM. L'algorithme que nous proposons pour intégrer les regroupements s'échelonne bien dans la dimension de données et le nombre de composés de mélanges, avec complexité indépendante de la taille d'échantillon. Nous démontrons que la nouvelle méthode réussit à obtenir de meilleurs résultats de regroupement dans plusieurs jeux de données de séquençage de l'ARN à cellule unique par rapport à d'autres méthodes bien connues.

[14:00-14:30]

Lihui Zhao (Northwestern University)

Dynamic Risk Prediction for Cardiovascular Events

Prédiction de risque dynamique pour les événements cardiovasculaires

Cardiovascular disease (CVD) is a leading cause of morbidity and mortality. CVD risk prediction plays a central role in clinical CVD prevention strategies, by aiding decision making for lifestyle modification and to match the intensity of therapy to the absolute risk of a given patient. Various CVD risk factors have been identified

La maladie cardiovasculaire (CVD) est l'une des principales causes de morbidité et de mortalité. La prédiction de risque de CVD joue un rôle central dans les stratégies cliniques de prévention de CVD, en orientant la prise de décision relative au changement de mode de vie et en ajustant l'intensité du traitement au risque absolu d'un patient donné. De nombreux facteurs de

and used to construct multivariate risk prediction algorithms. However, these algorithms are generally based on the risk factors measured at a single time. Since risk factors like blood pressure are regularly collected in clinical practice, and electronic medical records are making longitudinal data on these risk factors available to clinicians, dynamic prediction of CVD risk on a real-time basis using the history of CV risk factors will likely improve the precision of personalized CVD risk prediction. We will present statistical methods to build dynamic CVD risk prediction models using repeated measured risk factor levels. The pooled data from multiple community-based CVD cohorts will be used.

risque de CVD ont été repérés et utilisés pour construire des algorithmes multivariés de prédiction de risque. Cependant, ces algorithmes se basent généralement sur les facteurs de risque mesurés à un certain temps. Puisque les facteurs de risque comme la pression artérielle sont régulièrement recueillis en pratique clinique, et que les dossiers médicaux électroniques procurent des données longitudinales à partir de ces facteurs de risque aux médecins, il semble avantageux d'employer une prédiction dynamique du risque de CVD à temps réel à partir de l'historique des facteurs de risque cardiovasculaire pour améliorer la précision de la prédiction de risque de CVD personnalisée. Nous présenterons les méthodes statistiques pour construire des modèles de prédiction dynamique de risque de CVD à l'aide de niveaux de facteur de risque à mesures répétées. Nous exploiterons les données regroupées provenant des multiples cohortes de CVD basées sur une communauté.

[14:30-15:00]

Kailun Bai (University of Victoria)

scSorterDL: a cell type annotation tool for single-cell RNA sequencing data

scSorterDL : outil d'annotation du type de cellule pour données de séquençage d'ARN unicellulaire

With the rise of single-cell transcriptome sequencing technology, more and more studies focus on single-cell level, and at the same time. This creates not only new opportunities for biologists to study cells at higher resolution but also demands for bioinformaticians to develop automated cell annotation methods to analyze these new data types. scRNA-seq datasets usually contain the expression levels of tens of thousands of genes, and often contain a lot of technical noise, which renders cell annotation very challenging. Here, we describe our cell annotation method, which relies on the novelty of the ensemble methods and deep learning to achieve its capabilities, that can appropriately address the sparsity and high-dimensionality of scRNA-seq data without sacrificing classification accuracy or speed. Our tool has been implemented in pytorch, and has the additional advantage of being GPU and CPU parallelizable and is well suited for large datasets.

Avec l'essor de la technologie de séquençage du transcriptome de la cellule unique, de plus en plus d'études se concentrent sur le niveau de la cellule unique en même temps. Cela permet aux biologistes d'étudier les cellules à plus haute résolution, mais exigent également des bioinformaticiens qu'ils développent des méthodes d'annotation cellulaire automatisées pour analyser ces nouveaux types de données. Les ensembles de données scRNA-seq contiennent généralement les niveaux d'expression de dizaines de milliers de gènes et contiennent souvent beaucoup de bruit technique, ce qui rend l'annotation cellulaire très difficile. Nous décrivons ici notre méthode d'annotation cellulaire, qui s'appuie sur la nouveauté des méthodes d'ensemble et de l'apprentissage profond pour atteindre ses capacités, qui peut traiter de manière appropriée la sparsité et la haute dimensionnalité des données scRNA-seq sans sacrifier la précision ou la vitesse de classification. Notre outil a été implémenté dans pytorch et a l'avantage supplémentaire d'être parallélisable par GPU et CPU et d'être bien adapté aux grands ensembles de données.

Statistical Challenges in Deep Learning
Défis statistiques en apprentissage profond

Chair/Président: Vahid Partovi Nia

Organizer/Responsable: Vahid Partovi Nia

Date: Thursday June 2 / jeudi 2 juin

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-14:00]

Masoud Asgharian (McGill University)

Machine Learning and Neural Networks: Foundations and Some Fundamental Questions

Apprentissage automatique et réseaux neuronaux : fondements et questions fondamentales

Statistical learning theory is by now a mature branch of data science that hosts a vast variety of practical techniques for tackling data-related problems. In this talk we present some fundamental concepts upon which statistical learning theory has been based. Function estimation as the heart of learning theory is emphasized. Accordingly, we pay a closer attention to the so-called mapping neural networks and try to shed some light on certain theoretical aspects of them. We highlight some of the fundamental challenges that have attracted the attention of researchers and are yet to be fully resolved. One of these challenges is estimation of the intrinsic dimension of data that will be discussed in details. Another challenge is inferring causal direction when the training data set is not representative of the target population.

La théorie de l'apprentissage statistique est désormais une branche mature de la science des données qui inclut une grande variété de techniques pratiques qui permettent de résoudre divers problèmes liés aux données. Dans cet exposé, nous présentons certains concepts sur lesquels la théorie de l'apprentissage statistique a été fondée. Nous nous concentrons sur l'estimation des fonctions, qui est au cœur de la théorie de l'apprentissage. Nous nous intéressons donc de plus près aux réseaux neuronaux dits de cartographie et tâchons de faire la lumière sur certains de leurs aspects théoriques. Nous soulignons des défis fondamentaux qui n'ont pas encore été résolu et qui ont attiré l'attention des chercheurs, notamment l'estimation de la dimension intrinsèque des données. Un autre défi consiste à déduire la direction causale lorsque l'ensemble de données d'entraînement n'est pas représentatif de la population cible.

[14:00-14:30]

Ali Ghodsi (University of Waterloo) **Mojtaba Valipour** (Cornell University) **Bowen You** (Cornell University) **Maysum Panju** (Cornell University)

SymbolicGPT: A Generative Transformer Model for Symbolic Regression

SymbolicGPT : Un modèle de transformateur génératif pour la régression symbolique

Symbolic regression is the task of identifying a mathematical expression that best fits a provided dataset of input and output values. Due to the richness of the space of mathematical expressions, symbolic regression is generally a challenging problem. While conventional approaches based on genetic evolution algorithms have been used for decades, deep learning-based methods are relatively new and an active research area. In this work, we present SymbolicGPT, a novel transformer-based language model for symbolic regression. This model exploits the advantages of probabilistic language models like GPT, including strength in performance and flexi-

La régression symbolique consiste à identifier une expression mathématique qui s'adapte le mieux à un jeu de données d'entrées et de sorties. En raison de la richesse de l'espace des expressions mathématiques, la régression symbolique est habituellement un problème difficile à résoudre. Les méthodes basées sur l'apprentissage profond sont relativement nouvelles et représentent un domaine de recherche actif par rapport aux approches conventionnelles fondées sur des algorithmes d'évolution génétiques. Dans ce travail, nous présentons ≪symbolicGPT≫, un nouveau modèle de langage basé sur un transformateur pour la régression symbolique. Ce modèle exploite les avantages des modèles de langage probabilistes comme le GPT, y compris leurs fortes performance et flexibi-

bility. Through comprehensive experiments, we show that our model performs strongly compared to competing models with respect to the accuracy, running time, and data efficiency.

lité. Par l'entremise d'expériences approfondies, nous démontrons que notre modèle performe de façon solide par rapport à d'autres modèles concurrents en termes de précision, temps d'exécution et efficacité de données.

[14:30-15:00]

Yaoliang Yu (University of Waterloo) **Dockhorn Tim** (University of Waterloo) **Eyyüb Sari** (Huawei Noah's Ark Lab) **Mahdi Zolnouri** (Huawei Noah's Ark Lab) **Vahid Nia** (Huawei Noah's Ark Lab)

Demystifying and Generalizing BinaryConnect

Démystification et généralisation de la méthode BinaryConnect

BinaryConnect (BC) and its many variations have become the de facto standard for neural network quantization. However, our understanding of the inner workings of BC is still quite limited. We attempt to close this gap in four different aspects: (a) we show that existing quantization algorithms, including post-training quantization, are surprisingly similar to each other; (b) we argue for proximal maps as a natural family of quantizers that is both easy to design and analyze; (c) we refine the observation that BC is a special case of dual averaging, which itself is a special case of the generalized conditional gradient algorithm; (d) consequently, we propose ProxConnect (PC) as a generalization of BC and we prove its convergence properties by exploiting the established connections. We conduct experiments on CIFAR-10 and ImageNet, and verify that PC achieves competitive performance.

La méthode BinaryConnect et ses nombreuses variantes sont devenues la norme de facto pour quantifier les réseaux neuronaux. Cependant, notre compréhension du fonctionnement interne de BinaryConnect est encore assez limitée. Nous tentons de combler cette lacune sous quatre aspects différents : a) nous montrons que les algorithmes de quantification actuels et ceux de quantification post-formation se ressemblent étonnamment ; b) nous préconisons les cartes proximatives comme une famille naturelle de quantificateurs qui sont à la fois faciles à concevoir et à analyser ; c) nous affinons l'observation selon laquelle BinaryConnect est un cas particulier de moyennage double, qui est aussi un cas particulier de l'algorithme du gradient conditionnel généralisé ; d) par conséquent, nous proposons la méthode ProxConnect comme généralisation de la méthode BinaryConnect et nous démontrons ses propriétés de convergence en exploitant les connexions établies. Nous menons des expériences sur CIFAR-10 et ImageNet, puis nous démontrons que ProxConnect atteint des résultats concurrentiels.

Chair/Président: Yingwei (Paul) Peng

Organizer/Responsable: Joan X. Hu

Date: Thursday June 2 / jeudi 2 juin

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-14:00]

Trevor Thomson (Simon Fraser University) **X. Joan Hu** (Simon Fraser University) **Bohdan Nosyk** (Simon Fraser University)

Recent Advances in Modelling Time-to-Event Data with Internal Covariates

Avancées récentes dans la modélisation de données de durée de vie avec covariables internes

Previous studies indicate that retention on an opioid agonist treatment (OAT) can reduce the mortality risk of people with opioid use disorder. To account for the dynamic nature of OAT use, we considered an extended Cox proportional hazards model to account for treatment history through a time-dependent stratification variable. As the model makes explicit use of the time-dependent internal covariate, estimating survival probabilities based on the model is no longer feasible. With the aim of obtaining such probabilities, a Cox proportional hazards model is considered, where the internal covariate is replaced with a latent random variable, in which its distribution depends on the entire history of the covariate process. We modelled the internal covariate process over time and utilized the resulting estimates to infer parameters in the Cox proportional hazards model. The resulting procedure serves as an alternative to the conventional likelihood / partial likelihood based methods.

Des études précédentes indiquent que le maintien d'un traitement agoniste opioïdes (TAO) peut réduire le risque de mortalité de personnes souffrant d'un trouble lié à l'utilisation d'opioïdes. Afin de tenir compte de la nature dynamique de l'utilisation de TAO, nous examinons un modèle de risques proportionnels étendu pour prendre en considération l'historique de traitement par l'entremise d'une variable de stratification dépendante du temps. Vu que le modèle se sert explicitement de la covariable interne dépendante du temps, il est impossible d'estimer les probabilités de survie selon le modèle. Dans le but d'obtenir ces probabilités, nous examinons un modèle de risque proportionnel de Cox dans lequel la covariable interne est remplacée par une variable aléatoire latente, dont la distribution dépend de l'historique complet du processus de covariable. Nous avons modélisé le processus de covariable interne au fil du temps et avons utilisé les estimations obtenues pour inférer les paramètres dans le modèle de risques proportionnels de Cox. La procédure obtenue servira en guise de solution de rechange pour les méthodes conventionnelles basées sur la vraisemblance/vraisemblance partielle.

[14:00-14:30]

Leilei Zeng (University of Waterloo)

Response Dependent Sampling in Observational Cohort Studies

Échantillonnage dépendant de la réponse dans les études observationnelles de cohortes

Observational cohort studies of chronic disease involve the recruitment and follow-up of a sample of individuals with the goal of learning about the course of the disease, the effect of fixed and time-varying risk factors. Analysis of this information is often facilitated by using multistate models with intensity functions governing transition between disease states. Chronic disease studies often involve conditions for recruitment, for example

Les études observationnelles de cohortes sur les maladies chroniques nécessitent le recrutement et le suivi d'un échantillon d'individus dans le but de connaître l'évolution de la maladie et l'effet des facteurs de risque fixes et variables dans le temps. L'analyse de ces données est souvent facilitée par l'utilisation de modèles multi-états avec des fonctions d'intensité régissant la transition entre les états des maladies. Les études sur les maladies chroniques comportent souvent des critères de recrutement.

incident cohort involves individuals who are healthy at accrual, prevalent cohort samples individuals who have already developed the disease, and a length biased sampling includes individual who are alive at the time of recruitment. In this talk we discuss the impact of ignoring state-dependent sampling in multistate analysis and the ways of addressing the issue using auxiliary information. A longitudinal study of aging and cognition among religious sisters are used to illustrate the related methodologies.

Par exemple, une cohorte d'incidents comprend des individus en bonne santé au moment du recrutement, une cohorte prévalente comprend des individus qui ont déjà développé la maladie, et un échantillonnage biaisé par la durée comprend des individus qui sont en vie au moment du recrutement. Dans cette présentation, nous examinons les répercussions liées au fait de ne pas tenir compte de l'échantillonnage dépendant de l'état dans l'analyse multi-états et les moyens de résoudre ce problème à l'aide de données complémentaires. Nous utilisons une étude longitudinale du vieillissement et de la cognition chez des religieuses pour illustrer les méthodologies utilisées.

[14:30-15:00]

Rong Chen (Rutgers University)

Two Factor Models for High-Dimensional Tensor Time Series

Deux modèles factoriels pour séries chronologiques tensorielles en haute dimension

Large tensor data (multi-dimensional array) routinely appear nowadays in a wide range of applications, due to modern data collection capabilities. Often such observations are taken over time, forming tensor time series. In this talk we discuss two factor model approaches to the analysis of high-dimensional dynamic tensor time series. One approach assumes a Tucker-decomposition form, with a small core tensor being the time varying factor process, driving the co-moments of all individual time series that formed the tensor time series. The other approach assumes a CP-decomposition form, with a small number of independent univariate time series as the factors. The model in Tucker form is easy to estimate, but its factor process, in a tensor form, is difficult to model and interpret. The model in CP form is more difficult to estimate, but is easier to interpret and its factor process is easy to model. Estimation methods and their numerical and theoretical properties are presented. Applications are used to illustrate the models, their interpretations and comparison.

De grandes données tensorielles (réseaux multidimensionnels) apparaissent souvent de nos jours dans un large éventail d'applications, grâce aux capacités modernes de collecte de données. Souvent, ces observations sont prises au fil du temps, formant des séries chronologiques tensorielles. Dans cet exposé, nous discutons de deux approches de modèles factoriels pour l'analyse de séries chronologiques tensorielles dynamiques en haute dimension. Une approche suppose une forme de décomposition de Tucker, avec pour processus factoriel variant dans le temps un petit tenseur central qui dirige les co-moments de toutes les séries chronologiques individuelles qui forment la série chronologique tensorielle. L'autre approche suppose une forme de décomposition CP, avec pour facteurs un petit nombre de séries chronologiques indépendantes à une variable. Le modèle sous forme Tucker est facile à estimer, mais son processus factoriel, sous forme tensorielle, est difficile à modéliser et à interpréter. Le modèle sous forme CP est plus difficile à estimer, mais il est plus facile à interpréter et son processus factoriel est facile à modéliser. Nous présentons les méthodes d'estimation et leurs propriétés numériques et théoriques. Des applications sont utilisées pour illustrer les modèles, leurs interprétations et leur comparaison.

Chair/Président: Yaozhong Hu

Organizer/Responsable: Yaozhong Hu

Date: Thursday June 2 / jeudi 2 juin

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-14:00]

Xia Chen (University of Tennessee)

Necessary and sufficient condition for the solvability of the hyperbolic Anderson models with Gaussian noise that is fractional in times

Condition nécessaire et suffisante pour la solvabilité des modèles hyperboliques d'Anderson avec bruit gaussien fractionné en temps

In the Ito-Skorohkod regime, the solution of the hyperbolic Anderson models is formally written in the form of Ito-Wiener chaos expansion and the uniqueness/existence of the system is equivalent to the convergence of the expansion in L_2. In this report, we find the condition that is necessary and sufficient for such convergence in the setting where the system is driven by a Gaussian noise that is fractional in times. The work is based on part of the collaborative project with Deya, A., Jian Song and Samy Tindal.

Dans le régime d'Ito-Skorohkod, la solution des modèles hyperboliques d'Anderson s'écrit formellement sous la forme d'une expansion de chaos d'Ito-Wiener et l'unicité/existence du système est équivalente à la convergence de l'expansion dans L_2. Dans ce rapport, nous trouvons la condition nécessaire et suffisante pour une telle convergence dans la situation où le système est piloté par un bruit gaussien fractionnaire en temps. Ce travail est basé sur une partie du projet de collaboration avec Deya, A., Jian Song et Samy Tindal.

[14:00-14:30]

Jian Song (Shandong University) **Guanglin Rang** (Wuhan University)

The Scaling Limit of a Long-range Random Walk in Correlated Random Medium

La limite d'échelle d'une promenade aléatoire de longue portée dans un milieu aléatoire corrélé

The scaling limit of a long-range random walk in random medium with correlations in both time and d-dimensional space is investigated. We show that the rescaled partition function converges weakly to the Stratonovich solution of some stochastic fractional heat equation with multiplicative Gaussian noise, where the noise is fractional with Hurst parameter vector determined by the algebraic exponent of the medium correlation. This talk is based on a joint work with Guanglin Rang.

Nous étudions la limite d'échelle d'une promenade aléatoire de longue portée dans un milieu aléatoire avec corrélation dans le temps et l'espace à d dimensions. Nous démontrons que la fonction de partition rééchelonnée converge faiblement vers la solution Stratonovich de certaines équations de chaleur fractionnaire avec bruit gaussien multiplicatif, dont le bruit est fractionnaire avec le vecteur de paramètre Hurst déterminé par l'exposant algébrique de la corrélation moyenne. Cet exposé est basé sur un travail conjoint avec Guanglin Rang.

[14:30-15:00]

Samy Tindel (Purdue University)

A coupling between Sinai's random walk and Brox diffusion

Couplage entre la marche aléatoire de Sinai et la diffusion de Brox

Sinai's random walk is a standard model of 1-dimensional random walk in random environment. Brox diffusion is

La marche aléatoire de Sinai est un modèle standard de marche aléatoire unidimensionnelle en environnement aléatoire. La diffu-

its continuous counterpart, that is a Brownian diffusion in a Brownian environment. The convergence in law of a properly rescaled version of Sinai's walk to Brox diffusion has been established 20 years ago. In this talk, I will explain a strategy which yields the convergence of Sinai's walk to Brox diffusion thanks to an explicit coupling. This method, based on rough paths techniques, opens the way to rates of convergence in this demanding context. Notice that I'll try to give a maximum of background about the objects I'm manipulating, and will keep technical considerations to a minimum.

sion de Brox est son équivalent en continu, c'est-à-dire un mouvement brownien en environnement aléatoire brownien. Après renormalisation, il est connu que la marche de Sinai converge vers la diffusion de Brox. Dans mon exposé, j'expliquerai une stratégie montrant la convergence de la marche de Sinai grâce à un couplage explicite. La méthode est basée sur des techniques de trajectoires rugueuses. Elle donne des vitesses de convergence dans ce contexte exigeant. J'essaierai d'introduire soigneusement les objets que je manipule. J'essaierai aussi de réduire les considérations techniques à leur strict minimum.

Chair/Président: Liqun Diao

Date: Thursday June 2 / jeudi 2 juin

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-13:45]

Dylan Z. Spicker (University of Waterloo) **Michael Wallace** (University of Waterloo) **Grace Y. Yi** (University of Western Ontario)

Nonparametric Simulation Extrapolation for Measurement Error Models

Extrapolation par simulation non paramétrique pour des modèles d'erreur de mesure

Measurement error is a widespread issue which can render estimators inconsistent and reduce power in inference. Simulation extrapolation (SIMEX) is a comparatively straightforward procedure to correct for the effects of measurement error which can be applied in a wide variety of models. SIMEX typically assumes that errors are normally distributed, but this assumption is often violated in practice. Alternative parametric forms of SIMEX have been proposed using relevant distributional assumptions for particular settings. We propose a non-parametric extension to SIMEX, which uses a similar procedure to the standard SIMEX, but is generally applicable regardless of the error distribution. We demonstrate the utility of the proposed correction through theory, simulation, and by an application to data from the Korean Longitudinal Study of Ageing.

L'erreur de mesure est un problème très répandu qui peut rendre les estimateurs non consistants et réduire la puissance en inférence. L'extrapolation par simulation (SIMEX) est une procédure relativement simple permettant de corriger les effets de l'erreur de mesure qui peut être appliquée à une grande variété de modèles. SIMEX suppose généralement que les erreurs suivent une loi normale, mais cette hypothèse n'est pas toujours respectée en pratique. Des formes paramétriques alternatives de SIMEX ont été proposées en utilisant des hypothèses de distribution pertinentes pour des contextes particuliers. Nous proposons une extension non paramétrique de SIMEX, qui utilise une procédure similaire au SIMEX standard, mais qui est généralement applicable quelle que soit la distribution des erreurs. Nous démontrons l'utilité de la correction proposée par la théorie, la simulation et une application aux données de l'étude longitudinale coréenne sur le vieillissement.

[13:45-14:00]

Jingyu Cui (Western University) **Grace Y. Yi** (Western University)

Multivariate Regression Model with Measurement Error

Modèle de régression multivarié avec erreur de mesure

Multivariate regression models are useful for learning the relationship between multiple responses and multiple predictor variables. Inference under such models is however invalidated if the variables are subject to measurement error. In this talk, we discuss bias analyses of the least squares method under multivariate measurement error models. We propose consistent estimators for the response model parameters with measurement error effects accommodated. Numerical studies are carried out to assess the finite sample performance of the proposed method, in contrast to that of the naive method which ignores the measurement error effects.

Les modèles de régression multivariés sont pratiques pour connaître le lien entre des réponses multiples et des variables de prédicteurs multiples. L'inférence selon ce type de modèle est toutefois invalide si les variables sont sujettes à des erreurs de mesure. Lors de cet exposé, nous aborderons les analyses de biais de la méthode des moindres carrés selon des modèles d'erreur de mesure multivariés. Nous proposons des estimateurs convergents pour les paramètres de modèle de réponse en tenant compte des effets d'erreurs de mesure. Des études numériques sont entreprises dans le but d'évaluer la performance sous échantillons de taille finie de la méthode proposée, par rapport à celle de la méthode naïve qui ne tient pas compte des effets des erreurs de mesure.

[14:00-14:15]

Alexandra S. Bushby (University of Toronto) **Eleanor M. Pullenayegum** (The Hospital for Sick Children)
Measurement Error in Longitudinal Data with Irregular Observation

Erreur de mesure des données longitudinales avec des observations irrégulières

Biased estimates of the associations of covariates with outcomes in longitudinal data can arise for several reasons, for example, irregular observation times and the presence of measurement error. In longitudinal data, irregularity is often informative: If a patient is feeling unwell, their visit intensity is likely to be higher. To account for the informative visit process, we use an inverse-intensity weighted generalized estimating equation. Additionally, measurement error occurs when some measured covariates are recorded incorrectly due to inaccurate reporting, reading errors, etc. We extend a currently available measurement error method, simulation-extrapolation, to longitudinal data with irregular observation. We illustrate performance through simulation and apply it to a data set of treatment of major depressive disorder.

Des estimations biaisées des associations entre les covariables et les résultats des données longitudinales peuvent survenir pour plusieurs raisons, comme des temps d'observation irréguliers et la présence d'erreurs de mesure. Dans les données longitudinales, l'irrégularité est souvent informative : si un patient ne se sent pas bien, la fréquence de ses visites sera probablement plus élevée. Pour tenir compte de l'information contenu dans le processus de visite, nous utilisons une équation d'estimation généralisée pondérée par une intensité inverse. En outre, une erreur de mesure se produit lorsque certaines covariables mesurées sont enregistrées de manière incorrecte en raison de déclarations inexactes, d'erreurs de lecture, etc. Nous étendons une méthode d'erreur de mesure existante (simulation-extrapolation) à des données longitudinales dans le cadre de temps d'observation irréguliers. Nous illustrons la performance de la méthode par une simulation et l'appliquons à un ensemble de données sur le traitement des troubles dépressifs majeurs.

[14:15-14:30]

Melina Ribaud (HEC Montréal) **Aurélie Labbe** (HEC Montreal) **Karim Oualkacha** (Université du Québec à Montréal)
Imputation in genetic methylation studies: A linear model of coregionalization (LMC) with informative covariates

Problèmes d'imputation dans les études génétiques de méthylation : un modèle de corrégionalisation linéaire (LMC) avec covariables.

DNA methylation is a process that modifies the CpG sites of DNA by the addition of a methyl group. This phenomenon is necessary for a healthy functioning body. Methylation levels are quantified at every genomic CpG site, but they are subject to missing values. Our goal is to impute the level of methylation on the missing sites; it is a high dimensional imputation problem with covariates. In this presentation, we propose a method to predict the missing methylation levels. This method catches correlation structures among the methylation levels across genome sites and across samples. The regression function linking the methylation level to the covariates is modeled through a linear combination of the covariates together with latent factors. We assume the covariates' and latent factors' effects to be Gaussian random processes. We implement a fast two-step likelihood-based algorithm to estimate the model parameters. We predict missing values via equations conditional on the observed data.

La méthylation est un processus qui modifie les sites CpG de l'ADN par l'addition d'un groupe méthyle. Ce phénomène est nécessaire au fonctionnement du corps. La méthylation est mesurée sur tous les sites, mais sujette aux valeurs manquantes. L'objectif est d'imputer le niveau de méthylation sur les sites manquants ; c'est un problème d'imputation en grande dimension avec covariables. Dans cette présentation, nous proposons une méthode pour prédire les niveaux de méthylation manquants à partir de ceux observés et des covariables. Cette méthode capture les structures de corrélation du niveau de méthylation entre les sites et les échantillons. La fonction de régression reliant le niveau de méthylation aux covariables est modélisée par une combinaison linéaire des facteurs observés et latents (LMC). Nous supposons que les effets des facteurs sont des processus Gaussiens. Les prédictions pour les données manquantes sont obtenues par des équations conditionnelles aux données observées.

[14:30-14:45]

Mei Dong (University of Toronto) **Aya A. Mitani** (University of Toronto)

Multiple imputation methods for missing multilevel ordinal outcomes

Méthodes d'imputation multiple de résultats manquants ordinaux à plusieurs niveaux

Multiple imputation (MI) is an established technique to handle missing data in observational studies. Joint modeling (JM) and fully conditional specification (FCS) are commonly used methods for imputing multilevel data. However, MI methods for multilevel ordinal outcome variables have not been well studied, especially when there is informative cluster size (ICS). The purpose of this study is to describe different imputation and marginal analysis strategies for the clustered ordinal outcome when ICS exists. We compare five different imputation methods: complete case analysis, FCS, FCS+N (include cluster size when imputation), JM, and JM+N; and two analysis methods: generalized estimating equations (GEE) and cluster-weighted GEE. The simulation results show that FCS yields less biased estimates than JM, and including cluster size in the imputation model can significantly improve the accuracy when ICS exists. We further applied those methods to a real dental study.

L'imputation multiple est une technique reconnue pour traiter les données manquantes dans les études d'observation. La modélisation conjointe (JM) et la spécification entièrement conditionnelle (FCS) sont des méthodes souvent utilisées pour l'imputation de données à plusieurs niveaux. Cependant, les méthodes d'imputation multiple pour les variables de résultats ordinaux à plusieurs niveaux n'ont pas été bien étudiées, notamment en cas de taille de groupe informative (ICS). L'objectif de cette étude est de décrire différentes stratégies d'imputation et d'analyse marginale pour les résultats ordinaux en grappes lorsqu'il existe une taille de groupe informative. Nous comparons cinq méthodes d'imputation différentes (étude de cas complète, FCS, FCS+N, y compris la taille de la grappe lors de l'imputation, JM et JM+N) ainsi que deux méthodes d'analyse (équations d'estimation généralisées et estimation généralisées pondérées par groupe). Les résultats de la simulation montrent que la spécification entièrement conditionnelle produit des estimations moins biaisées que la modélisation conjointe, et que l'inclusion de la taille du groupe dans le modèle d'imputation peut considérablement améliorer la précision en cas de taille de groupe informative. Ensuite, nous appliquons ces méthodes à une étude dentaire réelle.

[14:45-15:00]

Jinhui Ma (McMaster University) **Parminder Raina** (McMaster University) **Lauren Griffith** (McMaster University) **Mylinh Duong** (McMaster University) **Alexandra Mayhew** (McMaster University) **Carol Bassim** (McMaster University) **Chris Verschoor** (Health Sciences North Research Institute) **Lehana Thabane** (McMaster University) **Hon-Yiu So** (Oakland University)

Imputation of Missing Spirometry Data in Population-based Studies

Imputation de données de spirométrie manquantes en études sur la population

Spirometry is the most reliable and objective measure of lung capacity. It has been increasing used in population-based studies for diagnosis and estimating the prevalence and incidence of chronic obstructive pulmonary disease. The rate of missing spirometry data is high especially in aging studies due to contraindication or poor quality of spirometry test. Two spirometry measures – forced expiratory volume in one second (FEV1) and forced vital capacity (FVC) – are collected from one spirometry test and the ratio of FEV1 to FVC together with FEV1 and FVC are used to identify obstructive or restrictive lung defects. Traditional imputation approach is inappropriate to impute missing FEV1and FVC data since they are correlated differently in healthy individuals and those with lung impairment. To overcome this challenge, we proposed to involve machine learning techniques in the multiple imputation process

La spirométrie est la mesure de capacité pulmonaire la plus fiable et objective. Elle est de plus en plus employée dans des études sur la population servant à diagnostiquer et estimer la prévalence et l'incidence de bronchopneumopathie chronique obstructive. Le taux de données de spirométrie manquante est élevé tout particulièrement dans les études de vieillissement à cause de contradictions ou de la moindre qualité de l'examen spirométrique. Deux mesures spirométriques, le volume expiratoire maximal en 1 seconde (FEV1) et la capacité vitale forcée (CVF) sont recueillis à partir d'un examen spirométrique, puis on utilise le rapport FEV1-FVC conjointement au FEV1 et à la CVF pour repérer l'anomalie pulmonaire obstructive ou restrictive. L'approche d'imputation traditionnelle est inadéquate pour imputer les données du FEV1 et de la CVF vu qu'ils sont corrélés différemment chez les individus en santé et chez ceux ayant une déficience pulmonaire. Afin de surmonter ce défi, nous proposons d'intégrer des techniques d'apprentissage automatique dans les processus d'imputation multiple

and evaluate its performance using data from the Canadian Longitudinal Study on Aging.

et évaluons sa performance à l'aide de données tirées de l'étude longitudinale canadienne sur le vieillissement.

Chair/Président: Qingrun Zhang

Date: Thursday June 2 / jeudi 2 juin

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-13:45]

Patrick Fournier (Université du Québec à Montréal)

Accounting for Epistasis in PRSs Through the Coalescent

Prise en compte de l'épistasie dans les SRP grâce au coalescent

Polymeric risk scores are nowadays heavily used in medicine as a useful prognostic tool. Despite their success in the prediction of various conditions, many challenges still need to be addressed. Among those is accounting for epistasis. In its most common form, a PRS is a linear combination of genetic markers. This make accounting for gene-gene interaction a combinatorially challenging problem. Moreover,taking a step back, the presence of cryptic epistasis in the data on which the so called weights of the PRS are computed could create false associations or hinder real ones. While it is possible to use the coalescent as the basis of such a score, modeling the recombination process is key for accounting for non-linearity in the training data. Since this is in itself a task known to be computationally challenging, we approximate this process using higher order Markov chains.

Les scores de risque polymérique sont aujourd'hui largement utilisés en médecine comme outil pronostique utile. Malgré leur succès dans la prédiction de diverses conditions, de nombreux défis restent à relever, dont la prise en compte de l'épistasie. Dans sa forme la plus courante, un SRP est une combinaison linéaire de marqueurs génétiques. Cela fait de la prise en compte de l'interaction gène-gène un problème combinatoire difficile. De plus, en prenant un peu de recul, la présence d'une épistasie cryptique dans les données sur lesquelles les soi-disant poids du SRP sont calculés pourrait créer de fausses associations ou cacher les vraies. Bien qu'il soit possible d'utiliser le coalescent comme base d'un tel score, la modélisation du processus de recombinaison est essentielle pour tenir compte de la non-linéarité des données d'apprentissage. Comme il s'agit en soi d'une tâche connue pour être difficile sur le plan informatique, nous approximons ce processus à l'aide de chaînes de Markov d'ordre supérieur.

[13:45-14:00]

Olga Vishnyakova (Simon Fraser University) **Angela Brooks-Wilson** (Simon Fraser University, BC Cancer) **Lloyd T. Elliott** (Simon Fraser University)

Analysis of Homeostasis in Health

Analyse de l'homéostasie dans la santé

Homeostasis allows the maintenance of stable physiological conditions and body chemistry. We are examining the biological foundation of health through the lens of homeostasis. Our hypothesis is that at a given age, healthy individuals have values closer to an ideal value, or 'Sweet Spot' for certain phenotypes. I will perform a systematic analysis of data from the Canadian Longitudinal Study on Aging to identify phenotypes under homeostatic control and identify associated genes. To examine phenotypes for variance effects associated with health, I will use a robust Brown-Forsythe Levene-type procedure to test for a difference in variance pair-

L'homéostasie permet de garder des conditions physiologiques et une chimie du corps stables. Nous examinons les fondements biologiques de la santé sous l'angle de l'homéostasie. Selon notre hypothèse, à un âge donné, les individus en bonne santé ont des valeurs plus proches d'une valeur idéale (point idéal) pour certains phénotypes. J'effectuerai une analyse systématique des données de l'Étude longitudinale canadienne sur le vieillissement pour déterminer les phénotypes faisant l'objet d'un contrôle homéostatique et pour répertorier les gènes associés. Afin d'examiner les phénotypes des effets de variance associés à la santé, j'utiliserai une méthode robuste de type Brown-Forsythe-Levene pour effectuer un test de différence de variance par paire entre les

wise between the most healthy and least healthy instrument levels. After, I will characterize the relationship between each instrument and phenotype magnitude to identify the optimal value, by examination of inflection points, followed by GWAS on the absolute deviation from the sweet spot to explore the genetic architecture of homeostasis in health.

groupes d'instruments les plus sains et les moins sains. Ensuite, je caractériserai la relation entre chaque instrument et l'intensité du phénotype pour déterminer la valeur optimale, par l'examen des points d'inflexion et par une étude d'association à l'échelle du génome sur l'écart absolu du point d'inflexion pour explorer l'architecture génétique de l'homéostasie de la santé.

[14:00-14:15]

Quan Long (University of Calgary)

An Expression-directed Linear Mixed Model (edLMM) Discovering Low-effect Genetic Variant

Modèle mixte linéaire avec expression dirigée (edLMM) pour découvrir les variants génétiques à faible effet

Detecting low-effect genetic variants associated with diseases using a small or moderate sample is difficult, hindering the downstream efforts of explaining the heritability or forming accurate polygenic risk scores (PRS). In this work, by utilizing the functional weights learned from transcriptome data, we formed an alternative approach to estimate the polygenic term in a traditional linear mixed model (LMM), which can estimate the genetic background more accurately. As a result, our tool, namely expression-directed linear mixed model (edLMM), enables the discovery of subtle signals of low-effect variants using moderate samples. By applying it to cohorts of around a few thousand individuals with phenotype of either binary or quantitative traits, we demonstrated its power gain at the low-effect spectrum. Aggregately, the additional low-effect variants detected by edLMM substantially improved estimation of missing heritability and formed accurate PRS.

L'utilisation d'un échantillon de taille petite ou modérée rend difficile la détection de variants génétiques à faible effet associés à des maladies, ce qui les freine les efforts en aval pour expliquer l'héritabilité ou pour former des scores de risque polygénique (PRS) exacts. Dans le cadre de ce travail, en utilisant les poids fonctionnels appris des données du transcriptome, nous élaborons une approche alternative pour l'estimation du terme polygénique dans un modèle mixte linéaire (LMM) traditionnel, une approche qui peut estimer les antécédents génétiques avec une plus grande exactitude. Par conséquent, notre outil, soit le modèle mixte linéaire avec expression dirigée (edLMM), permet la découverte de signaux subtils de variants à faible effet en utilisant des échantillons de taille modérée. En l'appliquant à des cohortes de quelque milliers de sujets avec phénotype de traits binaires ou quantitatifs, nous avons montré son gain de puissance dans le spectre à faible effet. Globalement, les variants additionnels à faible effet détectés par edLMM ont amélioré sensiblement l'estimation de l'héritabilité manquante et formé des PRS exacts.

[14:15-14:30]

Guan Wang (University of Toronto: Dalla Lana School of Public Health)

Two-Phase Design for Regional Genetic Sequencing Using Polygenic Risk Scores

Plan en deux phases pour un séquençage génétique régional utilisant des scores de risque polygénique

Due to the high cost of DNA sequencing for large-scale data, I propose a two-phase design using polygenic risk scores (PRS) to inform selection of individuals in phase 1, followed by regional sequencing in a selected subsample in phase 2. Residual dependent sampling (RDS) design is implemented by regressing the phenotype of interest on the PRS and selecting individuals with extreme residuals as the phase 2 subsample. Efficient analysis can be carried out under semi-parametric modelling by the EM algorithm. A fine-mapping application in a genome-wide association study (GWAS) of triglyceride levels in 4,504 individuals from the Northern Finland Birth Cohort of 1966 shows the proposed method can

En raison du coût élevé du séquençage de l'ADN pour les données à grande échelle, je propose un plan en deux phases utilisant les scores de risque polygénique (PRS) pour informer la sélection des individus en phase 1, suivie du séquençage régional d'un sous-échantillon sélectionné en phase 2. Je crée un plan d'échantillonnage dépendant des résidus (RDS) en régressant le phénotype d'intérêt sur le PRS et en sélectionnant les individus présentant des résidus extrêmes comme sous-échantillon de la phase 2. Une analyse efficace peut être effectuée via une modélisation semi-paramétrique par l'algorithme EM. Une application de cartographie fine à une étude d'association pangénomique (GWAS) des niveaux de triglycérides de 4 504 individus de la Northern Finland Birth Cohort de 1966 montre que la méthode

reduce sequencing costs in post-GWAS analyses while maintaining statistical performance. Simulation studies show that the proposed RDS design gives more precise estimation than simple random sampling, with adequate type one error control, while performing more similarly to the complete sample.

proposée permet de réduire les coûts de séquençage des analyses post-GWAS tout en en maintenant la performance statistique. Des études de simulation montrent que le plan RDS proposé donne une estimation plus précise que l'échantillonnage aléatoire simple, avec un contrôle adéquat de l'erreur de type I, tout en ayant des performances plus similaires à celles de l'échantillon complet.

[14:30-14:45]

Ting Zhang (McGill University) **Jerome Dockes** (McGill University) **Nikhil Bhagwat** (McGill University) **Clara Moreau** (Pasteur Institute) **Celia Greenwood** (McGill University) **Jean-Baptiste Poline** (McGill University)

Kernel Selection for Linear Mixed Effect model on Estimating Variance Explained

Sélection de noyaux pour modèle linéaire à effets mixtes sur l'estimation de la variance expliquée

In statistical modelling, the proportion of variance explained (VE) is a key measure of goodness-of-fit. Recently, to estimate the phenotypic variance explained by brain morphology, neuroscience researchers have used linear mixed effect models (LME) with non-brain-related fixed covariates, and a random component based on a predetermined correlation structure that measures the between subject anatomical similarity (ASM). In practice, ASM can be estimated by different kernel functions. The estimated VE is affected by the choice of ASM. A very popular model selection criterion, conditional AIC, always favours the kernel that yields a higher estimated VE. This preference can consequently result in serious overestimation of VE. So far, the within-sample model selection criterion does not account for kernel complexity, and out-of-sample cross validation cannot reveal the bias as ASM is data-set dependent. We are seeking a method for kernel selection in LME.

En modélisation statistique, la proportion de variance expliquée (VE) est une mesure clé de la qualité de l'ajustement. Récemment, pour estimer la variance phénotypique expliquée par la morphologie cérébrale, des chercheurs en neurosciences ont utilisé des modèles linéaires à effets mixtes (LME) avec des covariables fixes non liées au cerveau et une composante aléatoire basée sur une structure de corrélation prédéterminée qui mesure la similarité anatomique entre les sujets (ASM). En pratique, l'ASM peut être estimée par différentes fonctions à noyau. La VE estimée est affectée par le choix de l'ASM. Un critère de sélection de modèle très populaire, l'AIC conditionnel, favorise toujours le noyau qui produit la VE estimée la plus élevée. Cette préférence peut par conséquent entraîner une surestimation importante de la VE. Jusqu'à présent, le critère de sélection du modèle intra-échantillon n'a pas tenu compte de la complexité du noyau, et la validation croisée hors échantillon ne peut pas révéler le biais car l'ASM dépend de l'ensemble de données. Nous recherchons une méthode de sélection du noyau pour les modèles LME.

Chair/Président: Sharandeep Singh Pandher

Date: Thursday June 2 / jeudi 2 juin

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-13:45]

Thai-Son Tang (University of Toronto) **Zhihui (Amy) Liu** (Princess Margaret Cancer Centre, University Health Network; Dalla Lana School of Public Health, University of Toronto) **Olli Saarela** (Dalla Lana School of Public Health, University of Toronto)

A marginal structural model for normal tissue complication probability

Modèle structurel marginal pour la probabilité de complication des tissus normaux

The goal of radiation therapy is delivery of prescribed radiation dose to target organ while minimizing exposure to surrounding tissue to avoid normal tissue complications. Dose-volume histograms (DVHs) characterize the functional relationship between radiation dose and organ volume and are used in treatment planning. Normal tissue complication probability (NTCP) modelling has centered around making patient-level predictions using DVHs, but few have considered evaluating comparative effectiveness of treatment plans in a causal modeling framework. We formulate causal estimands for functional DVH exposures and propose estimators based on marginal structural models that parametrize bivariable monotonicity between dose, volume, and NTCP to reflect the natural biological mechanisms between these quantities. The properties of these estimators are studied through simulations, along with an illustration on anal canal cancer patients in Ontario.

L'objectif de la radiothérapie est de délivrer la dose de radiation prescrite à l'organe cible tout en réduisant au minimum l'exposition des tissus voisins afin d'éviter des complications sur les tissus normaux. Les histogrammes dose-volume caractérisent la relation fonctionnelle entre la dose de radiation et le volume de l'organe, et sont utilisés dans la planification du traitement. La modélisation de la probabilité de complication des tissus normaux a été axée sur la réalisation de prédictions à l'échelle du patient à l'aide d'histogrammes dose-volume, mais peu d'entre elles ont pris en compte l'évaluation de l'efficacité comparative des plans de traitement dans un cadre de modélisation causale. Nous créons des estimateurs causaux pour les expositions fonctionnelles d'histogrammes dose-volume et proposons des estimateurs fondés sur des modèles structurels marginaux qui paramètrent la monotonicité à variable double entre la dose, le volume et la probabilité de complication des tissus normaux afin de refléter les mécanismes biologiques naturels entre ces quantités. Nous examinons les propriétés de ces estimateurs à l'aide de simulations, ainsi que d'une illustration sur des patients atteints de cancer du canal anal en Ontario.

[13:45-14:00]

Haixu Alex Wang (Simon Fraser University) **Jiguo Cao** (Simon Fraser University)

Functional Nonlinear Learning

Apprentissage non linéaire fonctionnel

Using representations of functional data can be more convenient and beneficial in subsequent statistical models than direct observations. These representations, in a lower-dimensional space, extract and compress information from individual curves. The existing representation learning approaches in functional data analysis usually use linear mapping in parallel to those from multivariate analysis, e.g., functional principal component analysis (FPCA). However, functions, as infinite-

Les représentations de données fonctionnelles peuvent être plus commodes et bénéfiques que les observations directes dans les modèles statistiques subséquents. Dans un espace de plus faible dimensionnalité, ces représentations extraient et compriment l'information de courbes individuelles. Les approches existantes d'apprentissage des représentations dans l'analyse des données fonctionnelles utilisent généralement le mappage linéaire parallèlement aux approches d'analyse multivariée, par exemple, l'analyse en composantes principales fonctionnelles (ACPF).

dimensional objects, sometimes have nonlinear structures that cannot be uncovered by linear mapping. Linear methods will be more overwhelmed given multivariate functional data. For that matter, this paper proposes a functional nonlinear learning (FunNoL) method to sufficiently represent multivariate functional data in a lower-dimensional feature space. Furthermore, we merge a classification model for enriching the ability of representations in predicting curve labels.

Comme objets de dimension infinie, les fonctions ont parfois des structures non linéaires qui ne peuvent pas être non couvertes par mappage linéaire. Les méthodes linéaires seront encore plus dépassées compte tenu de données fonctionnelles multivariées. Pour cette raison, notre exposé propose une méthode d'apprentissage non linéaire fonctionnel (en anglais, FunNoL) pour une représentation suffisante de données fonctionnelles multivariées dans un espace de fonctionnalités de dimensionnalité plus faible. De plus, nous y combinons un modèle de classification pour enrichir la capacité des représentations de prédire les étiquettes de courbe.

[14:00-14:15]

Shivani Bhardwaj (University of Manitoba) **Yuliya V. Martsynyuk** (University of Manitoba)

Finite-sample properties and applicability of functional CLT based confidence intervals for a population mean

Propriétés d'échantillon fini et applicabilité d'intervalles de confiance fonctionnels basés sur le théorème central limite d'une moyenne de population

We consider a Student process that is based on independent copies of a random variable X and has trajectories in the function space D[0,1]. If X is in the domain of attraction of the normal law, a weighted version of the Student process is known to follow a functional central limit theorem (FCLT). Accordingly, appropriate functionals of such a process converge in distribution to the same functionals of a weighted Wiener process. We use such a convergence for several functionals and derive asymptotic confidence intervals (CI) for the mean of X. Based on our investigation of the finite-sample coverage probabilities and expected lengths of the obtained CI's for different types of distributions of X, we suggest when these FCLT based CI's may be appealing alternatives to an asymptotic CI for the mean of X that is derived from the asymptotic normality of the Student t-statistic.

Nous analysons un processus de Student qui repose sur des copies indépendantes d'une variable aléatoire X et qui a des trajectoires dans l'espace des fonctions D[0,1]. Si X est dans le domaine d'attraction de la loi normale, on sait qu'une version pondérée du processus de Student suivra un théorème central limite fonctionnel. Par conséquent, les fonctionnelles correspondantes d'un tel processus convergent en distribution vers les mêmes fonctionnelles d'un processus de Wiener pondéré. Nous utilisons une telle convergence pour plusieurs fonctionnelles et obtenons des intervalles de confiance asymptotiques pour la moyenne de X. Nous étudions les probabilités de couverture d'échantillons finis et les longueurs attendues des intervalles de confiance obtenus pour différents types de distributions de X. À partir de cette étude, nous montrons à quel moment ces intervalles de confiance reposant sur le théorème de la limite centrale peuvent constituer des solutions de rechange intéressantes à un intervalle de confiance asymptotique pour la moyenne de X, obtenu à partir de la normalité asymptotique de la statistique t de Student.

[14:15-14:30]

Boyi Hu (Simon Fraser University) **Hua Liu** (Xi'an Jiaotong University) **Jinhong You** (Shanghai University of Finance and Economics) **Jiguo Cao** (Simon Fraser University)

Convolution Smoothed Functional Linear Quantile Regression with Locally Sparse Adaptation

Régression quantile linéaire fonctionnelle lissée par convolution avec adaptation localement éparse

Local sparseness of the estimated slope function in a functional quantile regression model is crucial in certain functional data context. In this paper, we propose a smoothed functional linear quantile regression model with regularizations. Using this framework, the estimator for the slope function is smooth and locally sparse. Our estimator can identify the region where the func-

L'éparsité locale de la fonction de pente estimée dans un modèle de régression quantile fonctionnelle est cruciale dans certains contextes de données fonctionnelles. Dans cet article, nous proposons un modèle de régression quantile linéaire fonctionnelle lissée avec régularisations. Dans ce cadre, l'estimateur de la fonction de pente est lisse et localement clairsemé. Notre estimateur peut identifier la région où les covariables fonctionnelles n'ont aucune

tional covariates have no apparent relationship with the response. We also develop a fast algorithm for the estimation procedure. Simulation studies show that our estimator has good performance. We then demonstrate the practical applications of our method on two real-world datasets.

relation apparente avec la réponse. Nous développons également un algorithme rapide pour la procédure d'estimation. Des études de simulation montrent que notre estimateur a de bonnes performances. Nous démontrons ensuite l'application pratique de notre méthode sur deux ensembles de données du monde réel.

[14:30-14:45]

Chi-Kuang Yeh (University of Waterloo) **Gregory Rice** (University of Waterloo) **Joel A. Dubin** (University of Waterloo)

Projection Based Model Validation and Identification Methods for Functional Time Series

Méthodes de validation et d'identification de modèles basés sur la projection pour séries chronologiques fonctionnelles

Measuring the serial dependence across time is critical in model identification and diagnosis in time series (TS) analysis. In classic TS analysis, the autocorrelation function is perhaps the most widely used method to examine the temporal relationship of the scalar or vector-valued observations. In functional TS (FTS), which refers to TS of functional data, their dependence is best summarised by an autocovariance operator. Evaluating the size and information contained in such an object can be difficult. Existing methods are relatively constrained and unable to capture certain characteristics contained in the FTS objects, such as the "direction" of dependence. We develop a new method to address this problem by projecting lagged pairs unit sphere and computing the angle between them, which we refer to as dynamic autocorrelation. We establish the asymptotic properties of the empirical dynamical autocorrelation, and we study its use in an application to European electricity data.

La mesure de la dépendance sérielle dans le temps est essentielle pour l'identification de modèles et le diagnostic dans l'analyse des séries chronologiques (SC). Dans l'analyse SC classique, la fonction d'autocorrélation est peut-être la méthode la plus utilisée pour examiner la relation temporelle des observations à valeur scalaire ou vectorielle. Dans l'analyse SC fonctionnelle (SCF), qui fait référence à l'analyse SC de données fonctionnelles, leur dépendance est mieux résumée par un opérateur d'autocovariance. L'évaluation de la taille et des informations contenues dans un tel objet peut être difficile. Les méthodes existantes sont relativement limitées et ne permettent pas de capturer certaines caractéristiques contenues dans les objets SCF, telles que la « direction » de la dépendance. Nous développons une nouvelle méthode pour résoudre ce problème en projetant la sphère unitaire de paires décalées et en calculant l'angle entre elles, ce que nous appelons l'autocorrélation dynamique. Nous établissons les propriétés asymptotiques de l'autocorrélation dynamique empirique et nous étudions son utilisation dans une application aux données européennes sur l'électricité.

[14:45-15:00]

Skye Paphora Griffith (Queen's University)

Transfer Function Estimates and their Phase Distributions under the Multitaper Method

Estimations de fonctions de transfert et de leurs distributions de phase selon la méthode multitaper

In time series analysis, regression in the frequency domain relates Fourier transforms of the (tapered) response and predictor time series via a transfer function, a complex-valued function of frequency. Multitaper spectrum estimation has been shown to minimize spectral leakage (broad-band bias) while providing flexibility in terms of bandwidth selection (variance). Similar techniques can be used to obtain multitaper transfer function estimates (MTFEs). In this paper, we consider the MTFE phase distribution. For models whose underlying noise is Gaussian and weakly stationary, the MTFE phase is derived exactly, and for more general models, derived approximately. Estimation is done us-

En analyse de séries chronologiques, la régression dans le domaine fréquentiel met en relation les transformées de Fourier des séries chronologiques de la réponse (taper) et du prédicteur via une fonction de transfert, fonction à valeur complexe de la fréquence. Il a été démontré que l'estimation du spectre multitaper minimise les fuites spectrales (biais à large bande) tout en offrant une certaine flexibilité en termes de sélection de la bande passante (variance). Des techniques similaires peuvent être utilisées pour obtenir des estimations de fonctions de transfert multitaper (MTFE). Dans cet article, nous considérons la distribution de phase MTFE. Pour les modèles dont le bruit sous-jacent est gaussien et faiblement stationnaire, la phase MTFE est dérivée exactement, et pour les modèles plus généraux, dérivée approximativement. Nous ef-

ing plug-in estimates obtained from multitaper estimates of the autocovariance function of the response tim e series.

fectuons l'estimation à l'aide d'estimations enfichables obtenues à partir d'estimations multitaper de la fonction d'auto-covariance de la série chronologique de la réponse.

Chair/Président: Jiguo Cao

Organizer/Responsable: Jiguo Cao

Date: Thursday June 2 / jeudi 2 juin

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-15:52]

Peijun Sang (University of Waterloo) **Zuofeng Shang** (New Jersey Institute of Technology) **Pang Du** (Virginia Polytechnic Institute and State University)

Statistical Inference for Functional Linear Quantile Regression

Inférence statistique pour la régression quantile linéaire fonctionnelle

We propose inferential tools for functional linear quantile regression where the conditional quantile of a scalar response is assumed to be a linear functional of a functional covariate. In contrast to conventional approaches, we employ kernel convolution to smooth the original loss function. The coefficient function is estimated under a reproducing kernel Hilbert space framework. A gradient descent algorithm is designed to minimize the smoothed loss function with a roughness penalty. With the aid of the Banach fixed-point theorem, we show the existence and uniqueness of our proposed estimator as the minimizer of the regularized loss function in an appropriate Hilbert space. Furthermore, we establish the convergence rate as well as the weak convergence of our estimator. As far as we know, this is the first weak convergence result for a functional quantile regression model. Pointwise confidence intervals and a simultaneous confidence band for the true coefficient function are then developed based on these theoretical properties. Numerical studies including both simulations and a data application are conducted to investigate the performance of our estimator and inference tools in finite sample.

Nous proposons des outils inférentiels pour la régression quantile linéaire fonctionnelle pour laquelle le quantile conditionnel d'une réponse scalaire se présente comme une fonction linéaire d'une covariable fonctionnelle. Contrairement aux approches classiques, nous utilisons une convolution de noyaux pour lisser la fonction de perte originale. La fonction du coefficient est estimée dans un cadre d'espace de Hilbert à noyau reproduisant. Un algorithme de descente du gradient est conçu pour minimiser la fonction de perte lissée avec une pénalité de rugosité. À l'aide du théorème du point fixe de Banach, nous montrons l'existence et l'unicité de l'estimateur proposé en tant que minimiseur de la fonction de perte régularisée dans un espace de Hilbert approprié. De plus, nous déterminons le taux de convergence ainsi que la convergence faible de notre estimateur. À notre connaissance, il s'agit du premier résultat de convergence faible d'un modèle de régression quantile fonctionnelle. Ensuite, sur la base de ces propriétés théoriques, nous mettons au point des intervalles de confiance ponctuels et une bande de confiance simultanée de la vraie fonction du coefficient. Enfin, nous réalisons des études numériques comprenant à la fois des simulations et une application à des données afin d'étudier l'efficacité de notre estimateur et de nos outils d'inférence dans un échantillon de taille finie.

[15:52-16:14]

Yafei Wang (University of Alberta)

M-estimation for varying coefficient model with functional response in reproducing kernel Hilbert space

Estimation M d'un modèle à coefficient variable avec réponse fonctionnelle dans un espace de Hilbert à noyau reproducteur

Motivated by dealing with imaging dataset that needs to model the dynamic relationship between functional response and a set of covariates, we consider a varying coefficient model with the functional response, and robust

Intéressés par le traitement d'un ensemble de données d'imagerie qui requiert une modélisation de la relation dynamique entre la réponse fonctionnelle et un ensemble de covariables, nous examinons un modèle à coefficient variable avec la réponse

estimates, M-estimate, of the model which is against outliers and heteroscedasticity are established. In the paper, we establish that the proposed estimators are minimax rate optimal. Alternating direction method of multipliers algorithm is adopted to optimize the objective function and its convergence is established. Furthermore, to characterize the spatial dependency among observation points of functional response within per subject, we propose weighted M-estimates by using the strength of both M-estimates and copula modeling. A general computation procedure in obtaining weighted M-estimates in practice is given. Simulation study and functional imaging dataset analysis examine the robustness of the proposed method against outliers.

fonctionnelle, puis nous élaborons des estimations robustes (estimations M) du modèle résistant aux valeurs aberrantes et à l'hétéroscédasticité. Dans cette présentation, nous démontrons que les estimateurs proposés sont optimaux à un taux minimax. Nous adoptons l'algorithme des directions alternées pour optimiser la fonction d'objectifs, puis nous déterminons sa convergence. Par ailleurs, pour caractériser la dépendance spatiale entre les points d'observation de la réponse fonctionnelle chez un même sujet, nous proposons des estimations M pondérées au moyen de la robustesse des estimations M et de la modélisation par copules. Nous proposons une procédure générale de calcul pour obtenir des estimations M pondérées concrètes. Nous examinons la robustesse de la méthode proposée par rapport aux valeurs aberrantes au moyen d'une étude de simulation et d'une analyse d'un ensemble de données d'imagerie fonctionnelle.

[16:14-16:36]

Evan Sidrow (The University of British Columbia) **Nancy Heckman** (University of British Columbia) **Sarah M.E. Fortune** (Dalhousie University) **Andrew W. Trites** (University of British Columbia) **Ian Murphy** (University of Florida) **Marie Auger-Mèthè** (University of British Columbia)

Modelling Functional Data with Hierarchical Hidden Markov Models: Applications to Animal Movement

Modélisation de données fonctionnelles avec modèles de Markov cachés hiérarchiques : applications au mouvement des animaux

Modern biologging sensors can record sequences of curves at very high frequencies, allowing researchers to observe biological processes such as animal movement at extremely fine scales. High-frequency data sets can exhibit state-switching, multi-scale dependence structures that are difficult to model with standard methods in functional data analysis. Inspired by data collected from a northern resident killer whale (Orcinus orca), we describe a hierarchical framework that treats curves as observations from a hidden Markov model. Each curve's distribution is defined by a fine-scale model whose parameters depend upon a coarse-scale latent process. Through simulations, we show that our model produces more interpretable state estimates and more accurate parameter estimates compared to existing methods. We also consider several computational challenges when modelling state-switching functional data with hidden Markov models.

Les capteurs biologiques modernes peuvent enregistrer des séquences de courbes à très haute fréquence, ce qui permet aux chercheurs d'observer des processus biologiques tels que le mouvement des animaux à des échelles extrêmement fines. Les ensembles de données à haute fréquence peuvent présenter des structures de dépendance multi-échelles à changement d'état qui sont difficiles à modéliser avec les méthodes standard d'analyse des données fonctionnelles. Inspirés par les données recueillies sur la population d'orques résidentes du Nord (Orcinus orca), nous décrivons un cadre hiérarchique qui traite les courbes comme des observations d'un modèle de Markov caché. La distribution de chaque courbe est définie par un modèle à échelle fine dont les paramètres dépendent d'un processus latent à échelle grossière. Par le biais de simulations, nous montrons que notre modèle produit des estimations d'état plus interprétables et des estimations de paramètres plus précises par rapport aux méthodes existantes. Nous examinons également plusieurs défis informatiques lors de la modélisation de données fonctionnelles à changement d'état avec des modèles de Markov cachés.

[16:36-16:58]

Haolun Shi (Simon Fraser University) **Jiguo Cao** (Simon Fraser University)

Robust Regression-Based Functional Principal Component Analysis

Analyse en composantes principales fonctionnelle par régression robuste

It is of great interest to conduct robust functional prin-

Il est primordial de réaliser une analyse en composantes princi-

cipal component analysis (FPCA) that can identify the major modes of variation in the stochastic process with the presence of outliers. A new robust FPCA method is proposed in a new regression framework. An M-estimator for the functional principal components is developed based on Huber's loss by iteratively fitting the residuals from the Karhunen-Loève expansion for the stochastic process under the robust regression framework. Our method can naturally accommodate sparse and irregularly-sampled data. When the functional data have outliers, our method is shown to render stable and robust estimates of the functional principal components; When the functional data have no outliers, we show via simulation studies that the performance of our approach is similar to that of the conventional FPCA method. The proposed robust FPCA method is demonstrated by analyzing some real data sets.

pales fonctionnelle (ACPF) robuste qui détermine les principaux modes de variation d'un processus stochastique en présence de valeurs aberrantes. Nous proposons une nouvelle méthode d'ACPF robuste dans un nouveau cadre de régression. Un estimateur M pour les composantes principales fonctionnelles est créé en fonction de la perte de Huber par l'ajustement itératif des résidus de l'expansion de Karhunen-Loève pour le processus stochastique dans le cadre de la régression robuste. Notre méthode peut naturellement s'adapter aux données éparses et échantillonnées irrégulièrement. Lorsque les données fonctionnelles présentent des valeurs aberrantes, notre méthode permet d'obtenir des estimations stables et robustes des composantes principales fonctionnelles. Lorsque les données fonctionnelles ne présentent pas de valeurs aberrantes, nous montrons par des études de simulation que l'efficacité de notre approche est semblable à celle de la méthode d'ACPF classique. Nous faisons une démonstration de la méthode d'ACPF robuste par l'analyse de certains ensembles de données réelles.

Recent Advances in Methodologies and Applications of Innovative Survival Models
Progrès récents en méthodes et applications de modèles de survie innovants

Chair/Président: Longhai Li

Organizer/Responsable: Longhai Li

Date: Thursday June 2 / jeudi 2 juin

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-16:00]

Yingwei (Paul) Peng (Queen's University) **Chyong-Mei Chen** (National Yang Ming Chiao Tung University) **Pao-sheng Shen** (Tunghai University) **Hsin-Jen Chen** (National Yang Ming Chiao Tung University)

Length-Biased and Interval-Censored Data with a Cure Fraction

Données biaisées en longueur et censurées par intervalle avec un taux de guérison

Length-biased data, a special case of left-truncated data, assume that the incidence of the initial event follows a homogeneous Poisson process. We will introduce an analysis of length-biased and interval-censored data with a cure fraction in this talk. The Cox proportional hazards model for the survival time of the susceptible individuals and the logistic regression model for the probability of being susceptible are employed to model the data. We construct the full likelihood function and obtain the nonparametric maximum likelihood estimates of the regression parameters by employing the EM algorithm. The large sample properties of the estimates are established. The performance of the method is assessed by simulations. The proposed model and method are applied to data from an early-onset diabetes mellitus study.

Les données biaisées en longueur, un cas particulier de données tronquées à gauche, reposent sur l'hypothèse que l'incidence de l'événement initial suit un processus de Poisson homogène. Dans cette présentation, nous présentons une analyse des données biaisées en longueur et censurées par intervalle avec un taux de guérison. Afin de modéliser les données, nous utilisons le modèle à risques proportionnels de Cox pour le temps de survie des individus à risque et le modèle de régression logistique pour la probabilité d'être à risque. Nous construisons la fonction de vraisemblance complète et obtenons les estimations non paramétriques par maximum de vraisemblance des paramètres de régression avec l'algorithme EM. Nous établissons les propriétés asymptotiques des estimations. Nous évaluons l'efficacité de la méthode par des simulations. Nous appliquons le modèle et la méthode proposés aux données d'une étude sur le diabète sucré à un stade précoce.

[16:00-16:30]

Shahedul Khan (University of Saskatchewan)

Accelerated Failure Time Models for Recurrent Event Data Analysis and Joint Modeling

Modèles à temps d'échec accélérés pour l'analyse de données d'événements récurrents et de modélisation conjointe

There are two commonly encountered problems in survival analysis: (a) recurrent event data analysis, where an individual may experience an event multiple times over follow-up; and (b) joint modeling, where the event time distribution depends on a longitudinally measured internal covariate. The proportional hazards (PH) family offers an attractive modeling paradigm for recurrent event data analysis and joint modeling. Although there are well-known techniques to test the PH assumption for standard survival data analysis, checking this assumption for joint modeling has received relatively less atten-

On rencontre couramment deux problèmes en analyse de survie : (1) l'analyse de données d'événements récurrents, où un individu peut vivre un événement plusieurs fois durant un suivi, et (2) la modélisation conjointe, lorsqu'une distribution de temps d'événement dépend d'une covariable interne mesurée de façon longitudinale. La famille de risque proportionnel (RP) apporte un paradigme de modélisation intéressant pour l'analyse de données d'événements récurrents et la modélisation conjointe. Bien qu'il existe des techniques reconnues pour tester l'hypothèse de RP dans les analyses de données de survie standard, on s'intéresse relativement moins à tester cette hypothèse pour des modélisations

tion. An alternative framework involves considering an accelerated failure time (AFT) model, which is particularly useful when the PH assumption fails. In this talk, I will describe methodologies to analyze these types of data using the AFT family of distributions. I will also present the computational algorithms for statistical inference.

conjointes. Un cadre différent doit tenir compte d'un modèle à temps d'échec accéléré (TÉA), qui est particulièrement pratique lorsque l'hypothèse de RP échoue. Lors de cet exposé, je décrirai des méthodologies d'analyse pour ces types de données au moyen de distributions à TÉA. Je présenterai aussi les algorithmes computationnels de l'inférence statistique.

[16:30-17:00]

Cindy Xin Feng (Dalhousie University) **Tingxuan Wu** (University of Saskatchewan) **Longhai Li** (University of Saskatchewan)

A Comparative Study of R packages for Semiparametric Shared Frailty Models

Étude comparative de paquets R pour des modèles semi-paramétriques à fragilités partagées

Frailty models are often used to model the unobserved heterogeneity and clustered survival data. A shared frailty model is a random-effect model where the frailties are common or shared among individuals within groups. Different R packages are available for fitting shared frailty models such as survival, frailtyEM, fraltypack, frailtysurv, and frailtyHL. However, little research has been conducted to compare the performance of various R packages for fitting shared frailty models, making it difficult for users to decide on an appropriate tool for analyzing clustered survival data. We aim to compare the performance of the R packages via a series of simulation studies. The bias and variance of the parameter estimates, rate of convergence, and computational time of the packages are compared. The advantages and limitations of the software are discussed in detail.

Les modèles à fragilités sont souvent utilisés pour modéliser l'hétérogénéité non observée et les données de survie groupées. Un modèle à fragilités partagées est un modèle à effet aléatoire dans lequel les fragilités sont communes ou partagées parmi les individus dans les groupes. Divers paquets R sont disponibles pour l'ajustement des modèles à fragilités partagées, comme survival, frailtyEM, frailtypack, frailtysurv et frailtyHL. Cependant, comme peu de recherche comparative a été effectuée sur la performance de divers paquets R pour l'ajustement des modèles à fragilités partagées, il devient difficile pour les utilisateurs de décider d'un outil adéquat pour analyser les données de survie groupées. Notre but est de comparer la performance des paquets R à l'aide d'une série d'études par simulations. La comparaison des paquets porte sur le biais et la variance des estimations des paramètres, leur taux de convergence et leur temps d'exécution. Les avantages et les limites du logiciel seront présentés en détail.

Challenges and Examples in Data Science Consultation
Défis et exemples sur la consulation en sciences des données

Chair/Président: Jean-François Plante

Organizer/Responsable: Jean-François Plante

Date: Thursday June 2 / jeudi 2 juin

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-16:00]

Steve Kanters (RainCity Analytics)

Challenges and Highlights of Data Science Consulting

Défis et points forts de la consultation en science des données

A critical choice that new statistics graduates face is whether to pursue a career in academia or the private industry. It is well understood that data science consulting can be more financially rewarding and is faster paced, but other details are often more nebulous. At the macro-level, the challenges of consulting are similar to academia. They include the need to develop and maintain a network of fellow professionals, building and maintaining a reputation, and securing financing. There are some important differences here, particularly with respect to financing, which we will discuss. At the day-to-day level, challenges in data science consulting often stem from the client-centered nature of the work. It is important to determine what the client knows so as to understand whether their request should be revised. Communication skills for dissemination are equally crucial. In this talk we will expand upon these challenges and share some lived experiences.

Les nouveaux diplômés en statistiques sont confrontés au choix déterminant de poursuivre une carrière dans le milieu universitaire ou dans le secteur privé. Tout le monde sait que le domaine de la consultation en science des données peut être plus payant et que le rythme y est plus soutenu, mais d'autres détails sont souvent plus nébuleux. Sur le plan général, les défis liés au conseil sont semblables à ceux du milieu universitaire (développer et entretenir un réseau de collègues professionnels, bâtir et maintenir sa réputation, obtenir du financement). Il existe quelques différences importantes, notamment en ce qui concerne le financement, dont nous allons parler. Au quotidien, les défis à relever en matière de consultation en science des données proviennent souvent de la nature du travail axé sur le client. Il est important de déterminer ce que le client sait afin de comprendre si sa demande doit être révisée. Les compétences en communication pour la diffusion sont tout aussi cruciales. Dans cette présentation, nous allons examiner ces défis et échanger quelques expériences vécues.

[16:00-16:30]

Ghislene Zerguini (HEC Montréal)

Setting Your Data Workforce up for Success

Comment contribuer au succès des responsables de données en entreprise

In this digital age where Analytics & AI are the top two focus areas for companies (Gartner 2021), one of the game changing factors to business resides in successfully leveraging data. As trusted advisors to our clients and leaders in our practices we create valuable solutions by engaging teams with hybrid profiles, focusing on role clarification and upscaling of resources with an emphasis on the individual strengths of each individual and the overall performance of the team. The new challenges we need to tackle require new perspectives. It is imperative

En cette ère numérique où l'analytique et l'intelligence artificielle sont les deux principaux champs d'intérêt des entreprises (Gartner 2021), une bonne exploitation des données est l'un des facteurs qui change la donne en affaires. À titre de conseillers de confiance de nos clients et de chefs de file dans nos pratiques, nous créons des solutions précieuses en mobilisant des équipes avec des profils hybrides, mettant l'accent sur la clarification des rôles et le rehaussement des ressources, tout en misant sur les forces propres à chacun et la performance globale de l'équipe. Les nouveaux problèmes que nous devons résoudre exigent de nouvelles perspectives. Il est

to adapt the management consulting approaches to address this predicament and the data workforce is at the center of the solution. As we have come to witness the ongoing war on talent and the "Jack of all trades" expectations of the market, it is essential to discuss the roles of data scientists and technology consultants and bridge the gap.

impératif d'adapter les approches de conseil en gestion pour faire face à cette difficulté et les responsables des données sont au cœur de la solution. Comme nous sommes témoins de cette quête fébrile de gens compétents et des attentes « touche-à-tout » du marché, il est essentiel de s'interroger sur les rôles distincts des scientifiques de données et des consultants en technologie afin de combler cet écart.

[16:30-17:00]

Sarah Legendre Bilodeau (Videns Analytics) **Sébastien Duguay** (Videns Analytics)

Challenges and Examples in Data Science Consultation

La consultation en science des données - défis et exemples

Data science has experienced a major boom in recent years. Indeed, with the rise in popularity in the industry of business intelligence a decade ago, advanced analytics a few years later and artificial intelligence more recently, data science is at the heart of the concerns of companies of all sizes. The common thread running through all these terms : statistics. In their ambition to better exploit their data and increase their productivity, many companies choose to rely on a consulting firm in the field of data science. This is how Videns Analytics is mobilized to support companies of all sizes in their needs for data valorization and artificial intelligence developments. In this talk, we will discuss the challenges faced by companies of all sizes in their projects of data valuation. These challenges will be presented through concrete examples using data science.

La science des données a connu un essor majeur dans les dernières années. En effet, avec la montée en popularité dans l'industrie de l'intelligence d'affaires il y a une dizaine d'années, l'analytique avancée quelques années plus tard et l'intelligence artificielle plus récemment, la science des données est au cœur des préoccupations d'entreprises de toutes tailles. Le fil conducteur au travers toutes ces dénominations : la statistique. Dans leur ambition de mieux valoriser leurs données et d'augmenter leur productivité, plusieurs entreprises choisissent de s'appuyer sur une entreprise de consultation dans le domaine de la science des données. C'est ainsi que Videns Analytics est mobilisée pour appuyer des entreprises de toutes tailles dans leurs besoins de valorisation des données et de développements d'intelligence artificielle. Dans cette conférence, il sera donc question des défis rencontrés par des entreprises de toutes tailles dans leurs projets de valorisation des données. Ces défis seront présentés au travers d'exemples concrets mobilisant la science des données.

Chair/Président: Joan X. Hu

Organizer/Responsable: Joan X. Hu

Date: Thursday June 2 / jeudi 2 juin

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-16:00]

Hongzhe Lee (University of Pennsylvania)

Transfer Learning in High-dimensional Linear Regression and Graphical Models

Apprentissage par transfert dans une régression linéaire de haute dimension et des modèles graphiques

This talk considers estimation and prediction of high-dimensional linear regression model in the setting of transfer learning, using samples from the target model as well as auxiliary samples from different but possibly related models. When the set of "informative" auxiliary samples is known, an estimator and a predictor are proposed and their optimality is established. The optimal rates of convergence for prediction and estimation are faster than the corresponding rates without using the auxiliary samples. This implies that knowledge from the informative auxiliary samples can be transferred to improve the learning performance of the target problem. When sample informativeness is unknown, a data-driven procedure for transfer learning, called Trans-Lasso is proposed, and its robustness to non-informative auxiliary samples and its efficiency in knowledge transfer is established. The proposed procedures are demonstrated in numerical studies and in analysis of the GTEx data sets.

Cet exposé aborde l'estimation et la prédiction d'un modèle de régression linéaire de grande dimension dans le cadre d'apprentissage par transfert, au moyen d'échantillons du modèle cible et d'échantillons auxiliaires tirés de différents modèles possiblement reliés. Lorsque l'ensemble des échantillons auxiliaires ≪informatifs≫ est connu, on propose un estimateur et un prédicteur, et leur caractère optimal est établi. Les taux optimaux de convergence de la prédiction et de l'estimation sont plus rapides que les taux correspondants sans l'emploi des échantillons auxiliaires. Cela sous-entend que l'information contenue dans les échantillons auxiliaires informatifs peut être transférée afin d'améliorer la performance d'apprentissage du problème cible. Lorsque le caractère informatif de l'échantillon est inconnu, nous proposons une procédure d'apprentissage par transfert basée sur les données, que l'on appelle Trans-Lasso, dont la robustesse relative aux échantillons auxiliaires non informatifs et l'efficacité du transfert d'information sont démontrées. Les procédures proposées sont illustrées par l'entremise d'études numériques et d'analyses de jeux de données GTEx.

[16:00-16:30]

Juxin Liu (University of Saskatchewan)

Bias Analysis for Misclassification Errors in both the Response Variable and Covariate

Analyse de biais pour les erreurs de classification dans la variable de réponse et la covariable

Misclassification in both the response variable and the covariate has received very limited attention in the literature. For situations where the response variable and the covariate are simultaneously subject to misclassification errors, often an assumption of independent misclassification errors is used without justification. The aim of our work is to show the harmful consequences of inappropriate adjustment for the joint misclassification errors, that is, ignoring the dependence between the

Peu de documentation aborde le sujet de la mauvaise classification dans la variable de réponse et la covariable. Lorsque la variable de réponse et la covariable génèrent simultanément des erreurs de classification, on adopte fréquemment une hypothèse d'indépendance des erreurs de classification sans justification. L'objectif de notre travail est de mettre en évidence les conséquences négatives d'un ajustement inadéquat des erreurs de classification conjointes, c'est-à-dire ignorer la dépendance entre les processus de classification incorrecte de la variable de réponse

misclassification process of the response variable and that of the covariate. Moreover, we propose likelihood ratio tests to check the nondifferential/independent misclassification assumption in main study/internal validation study designs. Our simulation studies indicate that only ignoring the dependent error structure can be even worse than ignoring all the misclassification errors when the validation data size is relatively small. We illustrate the methodology by a real data example.

et de la covariable. De plus, nous proposons un test du rapport des vraisemblances servant à vérifier l'hypothèse de classification incorrecte indépendante et non différentielle dans une étude principale et avec validation interne. Nos études en simulation indiquent qu'ignorer uniquement la structure d'erreur dépendante peut être pire qu'ignorer toutes les erreurs de classification lorsque la taille des données de validation est relativement petite. Nous illustrons la méthodologie à partir d'un exemple basé sur des données réelles.

[16:30-17:00]

Richard J. Cook (University of Waterloo) **Jerald F. Lawless** (University of Waterloo)

Analysis of Life History Data Obtained from Biased Sampling and Observation Schemes

Analyse de données de cycle de vie tirées d'échantillonnage et de schémas d'observation biaisés

A great deal of modern health research aims to exploit data from disease registries or administrative health records in order to supplement or replace the more costly collection of information from prospective cohort studies. Understanding sampling mechanisms and the factors governing the nature and duration of follow-up are critical for such data sources, however, to ensure valid inference and interpretable findings. This talk will discuss statistical challenges from, and approaches for dealing with, dependent delayed-entry, dependent intermittent observation schemes, and dependent loss to follow-up. Multistate models will be used to represent the sampling and observation processes, communicate the assumptions for standard analyses, and provide a framework for joint modelling which enables one to mitigate biases through simultaneous model fitting or two-stage procedures using inverse probability weighting.

Une part importante de la recherches moderne sur la santé tente d'exploiter les données provenant de registres des maladies ou de dossiers médicaux administratifs afin de supplémenter ou remplacer la collecte d'information coûteuse à partir d'études de cohorte prospective. Comprendre les mécanismes d'échantillonnage et les facteurs gouvernant la nature et la durée des suivis est toutefois primordial afin d'assurer la validité des inférences et l'interprétabilité des résultats relatifs à de telles sources de données. Cet exposé abordera les défis statistiques et les approches de gestion relatives aux entrées retardées dépendantes, aux schémas d'observation intermittents dépendants et pertes de suivi dépendantes. Nous employons des modèles multiétats pour représenter les processus d'observation et d'échantillonnage, communiquer les hypothèses d'analyses standards, et procurer un cadre de modélisation conjointe qui permet de minimiser le biais par l'entremise d'ajustement de modèle simultané ou de procédures à deux étapes se servant d'une pondération par probabilité inverse.

Chair/Président: Félix Camirand Lemyre

Organizer/Responsable: Félix Camirand Lemyre

Date: Thursday June 2 / jeudi 2 juin

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-16:00]

Juliana Schulz (HEC Montréal) **Erica E.M. Moodie** (McGill University)

Doubly Robust Estimation of Optimal Dosing Strategies

Estimation doublement robuste des stratégies de dosage optimales

The goal of precision medicine is to develop personalized treatment strategies which account for individual-level characteristics and reflect patient response heterogeneity. This notion certainly has important implications in health care practices and patient well-being. The growing interest in the subject has led to the development of several statistical frameworks for estimating optimal treatment regimes. This talk will focus on a regression-based method for estimating the optimal drug dose level. The proposed method is an extension to the dynamic weighted ordinary least squares regression approach, and is shown to be appropriate for treatment measured on both a continuous and categorical scale. We consider a broad class of weight functions satisfying a certain balancing condition and demonstrate that incorporating these weights in a linear regression model allows for doubly robust estimation of the optimal treatment strategy. This approach will be illustrated both through simulations and an application to Warfarin dose strategies.

L'objectif de la médecine de précision est d'élaborer des stratégies de traitement personnalisées qui tiennent compte des caractéristiques individuelles et qui reflètent l'hétérogénéité de la réponse des patients. Cette notion entraîne bien sûr des répercussions importantes sur les pratiques de soins de santé et le bien-être des patients. L'intérêt croissant pour le sujet a conduit à la création de plusieurs cadres statistiques pour estimer les régimes de traitement optimaux. Dans cette présentation, nous mettrons l'accent sur une méthode basée sur la régression pour estimer le niveau de dose de médicament optimal. La méthode proposée est une extension de l'approche de régression dynamique pondérée des moindres carrés ordinaires, et s'avère adéquate pour le traitement selon une échelle continue et catégorique. Nous nous penchons sur une vaste classe de fonctions poids qui répondent à une certaine condition d'équilibre, et nous démontrons que l'intégration de ces poids dans un modèle de régression linéaire permet une estimation doublement robuste de la stratégie de traitement optimale. Nous illustrons cette approche à l'aide de simulations et d'une application aux stratégies de dosage de la warfarine.

[16:00-16:30]

Kevin McGregor (York University) **Nneka Okaeme** (York University)

Proportionality-Based Association Measures in Count-Based Compositional Data

Mesures d'association basées sur la proportionnalité pour des données de compatage compositionnelles

Compositional data comprise vectors of quantitative measures that describe the constituent parts of a whole. Data arising from various -omics platforms are compositional in nature. For instance, 16S sequencing and single-cell RNA-sequencing are both examples of platforms that yield compositional data. However, correlations between features on raw counts have no meaningful interpretation. Measures of association based on pro-

Les données compositionnelles comprennent des vecteurs de mesures quantitatives qui décrivent les parties constituantes d'un tout. Les données découlant de diverses plateformes omiques sont compositionnelles par nature. Le séquençage 16S et le séquençage d'ARN de cellule unique sont deux exemples de plateformes qui fournissent des données compositionnelles. Cependant, les corrélations entre les caractéristiques des comptages bruts n'ont pas d'interprétation utile. Des mesures d'association basées sur

portionality have previously been proposed, and have been shown to outperform other association metrics in single-cell sequencing data. These metrics were designed for continuous compositional data and do not account for the additional sampling variability in sequencing data. In this talk I will discuss the pitfalls of using standard proportionality metrics in count-based platforms and show that models accounting for sequencing variability (in particular the multinomial logit-normal model) improve estimates of association.

la proportionnalité ont été proposées antérieurement et il a été montré qu'elles surpassaient d'autres mesures d'association dans le cas de données de séquençage de cellule unique. Ces mesures ont été conçues pour des données compositionnelles continues et ne prennent pas en compte la variabilité d'échantillonnage additionnelle des données de séquençage. Dans cet exposé, je vais présenter les embûches liées à l'utilisation de mesures standards de proportionnalité dans les plateformes basées sur des comptages et montrer que les modèles qui prennent en compte la variabilité du séquençage (en particulier le modèle logit-normal multinomial) améliorent l'estimation de l'association.

[16:30-17:00]

Samantha-Jo Caetano (University of Toronto) **Rohan Alexander** (University of Toronto)

Further Developments of a Toolkit for Learning R at All Levels.

Le développement supplémentaire d'une boîte à outils pour l'apprentissage du langage R à tous les niveaux

Upon teaching a senior level applied statistics course, we noticed a disparity in R programming levels of third- and fourth-year statistics major students. With the goal of filling programming knowledge gaps in these students and to have all students on a more similar programming level, upon entering senior level statistics courses, Dr. Alexander and Dr. Caetano embarked on assembling a team of graduate and undergraduate students to develop a toolkit to help students improve their programming in R. The toolkit is a set of interactive modules that students complete autonomously. Since the initial development of the toolkit about (one year ago) many updates and developments have been made. Namely, the toolkit is now being re-formatted into a textbook and more formative assessments have been added. The modules start from the very basics of installing R to working in tidyverse to employing git commands. In this talk, we will outline the new developments and uses of this toolkit and highlight some recommendations and future steps.

Lors de l'enseignement d'un cours de statistiques appliquées de niveau supérieur, nous avons remarqué un écart relatif au niveau de programmation en R entre les étudiants en statistiques de troisième et quatrième années. Afin de rétrécir cet écart en connaissances de programmation entre les étudiants et faire en sorte qu'ils ont tous un niveau de programmation semblable au moment de commencer un cours de statistique de niveau supérieur, le docteur Alexander et le docteur Caetano ont assemblé une équipe d'étudiants de cycle supérieur et de premier cycle dans le but de concevoir une boîte à outils qui aidera les étudiants à améliorer leurs aptitudes de programmation en R. La boîte à outils consiste en un ensemble de modules interactifs que les étudiants peuvent compléter de façon autonome. Depuis le développement initial de la boîte à outils (il y a environ un an), de nombreuses mises au point et mises à jour ont été appliquées. Entre autres, la boîte à outils est actuellement reformatée en un manuel scolaire et de nouvelles évaluations formatrices y sont ajoutées. Les modules commencent par la base de l'installation de R, puis progressent vers tidyverse jusqu'à l'emploi de méthodes GIT. Lors de cet exposé, nous décrirons les grandes lignes des nouveaux développements et de l'utilisation de cette boîte à outils, soulignerons quelques recommandations et aborderons les étapes à venir.

Copula-based Methods
Méthodes basées sur les copules

Chair/Président: Wesley S. Burr

Date: Thursday June 2 / jeudi 2 juin

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-15:45]

Robert Zimmerman (University of Toronto) **Vianey Leos Barajas** (University of Toronto) **Radu V. Craiu** (University of Toronto)

Copula Modelling of Serially Correlated Multivariate Data with Hidden Structures

Modélisation par copules de données multivariées sériellement corrélées avec des structures cachées

In applications where streams of data exhibit variable latent structures, it is natural to model the data-generating process as a finite-state hidden Markov model (HMM). When observing vector outcomes, we consider multivariate state-dependent distributions that are fused together by copulas. Such a "copula-within-HMM" framework is highly flexible, because it provides the freedom to vary both the marginal distributions of observed outcomes and the copula that determines the dependencies between them. However, inference for this model is not straightforward; while the EM algorithm is the standard technique for parameter estimation within HMMs, a direct application becomes unwieldy in the face of the additional model complexity brought about by the copula. We develop a robust and efficient EM algorithm for the copula-within-HMM model, and show that it performs well in both model estimation and state classification tasks on a variety of simulated and real-world datasets.

Dans les applications où les flux de données présentent des structures latentes variables, il est naturel de modéliser le processus de génération des données comme un modèle de Markov caché (MMC) à états finis. Lorsque nous observons des résultats vectoriels, nous considérons des distributions multivariées dépendantes de l'état qui sont fusionnées par des copules. Un tel cadre de « copule avec MMC » est très flexible, car il offre la liberté de faire varier à la fois les distributions marginales des résultats observés et la copule qui détermine les dépendances entre eux. Cependant, l'inférence pour ce modèle n'est pas simple ; alors que l'algorithme EM est la technique standard d'estimation des paramètres dans les MMC, une application directe devient difficile face à la complexité supplémentaire du modèle due à la copule. Nous développons un algorithme EM robuste et efficace pour le modèle copule-dans-MMC, et montrons qu'il est performant pour les tâches d'estimation de modèle et de classification d'état sur une variété d'ensembles de données simulées et réelles.

[15:45-16:00]

Guanjie Lyu (University of Windsor) **Mohamed Belalia** (University of Windsor)

Testing Symmetry for Bivariate Copulas using Bernstein Polynomials

Test de symétrie pour copules bivariées à l'aide des polynômes de Bernstein

In this talk, tests of symmetry for bivariate copulas are introduced and studied using empirical Bernstein copula. Three statistics are proposed and their asymptotic properties are established. Besides, a multiplier bootstrap Bernstein version is investigated for implementation purpose. Simulation study and real data application showed that the Bernstein tests outperform the tests based on the empirical copula.

Dans cet exposé, nous introduisons des tests de symétrie pour les copules bivariées et les étudions à l'aide de la copule empirique de Bernstein. Nous proposons trois statistiques et établissons leurs propriétés asymptotiques. De plus, nous étudions une version du bootstrap multiplicateur de Bernstein à des fins de mise en œuvre. Une étude de simulation et une application sur des données réelles montrent que les tests de Bernstein sont plus performants que les tests basés sur la copule empirique.

[16:00-16:15]

H. Roland G. Dossa (Université du Québec à Montréal)

Generalized Functional Linear Mixed Models for Binary Traits in Family-Based Designs via Copulas

Modèles mixtes linéaires fonctionnels généralisés pour les traits binaires dans les devis basés sur la famille via copules

In this work, we propose a flexible family-based association test for rare variants and a binary trait using a new marginal generalized functional linear mixed model with a Gaussian Copula (NRVATGFLMM) while adjusting for covariates. To assess the performance of our newly introduced statistics, simulation studies were conducted to evaluate type I error rates and power levels. We make a comparison with some other existing statistical methods for family-based association tests.

Dans ce travail, nous proposons un test d'association flexible basé sur la famille pour les variants rares et un trait binaire en utilisant un nouveau modèle linéaire mixte fonctionnel généralisé marginal avec copule gaussienne (NRVATGFLMM) tout en ajustant pour les covariables. Pour déterminer la performance de notre nouvelle statistique, nous menons des études de simulation pour évaluer les taux d'erreur de type I et les niveaux de puissance. Nous effectuons une comparaison avec d'autres méthodes statistiques existantes pour les tests d'association basés sur la famille.

[16:15-16:30]

Xinyao Fan (The University of British Columbia)

Proxies in High-dimensional Factor Copula Models

Proxys dans les modèles de copules factorielles à haute dimension

Factor models are a parsimonious way to explain the dependence of variables using several latent variables. In Gaussian factor models (1-factor, bi-factor, oblique factor) and their factor copula counterparts, we propose a way to estimate the latent variables with proxies. We show the proxies which are defined as the conditional expectation of the latent factors given the observed variables are consistent as the number of observed variables linked to each latent variable increases. In the Gaussian and copula case, the proxies can be calculated in the closed-form via matrix calculations and Gaussian-Legendre quadrature respectively. The proxies can help to select the linking copulas in the factor models and to obtain accurate parameter estimates efficiently when the number of observed variables linked to each latent variable is large. In practice, 20 to 40 is adequate for use of proxies provided the dependence in the model is "strong" enough. One application is the stock return data.

Les modèles factoriels sont un moyen parcimonieux d'expliquer la dépendance des variables à l'aide de plusieurs variables latentes. Pour les modèles factoriels gaussiens (1-facteur, bi-facteur, facteur oblique) et leurs homologues à copule factorielle, nous proposons une façon d'estimer les variables latentes avec des proxys. Nous montrons que les proxys, qui sont définis comme l'espérance conditionnelle des facteurs latents sachant les variables observées, sont convergents lorsque le nombre de variables observées liées à chaque variable latente augmente. Dans le cas gaussien et de la copule, les proxys peuvent être calculés sous forme analytique via des calculs matriciels et la quadrature de Gauss-Legendre respectivement. Les proxys peuvent aider à sélectionner les copules de liaison dans les modèles factoriels et à obtenir efficacement des estimations précises des paramètres lorsque le nombre de variables observées liées à chaque variable latente est important. En pratique, 20 à 40 est adéquat pour l'utilisation de proxys à condition que la dépendance dans le modèle soit suffisamment « forte ». Les données sur les rendements boursiers en sont une application.

[16:30-16:45]

Serge B. Provost (The University of Western Ontario) **Yishan Zang** (Western University)

On Modeling Multivariate Data from Marginal Distributions

Modélisation de données multivariées à partir de distributions marginales

According to Sklar's theorem, the joint density function of a set of random variables can be expressed in terms of their marginal densities and associated copula density. Noting that for a given sample, the latter can readily and accurately be secured from a Bernstein empirical copula density approximant, it then suffices to obtain estimates of the marginal density functions of the variables involved in order to model their joint distribution.

Selon le théorème de Sklar, la fonction de densité conjointe d'un ensemble de variables aléatoires peut être exprimée au moyen de la densité marginale de ces variables et de la densité de copule associée. Étant donné que pour un échantillon donné, la densité de la copule peut être facilement et précisément obtenue à partir d'une approximation de la densité de la copule empirique de Bernstein, il suffit alors d'obtenir des estimations des fonctions de densité marginale des variables concernées afin de

The proposed approach enables one to set an appropriate smoothness level for each marginal density estimate. What is more, it also involves the copula of the distribution, which completely encapsulates the dependencies between the variables. Additionally, a methodology enabling one to obtain an initial copula density estimate from Deheuvels' empirical copula estimator will be described and an alternative representation of copula density estimates will be introduced. Several illustrative examples will be presented.

modéliser leur distribution conjointe. Grâce à l'approche que nous proposons, il est possible de fixer un niveau de lissage approprié pour chaque estimation de densité marginale. De plus, notre approche intègre également la copule de la distribution, qui englobe complètement les dépendances entre les variables. Par ailleurs, nous décrirons une méthodologie permettant d'obtenir une estimation initiale de la densité de copule à partir de l'estimateur de la copule empirique de Deheuvels et nous présenterons également une représentation alternative des estimations de la densité de copule. Plusieurs exemples illustratifs seront présentés.

[16:45-17:00]

Salah El Adlouni (Université de Moncton) **A. Boukili-Makhoukhi** (Université de Moncton) **W. El Hannoun** (Université Mohamed) **A. Zoglat** (Université Mohamed)

Vine Copulas to Estimate Intensity-Duration-Frequency Curves

Copules en vignes et estimation des courbes Intensité-Durée-Fréquence

Intensity-Duration-Frequency (IDF) curves of precipitation are a reference decision support tool in urban hydrology. It allows to estimate extreme precipitation and its return periods. Classically, the IDF curves are estimated using univariate frequency analysis of the maximum annual intensities of precipitation for different durations. It is then assumed that the annual maxima of different durations are independent to simplify parameter estimation. This hypothesis is very strong and is not necessarily verified for several climatic regions. This study examines the effects of the independence hypothesis by proposing a multivariate model that considers the dependencies between precipitation intensities of different durations. The multivariate model is based on D-vine copulas and the generalized distribution of extreme values (GEV) is considered as marginal model. An illustration of the proposed approach is made for precipitation at Moncton in the province of New Brunswick in Eastern Canada.

Les courbes Intensité-Durée-Fréquence (IDF) des précipitations sont un outil d'aide à la décision de référence en hydrologie urbaine. Elles permettent d'estimer les précipitations extrêmes et leurs périodes de retour. Classiquement, les courbes IDF sont estimées à partir d'une analyse fréquentielle univariée des intensités maximales annuelles de précipitations pour différentes durées. On suppose alors que les maximas annuels de différentes durées sont indépendants pour simplifier l'estimation des paramètres. Cette hypothèse est très forte et n'est pas nécessairement vérifiée pour plusieurs régions climatiques. Cette étude examine les effets de l'hypothèse d'indépendance en proposant un modèle multivarié qui considère les dépendances entre les intensités de précipitations de différentes durées. Le modèle multivarié est développé avec des copules en vignes et la distribution généralisée des valeurs extrêmes (GEV) est considérée comme modèle marginal. Une illustration de l'approche proposée est faite pour les précipitations à Moncton dans la province du Nouveau-Brunswick dans l'Est du Canada.

Chair/Président: Zihang Lu

Date: Thursday June 2 / jeudi 2 juin

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-15:45]

Xiawen Zhang (University of Toronto: Dalla Lana School of Public Health) **Eleanor M. Pullenayegum** (University of Toronto&SickKids)

The Bias of Parameters in Inverse-Intensity Weighted GEEs when People Without a Visit are Excluded

Biais des paramètres dans les EEG à pondération par intensité inverse lorsque les personnes sans visite sont exclues

Longitudinal data can be used to study disease progression and is often collected through chart reviews that feature irregular visit times. Traditional methods such as generalized estimating equations (GEEs) and mixed effect models lead to biased estimates when visit and outcome processes are related. Inverse-intensity weighed GEEs (IIW-GEEs) account for dependence between the visit and outcome processes. A common issue with chart reviews is that individuals with no visits are excluded from the dataset. We set out to examine the bias of regression parameters in IIW-GEEs when people without a visit are excluded. We show analytically that there is bias when people with no events are excluded, and verify this in a simulation study. Moreover, we show that decreasing visit frequency leads to an increase in bias on omitting people with no events. We recommend that everyone should be recorded in datasets collected through chart reviews, regardless of whether there is follow-up visit.

Les données longitudinales, qui permettent d'étudier la progression d'une maladie, sont souvent tirées de l'examen de dossiers qui présentent des temps de visite irréguliers. Or les méthodes traditionnelles telles que les équations d'estimation généralisées (EEG) et les modèles à effets mixtes conduisent à des estimations biaisées lorsque les processus de visite et de résultat sont liés. Les EEG à pondération par intensité inverse (PII-EEG) tiennent compte de la dépendance entre les processus de visite et de résultat. Un problème courant avec les revues de dossiers est que les individus sans visite sont exclus de l'ensemble de données. Nous avons entrepris d'examiner le biais des paramètres de régression dans les PII-EEG lorsque les personnes sans visite sont exclues. Nous montrons analytiquement qu'il y a un biais lorsque les personnes n'ayant aucun événement sont exclues, et nous le vérifions par une étude de simulation. De plus, nous montrons que la diminution de la fréquence des visites entraîne une augmentation du biais lors de l'omission des personnes ne présentant aucun événement. Nous recommandons que toutes les personnes soient enregistrées dans les ensembles de données collectées par l'examen des dossiers, qu'il y ait ou non une visite de suivi.

[15:45-16:00]

Rose Garrett (University of Toronto)

Why Recommended Visit Intervals should be Extracted when Conducting Longitudinal Analyses using Electronic Health Record Data

Pourquoi extraire les intervalles de rendez-vous recommandés dans les analyses longitudinales à l'aide de données de dossiers médicaux électroniques

Using routinely collected data from electronic health records offers a low-cost approach to investigating disease progression over multiple years of follow-up. However, this study design can lead to a biased sample since patients interact with the healthcare system more often when they are unwell and thus there is an overrepresentation of measurements on sicker patients.

L'utilisation de données régulièrement collectées dans les dossiers médicaux électroniques est une approche à faible coût pour étudier le progression d'une maladie au cours de plusieurs années de suivi. Cette conception de recherche peut cependant donner un échantillon biaisé, puisque les patients interagissent avec le système de santé plus souvent en cas de maladie, ce qui entraîne une surreprésentation des mesures relatives aux patients plus ma-

We show how the rigour of longitudinal analyses can be enhanced by leveraging information that has never been used before but is often already recorded in patient charts: physician recommendations on when the next visit should occur. Specifically, we demonstrate how recommended intervals can be used in examining and classifying the irregular visit process, and in evaluating the robustness of conclusions to untestable assumptions about the visit process. We illustrate our approach using data from a clinic-based cohort of patients with juvenile dermatomyositis at the Hospital for Sick Children.

Nous montrons comment la rigueur des analyses longitudinales peut être renforcée en tirant parti d'une information jamais utilisée antérieurement, bien que souvent déjà notée dans les dossiers de patients : l'intervalle recommandé par le médecin avant un prochain rendez-vous. Nous démontrons précisément que les intervalles recommandés peuvent être utilisés pour l'examen et la classification du processus de rendez-vous irréguliers et l'évaluation de la robustesse des conclusions aux hypothèses non vérifiables quant au processus de rendez-vous. Nous illustrons notre approche à l'aide de données cliniques portant sur une cohorte de patients atteints de dermatomyosite juvénile à l'Hospital for Sick Children.

[16:00-16:15]

Omidali Aghababaei Jazi (University of Toronto Mississauga) **Eleanor M. Pullenayegum** (Hospital for Sick Children (Sickkids))

Dynamic Prediction for Longitudinal Data with Irregular and Outcome-dependent Follow-up

Prédiction dynamique pour données longitudinales avec suivi irrégulier et dépendant des résultats

Statistical methodologies for longitudinal data with irregular and outcome-dependent follow-up have received much attention over the last two decades as the standard methods such as generalized estimating equation method may lead to biased estimates in this situation. An important problem in this setting is how to predict subsequent outcomes in a prospective longitudinal study. In this talk, we will present a computationally efficient joint modelling framework for predicting subsequent outcome values. We will use some standard diagnostic tools to assess the performance of the prediction model and conduct simulation studies to examine the finite sample properties of the prediction procedure. We also will apply the procedure to a data set from a multistage randomized clinical trial for treating major depressive disorder. Keywords: Longitudinal Data, Informative Follow-up, H-Likelihood, Prediction.

Les méthodes statistiques applicables aux données longitudinales avec suivi irrégulier et dépendant des résultats ont fait l'objet d'une grande attention ces deux dernières décennies, car les méthodes standard telles que l'équation d'estimation généralisée peuvent conduire à des estimations biaisées dans cette situation. Un problème important dans ce contexte est de savoir comment prédire les résultats ultérieurs dans une étude longitudinale prospective. Dans cet exposé, nous présenterons un cadre de modélisation conjointe efficace sur le plan informatique pour prédire les valeurs des résultats ultérieurs. Nous utiliserons des outils de diagnostic standard pour évaluer la performance du modèle de prédiction et nous mènerons des études par simulation pour examiner les propriétés d'échantillon fini de la procédure de prédiction. Nous appliquerons également la procédure à un ensemble de données provenant d'un essai clinique randomisé à plusieurs étapes pour le traitement du trouble dépressif majeur. Mots-clés : Données longitudinales, suivi informatif, vraisemblance H, prédiction.

[16:15-16:30]

Menelaos Konstantinidis (University of Toronto: Dalla Lana School of Public Health) **Lily S. H. Lim** (Children's Hospital Research Institute of Manitoba, University of Manitoba) **Eleanor M. Pullenayegum** (Dalla Lana School of Public Health, University of Toronto)

Designing an Accelerated Longitudinal Cohort for the Employment trajectories of Systemic Lupus Erythematosus Patients: A simulation Study

Conception de cohortes longitudinales accélérées pour des trajectoires d'emploi de patients atteints de lupus érythémateux systémique : une étude par simulations

Longitudinal studies are useful for examining a population over time. However, such studies can be restrictive (e.g. due to long follow-up times and high rates of at-

Les études longitudinales sont utiles pour l'examen d'un population au fil du temps. De telles études peuvent toutefois être restrictives (par exemple en raison des délais de suivi et des taux

trition). An alternative option is the Accelerated Longitudinal Cohort (ALC) – a design in which multiple overlapping cohorts are followed for shorter periods of time. We present a simulation study for the design of an ALC to study the employment trajectory of Systemic Lupus Erythematosus young-adult patients, analyzed by Multi-state models. The simulation seeks to identify design parameters (i.e., frequency of follow-up, length of follow-up per cohort, and overlap between cohorts), evaluated against the precision and bias of transition intensity matrices. Throughout, we consider how age, period, and cohort effects influence the study design. The present set of simulations will provide the first simulation of ALDs when analyzed using multistate models and provide a basis for future research.

élevés d'attrition). Une solution alternative est la conception de cohortes longitudinales accélérées (ALC) dans laquelle de multiples cohortes qui se chevauchent sont suivies sur des périodes plus courtes. Nous présentons une étude par simulations d'une conception ALC afin d'étudier la trajectoire d'emploi de jeunes adultes atteints de lupus érythémateux systémique, analysée à l'aide de modèles multi-états. Cette simulation vise à identifier les paramètres de la conception (c.-à-d. la fréquence des suivis, la durée de suivi par cohorte et le chevauchement entre cohortes), évaluée par rapport à la précision et au biais des matrices de taux de transition. Tout au long, nous considérons comment l'âge, la période et les effets de la cohorte influencent la conception de l'étude. Cet ensemble de simulations fournit la première simulation d'une conception ALC analysée à l'aide de modèles multi-états, en plus de fournir une base de recherche ultérieure.

[16:30-16:45]

Lulu Guo (Simon Fraser University - Burnaby, BC) **Hui Xie** (Faculty of Health Sciences, Simon Fraser University)

A Latent Variable Model for Longitudinal Trials with Multiple Endpoints and Treatment Noncompliance: an Application to a Study of Arthritis Health Journal

Modèle à variables latentes pour des essais longitudinaux en présence de multiples critères d'évaluation et de non-conformité au traitement : une application à une étude du Arthritis Health Journal

Noncompliance often occurs in longitudinal randomized controlled trials (RCTs). The complier average causal effect (CACE) informs the intervention efficacy for the subpopulation who would comply regardless of assigned treatment and has been considered as patient-oriented treatment effects of interest under noncompliance. Real-world RCTs often employ multiple study endpoints to measure treatment success. We focus on longitudinal randomized trial studies in which multiple endpoints are measured repeatedly over time. Under potential outcome framework, we proposed a latent variable model for longitudinal trials to deal with multiple outcomes and noncompliance. The model is a combination of factor analysis model and mixed effects regression models under principal stratification. The proposed approach is illustrated by evaluating the treatment efficacy of Arthritis Health Journal online tool.

La non-conformité se produit souvent dans les essais contrôlés randomisés longitudinaux (RCT). Les effets causaux moyens du facteur de conformité (CACE) renseignent sur l'efficacité d'une intervention pour une sous-population qui serait conforme quel que soit le traitement prescrit et qui sont réputés être les effets d'intérêt d'un traitement axé sur le patient sous la non-conformité. Nous nous attardons à des essais randomisés longitudinaux dans lesquels les multiples critères d'évaluation sont mesurés à répétition dans le temps. Sous un cadre d'issue potentielle, nous proposons un modèle à variables latentes pour des essais longitudinaux afin de composer avec des issues multiples et avec la non-conformité. Le modèle est une combinaison d'un modèle d'analyse factorielle et de modèles de régression à effets mixtes sous stratification principale. L'approche proposée est illustrée par une évaluation de l'efficacité du traitement de l'outil en ligne de l'Arthritis Health Journal.

[16:45-17:00]

Marzia Angela Cremona (Université Laval) **Huy Dang** (The Pennsylvania State University) **Francesca Chiaromonte** (The Pennsylvania State University)

smoothEM: A New Approach for the Simultaneous Assessment of Smooth Curves and Spikes

smoothEM : une nouvelle approche pour l'évaluation simultanée des courbes lisses et des pics

Many longitudinal data comprise both smooth and irregular elements. We consider scenarios in which an underlying smooth curve is composed not just with Gaussian

De nombreuses données longitudinales comprennent à la fois des éléments lisses et irréguliers. Nous considérons le cas d'une courbe lisse sous-jacente composée non seulement d'erreurs gaus-

errors, but also with irregular spikes that (a) are themselves of interest, and (b) can negatively affect our ability to characterize the underlying curve. We propose an approach that, combining regularized spline smoothing and an EM algorithm, allows to both identify spikes and estimate the smooth component. We prove the convergence of EM estimates to the true population parameters under some assumptions. Next, we demonstrate the performance of our method on finite samples and its robustness to assumptions violations through simulations. Finally, we apply it to the analysis of two time series on the annual heatwaves index in the US and on the weekly electricity consumption in Ireland. In both datasets, we are able to characterize underlying smooth trends and to pinpoint irregular/extreme behaviors.

siennes, mais aussi avec des pics irréguliers qui (a) sont eux-mêmes d'intérêt et (b) peuvent affecter négativement l'estimation de la courbe. Nous proposons une approche qui combine les splines de lissage et un algorithme EM pour à la fois identifier les pics et estimer la composante lisse. Nous prouvons la convergence des estimations EM vers les vrais paramètres de population sous certaines hypothèses. Ensuite, nous démontrons par simulation les performances de notre méthode sur des échantillons finis et sa robustesse aux violations d'hypothèses. Enfin, nous analysons deux séries temporelles sur l'indice de canicule aux États-Unis et sur la consommation d'électricité en Irlande. Dans les deux cas, nous caractérisons les formes lisses et identifions les comportements irréguliers/extrêmes.

Chair/Président: Ilia Sucholutsky

Date: Thursday June 2 / jeudi 2 juin

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-15:45]

Li Yi (University of Western Ontario)

How Self-Supervised Contrastive Learning Helps Learning with Label Noise

Comment l'apprentissage auto-supervisé contrastif aide à apprendre lorsque les étiquettes sont bruitées

Label noise is ubiquitous in real-world datasets. In this paper, we reveal that learning with label noise can benefit from representations learned by self-supervised learning (SSL) methods. By constructing a motivating example of instance-dependent label noise, we theoretically show that a classifier trained on representations learned by SSL performs significantly better than one trained by supervised learning. Then, we systematically study the benefits of the representations learned for learning with label noise. (1) The label noise uniformly spreads over the SSL representations. (2) The representations learned by SSL exhibit an intrinsic cluster structure with respect to true labels. We further theoretically justify the benefit of training a classifier on representations with such a cluster structure. It encourages the classifier trained on noisy data to be aligned with the optimal classifier. We conduct extensive experiments to show the effectiveness of SSL representations.

Les étiquettes bruitées sont omniprésentes dans les ensembles de données réelles. Dans cette présentation, nous montrons que l'apprentissage avec des étiquettes bruitées peut être amélioré par les représentations apprises par les méthodes d'apprentissage auto-supervisé. Grâce à un exemple d'étiquettes bruitées dépendant de l'instance, nous démontrons, sur le plan théorique, qu'un classificateur entraîné sur des représentations apprises par l'apprentissage auto-supervisé est nettement plus efficace qu'un classificateur entraîné par apprentissage supervisé. Ensuite, nous étudions de manière systématique les avantages des représentations apprises pour l'apprentissage avec des étiquettes bruitées. 1) Le bruit des étiquettes se répand uniformément sur les représentations d'apprentissage auto-supervisé. 2) Les représentations apprises par l'apprentissage auto-supervisé présentent une structure de groupe intrinsèque par rapport aux vraies étiquettes. De plus, nous démontrons de manière théorique l'avantage de former un classificateur sur des représentations avec une telle structure de groupe. Cela permet au classificateur entraîné sur des données bruitées de s'aligner sur le classificateur optimal. Nous menons des expériences approfondies pour montrer l'efficacité des représentations d'apprentissage auto-supervisé.

[15:45-16:00]

Cansu Alakus (HEC Montréal) **Denis Larocque** (HEC Montréal) **Aurélie Labbe** (HEC Montreal)

RFpredInterval: An R Package for Prediction Intervals with Random Forests and Boosted Forests

RFpredInterval : une bibliothèque R pour les intervalles de prévisions avec forêts aléatoires et forêts améliorées

Like many predictive models, random forests provide a point prediction for a new observation. Besides the point prediction, it is important to quantify the uncertainty in the prediction. Prediction intervals (PI) provide information about the reliability of the point predictions. We propose a new method to build Prediction Intervals with Boosted Forests (PIBF). PIBF provides bias-corrected point predictions obtained with the one-step boosted forest and prediction intervals by using the

Comme pour beaucoup de modèles prédictifs, les forêts aléatoires fournissent une prédiction d'un point pour une nouvelle observation. Outre la prédiction ponctuelle, il est important de quantifier l'incertitude de la prédiction. Les intervalles de prédiction (IP) fournissent des renseignements sur la fiabilité des prédictions ponctuelles. Nous proposons une nouvelle méthode de construction d'intervalles de prédiction avec forêts améliorées (IPFA). Les IPFA fournissent des prédictions ponctuelles corrigées du biais obtenu avec les forêts améliorées en une étape, ainsi que des inter-

nearest neighbor out-of-bag observations to estimate the conditional prediction error distribution. The proposed method is investigated and compared to ten existing methods to build PIs with random forests through simulation studies and real data analyses. We have developed a comprehensive R package, RFpredInterval, that integrates 16 methods to build PIs with random forests and boosted forests, including PIBF and 15 different variants to produce PIs with random forests proposed by Roy and Larocque (2020).

valles de prédiction au moyen des observations hors-sac (≪out-of-bag≫) du voisin le plus proche pour estimer la distribution d'erreur de prédiction conditionnelle. La méthode proposée est étudiée et comparée à dix autres méthodes existantes servant à construire des IP avec forêts aléatoires à partir d'études en simulation et d'analyse de données réelles. Nous avons conçu une bibliothèque R complète (RFpredInterval) comprenant 16 méthodes pour construire des IP avec des forêts aléatoires et des forêts améliorées, y compris l'IPFA et 15 variantes différentes servant à produire des IP avec forêts aléatoires proposées par Roy et Laroque (2020).

[16:00-16:15]

Leslie G. Fell (University of Guelph) **Olaf Berke** (University of Guelph) **Lorna E. Deeth** (University of Guelph) **Lise A. Trotz-Williams** (Wellington-Dufferin-Guelph Public Health)

Predicting Bacterial Contamination of Private Well Water in Wellington-Dufferin-Guelph, Ontario

Prédiction de contamination bactérienne de l'eau de puits privé à Wellington-Dufferin-Guelph, en Ontario

One in ten Canadians rely on private well water as their primary source for drinking water. Bacterial contamination of well water poses a serious public health issue. The risk of bacterial contamination can be exacerbated by numerous factors, including physical well characteristics and the surrounding hydrogeology. Classification methods, including a predictive logistic regression model for bacterial contamination, were developed for private wells within the Wellington-Dufferin-Guelph Public Health region of Southern Ontario. Predictive accuracy was compared, and the identification of the predominant risk factors after variable selection was explored. Results from this study will allow public health to develop targeted educational campaigns and reduce the burden of water-borne diseases within the community.

Un Canadien sur dix dépend de l'eau d'un puits privé comme source principale d'eau potable. La contamination bactérienne de l'eau de puits constitue un sérieux problème de santé publique. Le risque de contamination bactérienne est amplifié par de nombreux facteurs, y compris les caractéristiques physiques du puits et l'environnement hydrogéologique. Des méthodes de classification, y compris un modèle de régression logistique pour la contamination bactérienne, ont été élaborées pour les puits privés de la région de la Santé publique de Wellington-Dufferin-Guelph, dans le sud de l'Ontario. L'exactitude prédictive a été comparée et l'identification des facteurs de risque prédominants a été explorée après une sélection de variables. Les résultats de cette étude permettront à la Santé publique de mettre au point des campagnes éducatives ciblées et de réduire le fardeau des maladies transmissibles par l'eau dans la communauté.

[16:15-16:30]

Henrik Stryhn (University of Prince Edward Island)

A Random Effects Model for Sparse Cross-Classification Data

Modèle à effets aléatoires pour données éparses de classification croisée

We study the application of a linear mixed model with crossed random effects to determine a ranking order for abstracts submitted to a scientific conference, each scored on a 0-100 scale by two reviewers. Simple averaging of scores across reviewers may seem plausible and easy to communicate, but does not account for reviewer effects. Conversely, within-reviewer standardization may introduce biases in the abstract scores and ranks. Model-based ranking relies on assumptions, e.g. normality of random effects, that allow standard ML or

Nous étudions l'application d'un modèle mixte linéaire à effets aléatoires croisés pour déterminer un ordre de classement des résumés soumis à une conférence scientifique, chacun étant noté sur une échelle de 0 à 100 par deux évaluateurs. La simple moyenne des notes entre les évaluateurs peut sembler plausible et facile à communiquer, mais elle ne tient pas compte des effets des évaluateurs. Inversement, la normalisation au sein d'un même évaluateur peut introduire des biais dans les notes et les classements des résumés. Le classement basé sur un modèle repose sur des hypothèses, par exemple la normalité des effets

MCMC estimation algorithms to smooth out any design singularities, which it may be important to be aware of. We compare the results and performance (by simulation) of these methods on real data comprising 119 abstracts and 27 reviewers. Variation among reviewers amounted to 36% of the total variance. Model-based ranking and simple average ranking (after 15 supplementary reviews were acquired) classified 16/119 abstracts differently at the desired cut-off.

aléatoires, qui permettent aux algorithmes classiques d'estimation ML ou MCMC d'atténuer toute singularité de conception, dont il faut peut-être être conscient. Nous comparons les résultats et les performances (par simulation) de ces méthodes sur des données réelles comprenant 119 résumés et 27 évaluateurs. La variation entre les évaluateurs représentait 36 % de la variance totale. Le classement basé sur un modèle et le simple classement moyen (après acquisition de 15 révisions supplémentaires) ont classé 16/119 résumés différemment au seuil souhaité.

[16:30-16:45]

Alessandro Maria Maria Selvitella (Purdue University Fort Wayne) **Kathleen Lois Foster** (Department of Biology - Ball State University)

Anolis Ecomorph Biomechanics across Arboreal Environments: What can Machine Learning tell us about Behavioral Plasticity in Lizards?

Biomécanique des écomorphes d'Anolis dans les environnements arboricoles : que peut nous apprendre l'apprentissage automatique sur la plasticité comportementale des lézards ?

Arboreal animals must learn to modulate their movements to overcome the challenges posed by the complexity of their heterogeneous environment. Anolis lizards are remarkable in the apparent ease with which they conquer this heterogeneity. Significant progress has been made towards understanding the impact of substrate structure on the behavioral plasticity of arboreal species, but it is unclear whether the same strategies are common across ecomorphs. Our approach is to leverage machine learning methods to analyze 3D limb kinematics of 6 Puerto Rican Anolis species running on different surfaces. By comparing the prediction accuracies of models trained on specific ecomorphs and tested on new ecomorphs, we can gain insight into which locomotor strategies are universally useful, and thus broadly transferred to new conditions in all species, and which strategies are unique to particular ecomorphs. This is a joint work with Kathleen Lois Foster (Department of Biology, Ball State University).

Les animaux arboricoles doivent apprendre à moduler leurs mouvements pour surmonter les défis posés par la complexité de leur environnement hétérogène. Les lézards Anolis sont remarquables par la facilité apparente avec laquelle ils surmontent cette hétérogénéité. Des progrès significatifs ont été réalisés pour comprendre l'impact de la structure du substrat sur la plasticité comportementale des espèces arboricoles, mais il n'est pas clair si les mêmes stratégies sont communes à tous les écomorphes. Notre approche consiste à utiliser des méthodes d'apprentissage automatique pour analyser la cinématique des membres en 3D de 6 espèces d'Anolis portoricains qui courent sur différentes surfaces. En comparant la précision des prédictions des modèles ajustés sur des écomorphes spécifiques et testés sur de nouveaux écomorphes, nous pouvons mieux comprendre quelles stratégies locomotrices sont universellement utiles, et donc largement transférées à de nouvelles conditions chez toutes les espèces, et quelles stratégies sont uniques à des écomorphes particuliers. Il s'agit de travaux conjoints avec Kathleen Lois Foster (Département de biologie de l'Université Ball State).

[16:45-17:00]

Yunfeng Yang (University of Waterloo)

Multimodel Bayesian Analysis of Load Duration Effects in Lumber Reliability

Analyse bayésienne multimodèle des effets de la durée de chargement dans la fiabilité du bois d'œuvre

The strength of lumber products may change over time if applied to stress. This is known as the duration-of-load (DOL) effect, which is considered in ensuring the long-term reliability of wood structures. In this talk, we use a multimodel Bayesian approach to perform reliability analysis of lumber and account for the DOL effect under different load profiles. The generating mech-

La résistance des produits en bois peut changer avec le temps s'ils sont soumis à des contraintes. Ce phénomène est connu sous le nom d'effet de durée de chargement (DC), qui est pris en compte pour assurer la fiabilité à long terme des structures en bois. Dans cet exposé, nous utilisons une approche bayésienne multimodèle pour effectuer une analyse de fiabilité du bois d'œuvre et tenir compte de l'effet de DC sous différents profils de charge. Les

anisms of residential load, snow load, and wind load are presented. Also, three individual DOL models are considered: the US model, the Canadian model, and the Gamma process model in the analysis. We propose Bayesian model-average (BMA) to combine the reliability estimates from these individual models under a given load profile. The BMA results are illustrated with its 95% credible intervals under the real data analysis of Hemlock experimental dataset.

mécanismes de génération de la charge résidentielle, de la charge de neige et de la charge de vent sont présentés. Nous examinons trois modèles de DC dans l'analyse : le modèle américain, le modèle canadien et le modèle de processus Gamma. Nous proposons la méthode du modèle moyen bayésien (BMA) pour combiner les estimations de fiabilité de ces modèles individuels sous un profil de charge donné. Nous illustrons les résultats du BMA avec ses IC à 95 % sur une analyse de données réelles de l'ensemble de données expérimentales Hemlock.

Chair/Président: You Liang

Organizer/Responsable: You Liang

Date: Friday June 3 / vendredi 3 juin

Time/Heure: 11:00-12:30

Abstract/Résumé

[11:00-11:30]

Wenqing He (University of Western Ontario) **Juan Xiong** (Shengzhen University)

Identification of Survival Relevant Genes with Measurement Error in Gene Expression Incorporated

Identification de gène pertinent de survie avec erreur de mesure dans l'expression génique intégrée

Modern gene expression technologies enable simultaneous measurement of thousands of genes, thus are important for predicting patient survival. However, survival analysis with gene expression data is challenging due to the high dimensionality. Proper identification of survival relevant genes is imperative for building suitable prediction models. Gene expressions are typically subject to measurement errors introduced from the complex experimental procedure and the measurement error is often ignored. In this talk, the effect of measurement error on the identification of survival relevant genes is explored under the accelerated failure time model. Survival relevant genes are identified by regularizing the weighted least square estimator with the adaptive LASSO penalty. The simulation-extrapolation method is applied to adjust for the impact of measurement error. The performance of the proposed method is assessed by simulation studies and illustrated by a real study.

Les technologies modernes d'expression génique permettent de mesurer simultanément des milliers de gènes, et sont donc importantes pour prédire la survie d'un patient. Cependant, l'analyse de survie avec des données d'expression génique représente un défi en raison de sa haute dimension. Il est impératif d'identifier adéquatement les gènes pertinents de survie pour construire des modèles de prédiction convenables. Les expressions géniques sont généralement sujettes à des erreurs de mesure générées par la procédure expérimentale complexe, mais on en tient pas souvent compte. Lors de cet exposé, nous abordons l'effet qu'a l'erreur de mesure sur l'identification des gènes pertinents de survie selon le modèle à temps d'échec accéléré. On repère les gènes pertinents de survie en régularisant l'estimateur du moindre carré pondéré avec la pénalité adaptative LASSO. Nous appliquons la méthode de simulation-extrapolation pour ajuster en fonction de l'influence de l'erreur de mesure. La performance de la méthode proposée est évaluée par des études en simulation et démontrée par une étude réelle.

[11:30-12:00]

Xuekui Zhang (University of Victoria)

Automated Cell-Type Annotation using scRNA-seq Data

Annotation automatique des types de cellules à l'aide de données de séquençage de l'ARN en cellule unique

I present a novel analysis pipeline that automatically annotates the cell types using single-cell RNA-seq data. To evaluate the performance of our method, we use multiple popular benchmark data to compare the performance of our method to popular competitors in the literature.

Je présente un nouveau système d'analyse qui permet d'annoter automatiquement les types de cellules à l'aide de données de séquençage de l'ARN en cellule unique. Pour évaluer les résultats de cette méthode, j'utilise plusieurs données de référence populaires afin de comparer les résultats de notre méthode à ceux de méthodes concurrentes populaires de la littérature.

[12:00-12:30]

Liangliang Wang (Simon Fraser University) **Shijia Wang** (Nankai University) **Alexandre Bouchard-Côté** (University of British Columbia)

Efficient Sequential Monte Carlo Methods for Bayesian Phylogenetic Inference
Méthodes de Monte Carlo séquentielles efficaces pour l'inférence phylogénétique bayésienne

In Bayesian phylogenetics, the goal is to approximate a posterior distribution of phylogenetic trees based on biological data. Standard Bayesian estimation of phylogenetic trees can handle rich evolutionary models but requires expensive Markov chain Monte Carlo (MCMC) simulations, which may suffer from the curse of dimensionality and the local-trap problem. Previous work has shown that the sequential Monte Carlo (SMC) method can serve as an excellent alternative to MCMC in posterior inference. In this talk, I will talk about our SMC methods for Bayesian Phylogenetic Inference and illustrate them using simulation studies and real data analysis.

En phylogénétique bayésienne, l'objectif est d'approximer une distribution a posteriori des arbres phylogénétiques à partir de données biologiques. L'estimation bayésienne standard des arbres phylogénétiques peut s'appliquer à des modèles d'évolution riches mais nécessite des simulations de Monte Carlo par chaîne de Markov (MCMC) coûteuses, qui peuvent souffrir du fléau de la dimension et du problème de piège local. Des travaux antérieurs ont montré que la méthode de Monte Carlo séquentielle (SMC) peut servir d'excellente alternative à la MCMC dans l'inférence a posteriori. Dans cet exposé, je parlerai de nos méthodes SMC pour l'inférence phylogénétique bayésienne et les illustrerai à l'aide d'études de simulation et d'analyses de données réelles.

Chair/Président: Rhonda J Rosychuk

Organizer/Responsable: Melody Ghahramani

Date: Friday June 3 / vendredi 3 juin

Time/Heure: 11:00-12:30

Abstract/Résumé

[11:00-12:00]

David R. Bellhouse (University of Western Ontario) **Christian Genest** (McGill University)

A Glimpse into SSC History

Un aperçu de l'histoire de la SSC

"What's past is prologue" says Shakespeare in The Tempest. We apply this quotation to the evolution of the SSC from its gestation to the present day. The talk will consist of three parts: (1) Before 1972, the Canadian statistical community was small, diverse and diffuse across the country; it was active in academia, the private sector, and the public sector. (2) The precursor to the SSC was formed in 1972 in Montreal by a small group of statisticians which focussed on producing a new publication, The Canadian Journal of Statistics. Until 1977, there was significant turmoil in the statistical community over the new society, which was only resolved when groups put aside their differences to form the SSC. (3) One of the major activities of the SSC has been to expand beyond the academic community with initiatives such as professional accreditation and the formation of various special interest sections.

« Le passé n'est qu'un prologue », écrit Shakespeare dans La Tempête. Nous appliquons cette citation à l'évolution de la SSC depuis sa gestation jusqu'à nos jours. L'exposé s'articulera en trois temps : (1) Avant 1972, la communauté statistique canadienne était petite, disparate et éparse à travers le pays ; elle œuvrait dans les secteurs privé, public et académique. (2) L'ancêtre de la SSC a été créé en 1972 à Montréal par quelques statisticiens qui ont mis sur pied une nouvelle publication, La revue canadienne de statistique. Jusqu'en 1977, la communauté statistique était divisée au sujet de la nouvelle association et l'unité ne fut retrouvée que lorsque les protagonistes mirent de côté leurs différences pour former la SSC. (3) Entre autres initiatives, la SSC a cherché depuis lors à s'étendre au-delà du milieu universitaire avec des initiatives telles que l'accréditation professionnelle et la création de divers groupes d'intérêt.

Chair/Président: Joanna Elizabeth Mills Flemming

Organizer/Responsable: Joanna Elizabeth Mills Flemming

Date: Friday June 3 / vendredi 3 juin

Time/Heure: 11:00-12:30

Abstract/Résumé

[11:00-11:30]

Jonathan Babyn (Dalhousie University)

Estimating both Population Effective and Census Size using Close-Kin Mark Recapture

Estimation de la taille effective de la population et de la taille de la population recensée à l'aide du marquage-recapture d'espèces qui ont un lien de parenté proche

Close-kin Mark Recapture (CKMR) extends Mark Recapture (MR) by replacing physical tags with genetic ones. Finding close-kin (such as parents, siblings and half-siblings) of the "tagged" individuals in future sampling of individuals both alive and dead acts as recapture events allowing for estimating the census population size (N_c). However CKMR models require the life history of the species (e.g. fecundity, breeding cycle, etc.) which complicates their implementation. Population Effective size (N_e) gives the equivalent number of breeders in a population assuming an ideal Wright-Fisher model (random mating, discrete generations, etc.). N_e can help provide a sense of the genetic health of a population. This presentation explores recent work on constructing a CKMR model to provide both estimates of N_c and N_e from the same model and data.

Le marquage-recapture d'espèces qui ont un lien de parenté proche est une extension de marquage-recapture qui remplace les marqueurs physiques par des marqueurs génétiques. La découverte d'un lien de parenté proche (parents, frères et sœurs, et demi-frères et sœurs) entre les individus ≪ marqués ≫ lors d'un échantillonnage ultérieur d'individus vivants et morts constitue un événement de recapture permettant d'estimer la taille de la population recensée (N_c). Cependant, les modèles de recapture-marquage des espèces ayant un lien de parenté proche requièrent le cycle de vie de l'espèce (par exemple, la fécondité et le cycle de reproduction), ce qui complique leur mise en œuvre. La taille effective de la population (N_e) donne le nombre équivalent de reproducteurs dans une population dans l'hypothèse d'un modèle idéal de Wright-Fisher (accouplement aléatoire, générations discrètes, etc.). La taille effective de la population peut aider à se faire une idée de la santé génétique d'une population. Cette présentation porte sur les récents travaux de conception d'un modèle de marquage-recapture d'espèces qui ont un lien de parenté proche afin de fournir à la fois des estimations de la taille de la population recensée et de la taille effective de la population à partir du même modèle et des mêmes données.

[11:30-12:00]

Ethan Lawler (Dalhousie University) **Chris Field** (Dalhousie University) **Joanna Mills Flemming** (Dalhousie University)

Species Distribution Modelling using Spatio-temporal Nearest Neighbour Gaussian Processes

Modélisation de la distribution des espèces à l'aide de processus gaussiens spatio-temporels du plus proche voisin

We develop a particular generalized linear mixed model for spatio-temporal point-referenced data that is flexible enough to accommodate data from most ecological surveys while being structured enough to facilitate analyses without advanced coding. Our implementation in the staRVe package uses a spatio-temporal version of a

Nous développons un modèle mixte linéaire généralisé particulier pour des données spatio-temporelles référencées par points qui est suffisamment flexible pour s'adapter aux données de la plupart des enquêtes écologiques tout en étant suffisamment structuré pour faciliter les analyses sans codage avancé. Notre implémentation dans le package staRVe utilise une version spatio-temporelle d'un

nearest neighbour Gaussian process enabling analysis of relatively large datasets. We introduce our modelling framework and present an example analysis using the staRVe package. We also discuss how the model can be generalized to a multivariate spatio-temporal model using copulas.

processus gaussien du plus proche voisin permettant l'analyse d'ensembles de données relativement importants. Nous introduisons notre cadre de modélisation et présentons un exemple d'analyse utilisant le package staRVe. Nous discutons également de la manière dont le modèle peut être généralisé à un modèle spatio-temporel multivarié en utilisant des copules.

[12:00-12:30]

Andrea Perreault (Fisheries and Oceans Canada) **Noel Cadigan** (Fisheries and Marine Institute of Memorial University)

Profile Likelihood Diagnostics for Integrated State-Space Models

Diagnostics du profile de vraisemblance pour les modèles d'espace d'états intégrés

State-space models (SSM) that separately include process and observation errors are often used to analyze ecological time-series data; however, diagnostics such as likelihood profiles are complicated for SSM due to the dependencies in the data caused by process errors. Often of interest are models that integrate many data sources, and understanding the contribution of each data source to the total likelihood is useful. In the non-state-space integrated setting, it is straightforward to construct profile likelihood plots that show the likelihood contributions of each data source versus changes in a model input, since the total negative log likelihood (nll) is simply the sum of the nll of each data source. SSM are often estimated using the marginal nll to account for random effects; however, the marginal nll cannot be directly split into data source nlls. This research develops a novel approach to provide profile nll contributions for SSM and is illustrated using a fisheries case study.

Les modèles d'espace d'état (MEE) qui comprennent séparément les erreurs de procédé et d'observation sont souvent employés pour l'analyse de données de séries temporelles écologiques. Toutefois, les diagnostics tels que les profiles de vraisemblance sont trop complexes pour les MEE en raison des dépendances dans les données causées par les erreurs de procédé. Les modèles intégrants plusieurs sources de données sont généralement pratiques, et la compréhension de la contribution de chaque source de données à la vraisemblance totale est utile. Dans le cadre intégré sans espace d'états, il est simple de construire un graphique du profile de vraisemblance qui présente les contributions de vraisemblance de chaque source de données par rapport aux changements dans une entrée de modèle, puisque le logarithme du rapport de vraisemblance négatif (nll) total est simplement la somme du nll de chaque source de données. Les MEE sont souvent estimés à l'aide de nll marginaux pour tenir compte d'effets aléatoires, cependant les nll marginaux ne peuvent pas être directement séparés en nll de sources de données. Cette recherche élabore une nouvelle approche pour obtenir les contributions des nll profilés pour les MEE et est illustrée à partir d'une étude de cas sur les établissements piscicoles.

2022 Pierre Robillard Award Address
Allocution du récipiendaire du prix Pierre-Robillard 2022

Chair/Président: Yingwei (Paul) Peng

Date: Friday June 3 / vendredi 3 juin

Time/Heure: 11:00-12:30

Abstract/Résumé

[11:00-12:00]

Janie Coulombe (McGill University)

Causal inference on the marginal effect of an exposure: Addressing biases due to covariate-driven monitoring times and confounders

Inférence causale sur l'effet marginal d'une exposition : Comment tenir compte des biais dus aux temps de visite qui dépendent du patient et aux facteurs confondants

Causal inference focuses on the estimation of effects due to specific, well-defined causes (like exposures on which we can intervene). Health data are collected abundantly, which provides a rich landscape for research on causal inference. However, the collection of these data does not always rely on a study design made expressly for answering a causal question. Consequently, longitudinal observational data from medical health records are filled with biasing associations that can affect inference. We focus on two such types of associations, the confounding bias, and the bias due to covariate-driven monitoring times, in the inference on the causal marginal effect of an exposure on a longitudinal outcome. Using causal diagrams, we describe how these biases can arise and how to account for them. Two novel estimators are proposed and demonstrated in extensive simulation studies, and asymptotic theory is developed. Two extensions are further proposed for more complex data scenarios.

L'inférence causale s'intéresse à l'estimation d'effets dus à des causes bien définies, comme une exposition sur laquelle intervenir. Les données de santé, comme celles des dossiers médicaux électroniques, sont collectées de façon abondante et offrent un large éventail pour la recherche en inférence causale. Cependant, leur collecte dépend rarement d'un plan d'expérience conçu expressément pour répondre à une question causale. Ainsi, ces données sont remplies d'associations trompeuses pouvant affecter l'inférence. On se concentre ici sur deux types d'associations, le biais de confusion et le biais dû aux temps de visite dépendant des variables du patient, dans l'inférence sur l'effet causal marginal d'une exposition sur une issue longitudinale. À partir de diagrammes causaux, on décrit les mécanismes de biais et la façon d'en tenir compte. Deux estimateurs sont proposés et démontrés à partir de simulations et deux extensions sont proposées pour des scénarios de données plus complexes.

Chair/Président: Xiaowen Zhou, Shui Feng

Organizer/Responsable: Xiaowen Zhou

Date: Friday June 3 / vendredi 3 juin

Time/Heure: 11:00-12:30

Abstract/Résumé

[11:00-11:30]

Shui Feng (McMaster University)

Kingman Coalescent and Bayesian Nonparametrics

Coalescent de Kingman et approche bayésienne non paramétrique

Consider a population whose genetic composition evolves in time according to the class of Fleming-Viot processes with parent independent mutation. Ancestral inference on the population can be done through Kingman coalescent based on random samples taking at each fixed time. An alternate method is through the Bayesian nonparametric predictive approach. In this talk we will discuss the relation between the Bayesian nonparametric analysis and Kingman coalescent. This is based on a joint work with Stefano Favaro and Paul Jenkins.

Nous examinons une population dont la composition génétique évolue dans le temps selon la classe des processus de Fleming-Viot avec mutation indépendante des parents. L'inférence ancestrale sur la population peut être faite par le coalescent de Kingman, qui repose sur des échantillons aléatoires prélevés à chaque temps fixe. Une autre méthode consiste à utiliser l'approche prédictive bayésienne non paramétrique. Dans cette présentation, nous traitons de la relation entre l'analyse non paramétrique bayésienne et le coalescent de Kingman. Ces travaux ont été réalisés en collaboration avec Stefano Favaro et Paul Jenkins.

[11:30-12:00]

Lam Ho (Dalhousie University)

Theory of Ancestral State Reconstruction

Théorie de reconstruction d'état ancestral

Ancestral state reconstruction is one of the most important tasks in evolutionary biology. Reconstructing the trait value at the root of a phylogenetic tree can provide answers to many macroevolution questions, such as the origin of an epidemic. Conditions under which we can reliably reconstruct the ancestral state have been studied for both discrete and continuous traits. However, the connection between these results is unclear, and it seems that each model needs different conditions. In this talk, I will discuss a unifying theory on the existence of a consistent ancestral state reconstruction method.

La reconstruction d'état ancestral est l'une des tâches les plus importantes en biologie de l'évolution. Reconstruire la valeur de trait à la racine d'un arbre phylogénétique peut offrir des réponses à plusieurs questions relatives à la macroévolution, comme l'origine d'une épidémie. Les conditions selon lesquelles on peut reconstruire de façon fiable l'état ancestral ont été étudiées pour les traits continus et discrets. Toutefois, le lien entre ses résultats reste ambigu, et il semblerait que chaque modèle demande des conditions différentes. Lors de cet exposé, je discuterai d'une théorie unificatrice sur l'existence d'une méthode de reconstruction d'état ancestral cohérente.

[12:00-12:30]

Xiaowen Zhou (Concordia University)

Continuous-state Nonlinear Branching Processes

Processus de branchement non linéaires à l'état continu

We consider a class of continuous-state branching processes whose branching rates depend on the current

Nous examinons une classe de processus de branchement à l'état continu dont les taux de branchement dépendent de la taille ac-

population sizes. They are nonnegative-valued Markov processes that can be obtained from spectrally positive Lévy processes via Lamperti type time changes. The non-additive branching mechanism allows the processes to have exotic behaviors such as coming down from infinity but at the same time requires new techniques in their study. In this talk we are going to introduce recent progresses on coming down from infinity, explosion and extinguishing behaviors for such processes. It is based on joint work with Clement Foucart, Bo Li, Junping Li, Pei-Sen Li and Yingchun Tang.

tuelle de la population. Il s'agit des processus de Markov à valeurs non négatives qui peuvent être obtenus à partir des processus de Lévy spectralement positifs à l'aide de changements temporels de type Lamperti. Le mécanisme de branchement non-additif permet aux processus d'avoir des comportements extrêmes (descente de l'infini), mais nécessite en même temps de nouvelles techniques pour les étudier. Dans cette présentation, nous présentons les progrès récents sur les comportements de descente de l'infini, d'explosion et d'extinction de ces processus. Cette étude repose sur des travaux conjoints avec Clément Foucart, Bo Li, Junping Li, Pei-Sen Li et Yingchun Tang.

Chair/Président: Utkarsh J. Dang

Date: Friday June 3 / vendredi 3 juin

Time/Heure: 11:00-12:30

Abstract/Résumé

[11:00-11:15]

Andrea Payne (Carleton University) **Anjali Silva** (University of Toronto) **Steven Rothstein** (University of Guelph) **Paul D. McNicholas** (McMaster University) **Sanjeena Dang (Subedi)** (Carleton University)

Clustering High Dimensional Multivariate Count Data Using a Family of Mixtures of Multivariate Poisson Log-Normal Distributions

Regroupement de données de dénombrement multivariées à haute dimension à l'aide d'une famille de mélanges de distributions log-normales multivariées de Poisson

Multivariate count data encountered in bioinformatics are high dimensional and often exhibit over-dispersion. Mixtures of multivariate Poisson lognormal (MPLN) models have been used to analyze these multivariate count measurements efficiently. In the MPLN model, the counts, conditional on the latent variable, are modeled using a Poisson distribution and the latent variable comes from a multivariate Gaussian distribution. Due to this hierarchical structure, the MPLN model can account for over-dispersion and allows for correlation between the variables. Here, we extend the mixture of multivariate Poisson-log normal distributions for high dimensional data by incorporating a factor analyzer structure in the latent space. A parsimonious family of mixtures of Poisson log-normal distributions are proposed by decomposing the covariance matrix and imposing constraints on these decompositions. We demonstrate the performance of the model using simulated and real datasets.

Les données de dénombrement multivariées en bio-informatique sont à haute dimension et présentent souvent une surdispersion. Les modèles de mélanges de multivariés log-normales de Poisson (MPLN) sont utilisés pour l'analyse efficace des mesures de dénombrement multivariées. Dans le modèle MPLN, les dénombrements, conditionnellement à la variable latente, sont modélisés à l'aide d'une distribution de Poisson et la variable latente provient d'une distribution gaussienne multivariée. En raison de cette structure hiérarchique, le modèle MPLN peut prendre en compte la surdispersion et permet la corrélation entre les variables. Nous étendons ici le mélange de distributions log-normales multivariées de Poisson pour les données à haute dimension en incorporant à l'espace latent une structure d'analyse des facteurs. Une famille clairsemée de mélanges de distributions log-normales de Poisson est proposée en décomposant la matrice de covariance et en imposant des contraintes à ces décompositions. Nous illustrons la performance du modèle à l'aide d'ensembles de données simulées et réelles.

[11:15-11:30]

Zahra Aghahosseinalishirazi (Western University) **Dr Camila De Souza** (The University of Western Ontario)

Clustering Single-Cell RNA Sequencing Data via the Expectation-Maximization Algorithm

Regroupement des données de séquençage de l'ARN de cellules uniques avec l'algorithme espérance-maximisation

Cells are the essential units in biology and can be distinguished by their phenotype, such as size and shape, or at the molecular level, based on their genome, epigenome, and transcriptome. In this thesis, we focus on the transcriptome, which includes all RNA transcripts present in a given cell population indicating the genes that are being expressed at a certain time. We consider single-cell RNA sequencing data and propose a novel model-based

Les cellules sont des unités essentielles en biologie et peuvent se distinguer par leur phénotype, comme la taille et la forme, ou sur le plan moléculaire, en fonction de leur génome, de leur épigénome et de leur transcriptome. Dans cette présentation, nous nous penchons sur le transcriptome, qui comprend toutes les transcriptions d'ARN présentes dans une population cellulaire donnée, et qui indique les gènes qui sont exprimés à un moment donné. Nous examinons les données de séquençage de l'ARN d'une seule cellule

clustering method to group cells based on their transcriptome profiles. The proposed clustering approach takes into account the large proportion of zeros present in the data, which can be either true biological zeros or technological noise. The assumed model for clustering is a mixture of either zero-inflated Poisson or zero-inflated negative binomial distributions, and inference is conducted via the EM algorithm. The performance of proposed methodology will be evaluated via simulation studies and analyses of published real datasets.

et proposons une nouvelle méthode de regroupement fondée sur un modèle visant à regrouper les cellules en fonction de leur profil transcriptomique. L'approche de regroupement proposée tient compte de la grande proportion de zéros présents dans les données, qui peuvent être soit de véritables zéros biologiques, soit du bruit technologique. Le modèle adopté pour le regroupement est un mélange de lois de Poisson avec excès de zéros ou de lois binomiales négatives avec excès de zéros, et l'inférence est effectuée par l'algorithme EM. Nous évaluons les résultats de la méthodologie proposée au moyen d'études de simulation et d'analyses d'ensembles de données réelles publiées.

[11:30-11:45]

Ashani N. Wickramasinghe (University of Manitoba) **Saman Muthukumarana** (University of Manitoba) **Dan Loewen** (ioAirFlow) **Matt Schaubroeck** (ioAirFlow)

Temperature Clusters in Commercial Buildings Using K-means and Time Series Clustering

Regroupement de températures dans les bâtiments commerciaux à l'aide de K-moyennes et de séries chronologiques

An efficient building should be able to control its internal temperature in a manner that considers both the building's energy efficiency and the comfort level of its occupants. Thermostats help to control the temperature within a building and identifying proper thermostat placement helps to improve efficiency. To determine the minimum number of thermostats required to accurately measure the internal temperature distribution of a building, it is necessary to find the locations that show similar environmental conditions. In this study, we analyzed high resolution temperature measurements from a commercial building to assess the performance of the building's HVAC zoning system. Then we conducted two cluster analyses to evaluate the efficiency of the existing zoning structure. K-means and time series clustering was used to identify the clusters per building floor. Based on statistical assessments, we observed that time series clustering showed better results than k-means clustering.

Un bâtiment efficace doit pouvoir contrôler sa température interne d'une manière qui tient compte à la fois de l'efficacité énergétique du bâtiment et du niveau de confort de ses occupants. Les thermostats aident à contrôler la température à l'intérieur d'un bâtiment et l'identification du bon emplacement des thermostats aide à améliorer l'efficacité. Pour déterminer le nombre minimum de thermostats requis pour mesurer avec précision la distribution de la température interne d'un bâtiment, il est nécessaire de trouver les emplacements qui présentent des conditions environnementales similaires. Dans cette étude, nous avons analysé des mesures de température haute résolution d'un bâtiment commercial afin d'évaluer les performances du système de zonage HVAC du bâtiment. Nous avons ensuite effectué deux analyses de regroupement pour évaluer l'efficacité de la structure de zonage existante. Nous avons utilisé la méthode des K-moyennes et le regroupement des séries temporelles pour identifier des grappes par étage du bâtiment. Sur la base d'évaluations statistiques, nous avons observé que le regroupement des séries temporelles présentait de meilleurs résultats que la méthode des k- moyennes.

[11:45-12:00]

Michelle Wu (University of Toronto) **Hyejung Jung** (University of Toronto)

Correlation Analysis and Machine Learning-Based Approaches to Assess Depression Severity

Analyse de corrélation et approches basées sur l'apprentissage machine pour évaluer la gravité de la dépression

There exists quantitative ways to measure depression via questionnaires and assessments, evaluated either by physicians or completed by the individual themselves. Literature suggests though that patient responses to questions can vary from those administered by physicians, proposing that these quantitative instruments are subject to interpretation. Using parametric correlation

Il existe des moyens quantitatifs de mesurer la dépression par le biais de questionnaires et d'examens, évalués soit par des médecins, soit remplis par la personne elle-même. La littérature suggère cependant que les réponses des patients aux questions peuvent varier de celles administrées par les médecins, ce qui laisse entendre que ces instruments quantitatifs sont sujets à interprétation. En utilisant des méthodes de corrélation pa-

methods, we will compare and cluster scores of physician vs patient administered tests and present them via graphical methods. By examining the difference in questionnaires scores over time while considering the patient's mental health status and other biological factors, we will look for discernible characteristics that may be influencing survey results. Later, feature selection methods will be utilized to detect specific survey questions that may be stronger indicators of depression. The intention is to build a basis for re-weighing current questions to better capture symptoms for depression diagnosis.

ramétrique, nous comparerons et regrouperons les scores des tests administrés par les médecins par rapport à ceux autoadministrés par les patients et nous les présenterons par le biais de méthodes graphiques. En examinant la différence des scores des questionnaires au fil du temps tout en tenant compte de l'état de santé mentale du patient et d'autres facteurs biologiques, nous rechercherons des caractéristiques discernables qui pourraient influencer les résultats des enquêtes. Plus tard, nous utiliserons des méthodes de sélection de caractéristiques pour détecter des questions d'enquête spécifiques qui pourraient être des indicateurs plus forts de la dépression. L'intention est d'établir une base pour repondérer les questions actuelles afin de mieux saisir les symptômes contribuant à un diagnostic de dépression.

[12:00-12:15]

Michael John Ilagan (McGill University) **Carl F. Falk** (McGill University)

Supervised Components, Unsupervised Mixing Proportions: Detection of Bots in Likert-type Surveys

Composantes supervisées, proportions du mélange non supervisé : la détection des bots dans les enquêtes de type Likert

To detect bot responses in Likert-type survey data, researchers typically take a training sample, stratified by class: to produce human examples, the survey is administered to verified humans; but to produce bot examples, random data are simulated. On one hand, because the training prevalence of bots is arbitrary, conventional supervised classifiers fail. In the limit, as bots approach 100% of the sample, every observation is indiscriminately classified as bot. On the other hand, fully unsupervised solutions, such as the expectation-maximization algorithm, ignore valuable training data. In this work, we propose a classifier that estimates Gaussian components supervised from training data while estimating mixture proportions unsupervised from the test set. In a simulation study, our classifier maintained accuracy across varying test bot prevalence rates. More generally, our approach is useful when true mixture proportions are unknown while training data is sampled stratified by class.

Afin de détecter les réponses émanant de bots dans les données d'enquêtes de type Likert, les chercheurs utilisent généralement un échantillon de formation avec stratification de classe : pour produire des exemples humains, l'enquête est menée auprès d'humains certifiés, mais pour obtenir des exemples de bots, des données aléatoires sont simulées. Comme d'une part la prévalence de formation des bots est arbitraire, les classificateurs supervisés conventionnels échouent. À la limite, quand les bots approchent les 100 % de l'échantillon, chaque observation est classifiée comme bot de façon indiscriminée. D'autre part, des solutions complètement non supervisées, comme l'algorithme espérance-maximisation, ne prennent pas en compte de précieuses données de formation. Nous proposons un classificateur qui estime les composantes gaussiennes supervisées tirées des données de formation, tout en estimant les proportions du mélange non supervisé découlant du jeu-test de données. Dans une étude en simulation, l'exactitude de notre classificateur s'est maintenue dans des tests variés de taux de prévalence des bots. De façon plus générale, notre approche est utile lorsque les vraies proportions du mélange sont inconnues, tandis que les données de formation sont échantillonnées par stratification de classe.

[12:15-12:30]

Wanhua Su (MacEwan University)

Classification With Imbalanced Data

Classification avec données déséquilibrées

In some binary classification applications such as fraud detection and direct marketing, the two classes are extremely imbalanced. There are two popular approaches to handle classification with imbalanced data: use cost-

Dans certaines applications de classification binaire comme en détection des fraudes et en marketing direct, les deux classes sont extrêmement déséquilibrées. Il existe deux approches populaires pour traiter la classification avec données déséquilibrées :

sensitive performance metric to train the classifiers and resampling methods such as SMOTE (synthetic minority oversampling technique). It is well known that the average precision (i.e., the area under the precision-recall curve) is an informative performance metric for imbalanced data. We propose a family of learning procedures that combines the average precision and resampling techniques. The performance of the proposed method is investigated through comprehensive simulation studies and a variety of benchmark data sets.

la mesure de performance sensible au coût pour former le classificateur et des méthodes de rééchantillonnage comme SMOTE (technique de suréchantillonnage synthétique des minorités). Il est bien connu que la précision moyenne (c.-à-d. l'aire sous la courbe précision-rappel) est une mesure informative de la performance pour les données déséquilibrées. Nous proposons une famille de procédures d'apprentissage qui combine la précision moyenne et des techniques de rééchantillonnage. La performance de la méthode proposée est étudiée à l'aide d'études de simulation approfondies et divers ensembles de données de référence.

Recent Advances in Regression Methods
Progrès récents en méthodes de régression

Chair/Président: Max Turgeon

Date: Friday June 3 / vendredi 3 juin

Time/Heure: 11:00-12:30

Abstract/Résumé

[11:00-11:15]

Jason Hou-Liu (University of Waterloo) **Ryan P. Browne** (University of Waterloo)

Fast Estimation of Generalized Linear Models via Sketching

Estimation rapide de modèles linéaires généralisés par esquisse

Generalized linear models form the backbone of many statistical and business analyses and yield readily interpretable results. Contemporary datasets can exhibit a very large number of observations, making conventional iterative estimation procedures difficult. In these massive data contexts, data processing, access, and transfer can become time-prohibitive and hamper model prototyping and analysis. We apply sketching as a stochastic data reduction method to generate surrogate datasets upon which estimation can be more tractably performed. Probabilistic results and practical applications are discussed with empirical simulations and comparison against existing methodologies. We extend the simulation study to consider different magnitudes of massive data and the effect on the associated handling procedures, with attention to computational infrastructure configurations.

Les modèles linéaires généralisés constituent l'épine dorsale de nombreuses analyses statistiques et commerciales et produisent des résultats facilement interprétables. Les ensembles de données contemporains peuvent inclure un très grand nombre d'observations, ce qui rend l'estimation itérative classique difficile. Dans ces contextes de données massives, le traitement, l'accès et le transfert des données peuvent devenir prohibitifs en termes de temps et entraver le prototypage et l'analyse des modèles. Nous appliquons l'esquisse comme méthode de réduction stochastique des données pour générer des ensembles de données de substitution sur lesquels l'estimation peut être réalisée de manière plus efficace. Nous discutons des résultats probabilistes et des applications pratiques via des simulations empiriques et une comparaison avec les méthodes existantes. Nous étendons l'étude par simulation pour considérer différentes magnitudes de données massives et l'effet sur les procédures de traitement associées, en prêtant attention aux configurations de l'infrastructure de calcul.

[11:15-11:30]

Hui Guo (Western University)

Variable Selection for Logistic Regression Models with Misclassified Response

Sélection de variables pour les modèles de régression logistique avec réponse classée incorrectement

In contrast to extensive attention to variable selection with measurement error in covariates, variable selection in the presence of response measurement error remains relatively underexplored. In this talk, we study this problem in the context of logistic regression with the error-prone response. We develop variable selection procedures using validation sample via penalized strategies with response mismeasurement effects accommodated. We theoretically examine the root n consistency of the proposed procedures and derive the asymptotic normality for both parametric and semi-parametric settings. With properly chosen penalty function and the regularization parameter, the oracle property holds for

Contrairement à la forte attention accordée à la sélection de variables avec erreur de mesures dans les covariables, la sélection de variables en présence d'erreurs de mesure de réponse reste relativement inexplorée. Lors de cet exposé, nous étudions ce problème en contexte de régression logistique avec la réponse sujette à erreur. Nous élaborons les procédés de sélection de variables à l'aide d'échantillon de validation par l'entremise de stratégies pénalisées avec effets de réponse mal mesurée et adaptée. Nous examinons théoriquement la convergence d'ordre racine n des procédés proposés et dérivons la normalité asymptotique dans des contextes paramétriques et semi-paramétriques. Au moyen d'une fonction de pénalité adéquatement choisie et de paramètre de régularisation, la propriété de l'oracle reste valide en ce qui concerne les estima-

the resulting estimators. To assess the finite sample properties of the proposed variable selection procedures, we carry out numerical studies, and the empirical results verify the effectiveness of our proposed algorithm.

teurs obtenus. Nous menons des études numériques afin d'évaluer les propriétés de l'échantillon fini par le procédé de sélection de variables proposé, et les résultats empiriques confirment l'efficacité de l'algorithme proposé.

[11:30-11:45]

Yansan Han (Brock University) **Mei Ling Huang** (Brock University) **William Marshall** (Brock University)

Quantile Regression Analysis on COVID-19

Analyse de régression quantile sur la COVID-19

Quantile regression (QR) estimates conditional quantiles with applications in the real world. Estimating extreme conditional quantiles is an important and difficult problem. The regular quantile regression method often sets a linear model with estimating 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 two-stage nonparametric quantile regression method with a 5-step algorithm by using extrapolation. Monte Carlo simulations show good efficiency for the proposed nonparametric QR extrapolation estimator relative to the regular linear QR extrapolation estimator. The paper also investigates on COVID-19 in Ontario Canada by using the proposed method. Comparisons of the proposed method and existing methods are given.

La régression quantile estime les quantiles conditionnels avec des applications du monde réel. L'estimation des quantiles conditionnels extrêmes est un problème important et difficile à résoudre. La méthode de régression quantile classique définit souvent un modèle linéaire à l'aide d'une estimation des coefficients pour obtenir le quantile conditionnel estimé. Cette approche risque d'être limitée par le paramétrage du modèle. Pour surmonter ce problème, nous proposons une méthode de régression quantile non paramétrique en deux étapes avec un algorithme d'extrapolation en cinq étapes. Les simulations de Monte-Carlo mettent en évidence la bonne efficacité de l'estimateur d'extrapolation par régression quantile non paramétrique proposé par rapport à l'estimateur d'extrapolation par régression quantile linéaire classique. Dans cette présentation, à l'aide de notre méthode, nous étudions également la COVID-19 en Ontario, au Canada. Enfin, nous comparons notre méthode à celles qui sont actuellement utilisées.

[11:45-12:00]

Zeyu Chen (University of Toronto: Dalla Lana School of Public Health) **Osvaldo Espin-Garcia** (University of Toronto: Dalla Lana School of Public Health)

The Perils of Ignoring the Study Design in High-Dimensional Settings: A Simulation-based Evaluation

Les dangers d'ignorer la conception de recherche dans des contextes à haute dimension : une évaluation en simulation

The nested case-control (NCC) and the case-cohort (CCH) are two commonly used study designs in resource-limited scenarios, however, their statistical properties (e.g., variable selection and predictive performance) in high-dimensional settings are less understood. In particular, it is unclear how commonly used approaches for high-dimensional data can take the design into account. This project evaluates the consequences of ignoring these two design structures in high-dimensional settings via Monte Carlo simulations with a focus on LASSO regression models and random survival forests. Preliminary results indicate that the sensitivity of variable selection and the prediction performance tend to decrease when the study design is ignored for both NCC and CCH. More simulation results are coming out in the next few months to give a more comprehensive summary of the perils. Our project demonstrates the im-

Même si l'étude cas-témoins (NCC) et l'étude cas-cohorte (CCH) sont deux plans d'étude couramment utilisés dans des contextes à ressources limitées, leurs propriétés statistiques (par ex. : la sélection des variables et la performance prédictive) dans des contextes à haute dimension sont moins comprises. En particulier, la façon dont les approches couramment utilisées pour les données à haute dimension peuvent prendre en compte le plan d'étude n'est pas clair. À l'aide de simulations de Monte Carlo axées sur des modèles de régression LASSO et des forêts de survie aléatoires, ce projet évalue les conséquences d'ignorer ces deux structures de plan dans des contextes à haute dimension. Des résultats préliminaires indiquent que la sensibilité de la sélection des variables et de la performance prédictive tend à diminuer lorsque le plan d'étude est ignoré à la fois dans la NCC et la CCH. Dans les prochains mois, d'autres résultats de simulation viendront préciser davantage ces dangers. Notre projet montre l'importance de prendre en compte la structure de conception de recherche

portance of considering the study design structure for high-dimensional data analysis.

pour l'analyse de données à haute dimension. Même si l'étude cas-témoins (NCC) et l'étude cas-cohorte (CCH) sont deux conceptions de recherche couramment utilisées dans des contextes à ressources limitées, leurs propriétés statistiques (par ex. : la sélection des variables et la performance prédictive) dans des contextes à haute dimension sont moins comprises. En particulier, la façon dont les approches couramment utilisées pour les données à haute dimension peuvent prendre en compte la conception de recherche est moins claire. À l'aide de simulations de Monte Carlo axées sur des modèles de régression LASSO et des forêts de survie aléatoires, ce projet évalue les conséquences d'ignorer ces deux structures de conception dans les contextes à haute dimension. Des résultats préliminaires indiquent que la sensibilité de la sélection des variables et de la performance prédictive tend à diminuer lorsque la conception de recherche est ignorée à la fois dans la NCC et la CCH. D'autres résultats en simulation attendus dans les prochains mois préciseront davantage ces dangers. Notre projet montre l'importance de prendre en compte la structure de conception de recherche pour l'analyse de données à haute dimension.

[12:00-12:15]

Chong Gan (University of Guelph) **Zeny Feng** (University of Guelph) **Jiahua Chen** (University of British Columbia)

Association Tests under Gaussian Mixture Regression Models

Tests d'association selon les modèles de régression de mélange gaussien

Gaussian mixture model is being increasingly used to cluster the unobserved heterogeneous data. It is common that some covariates are related to the observed outcomes of interests such that they provide valuable information to cluster the responses, so we performed the likelihood ratio test for the association between these variables. However, the number of subgroups is usually unknown and the numbers of subgroups selected under null and alternative hypotheses can be different. This generates challenges in designing testing procedures and the interpretation of testing results. We proposed Naïve I, II, III and weighted significance procedures to address these problems, and used criteria of type I error, power and adjusted rand index to compare the performance of these procedures in simulation studies. Our proposed methods will be applied to the diabetes data, revealing how glucose tolerance impacts insulin concentrations for different underlying groups of people.

Les modèles de mélange gaussien servent de plus en plus à regrouper les données hétérogènes non observées. Certaines covariables sont fréquemment liées aux résultats observés pertinents, ce qui leur permet de fournir des renseignements précieux pour grouper les réponses. C'est pourquoi nous avons réalisé un test du rapport des vraisemblances pour les associations entre ces variables. Cependant, le nombre de sous-groupes est généralement inconnu et les nombres de sous-groupes sélectionnés selon des hypothèses nulle et alternative peuvent être différents. Ce qui pose donc problème dans la conception de procédures de test et l'interprétation des résultats de test. Nous proposons des procédures de signification pondérée et naïves I, II et III pour résoudre ces problèmes, et utilisons des critères d'erreur de type I et l'indice de Rand ajusté et de puissance pour comparer la performance de ces procédures dans des études par simulation. Les méthodes proposées seront appliquées à des données sur le diabète, révélant comment la tolérance au glucose influence les concentrations d'insuline pour différents groupes sous-jacents de personnes.

[12:15-12:30]

Gunho Bae (University of Manitoba) **Saumen Mandal** (University of Manitoba)

Optimal Experimental Designs for Estimating Parameters Independently of Each Other

Plans d'expérience optimaux pour l'estimation de paramètres indépendamment les uns des autres

We construct optimal designs for some regression models in which it is desired to estimate certain parameters independently of each other. Motivated by this fact, we construct designs by minimizing absolute correlations among the least squares estimators of the parameters or linear functions of the parameters in a linear model. In the case of estimating a parameter independently of another two parameters, we create a compound optimization criterion and then solve the problem by means of a simultaneous optimization technique. This is an optimization problem in which we maximize two functions of the design weights simultaneously. The functions have a common maximum of zero which is simultaneously attained at the optimal design weights. In order to construct the designs, we use a class of algorithms, indexed by a function which satisfies certain conditions. In conclusion, some results will be reported and discussed.

Nous construisons des plans optimaux pour des modèles de régression dans lesquels une estimation des paramètres indépendamment les uns des autres est souhaitée. Par conséquent, nous construisons ces plans en minimisant les corrélations absolues entre les estimateurs des moindres carrés des paramètres ou de fonctions linéaires des paramètres dans un modèle linéaire. Dans le cas d'une estimation d'un paramètre indépendamment de deux autres, nous créons un critère d'optimisation composé, puis résolvons le problème au moyen d'une technique d'optimisation simultanée. Il s'agit d'un problème d'optimisation dans lequel nous maximisons simultanément deux fonctions des pondérations du plan d'expérience. Les fonctions ont un maximum de zéros communs atteint simultanément aux pondérations optimales du plan. Pour la construction des plans, nous utilisons une classe d'algorithmes indexée par une fonction qui répond à certaines conditions. En conclusion, certains résultats seront présentés et discutés.

CANSSI Programs and Plans
Programmes et plans de l'INCASS

Chair/Président: Donald Estep

Organizer/Responsable: Donald Estep

Date: Friday June 3 / vendredi 3 juin

Time/Heure: 12:30-13:30

Abstract/Résumé

[12:30-13:30]

Andrea Benedetti (McGill University) **Joanna Mills Flemming** (Dalhousie University) **Donald Estep** (CANSSI)

CANSSI Programs and Plans

Programmes et plans de l'INCASS

The session will start with a presentation about CANSSI programs. It will focus on the new programs, covering goals, eligibility, and application. Next, the session will describe developments with CANSSI Regional Centres, including a new program for CANSSI Atlantic. Finally we will discuss recent and future plans for CANSSI.

La session débutera par une présentation des programmes de l'INCASS. Elle se concentrera sur les nouveaux programmes, les objectifs à couvrir, l'admissibilité et la demande. Ensuite, la session décrira les développements avec les centres régionaux de l'INCASS, y compris un nouveau programme pour l'INCASS Atlantique. Enfin, nous discuterons des plans récents et futurs de l'INCASS.

Chair/Président: Andrea Benedetti

Organizer/Responsable: Andrea Benedetti

Date: Friday June 3 / vendredi 3 juin

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-14:00]

Kaiqiong Zhao (McGill University) **Linglong Kong** (University of Alberta) **Yanchun Bao** (University of Essex)

A Multi-Channel Fusion Framework for Statistical Learning and Inference with its Application in Multi-Omics Data Analysis

Cadre de fusion multicanal pour l'apprentissage et l'inférence statistiques et application à l'analyse de données multi-omiques

Multiple types of genomics data are now increasingly available. Existing statistical methods for integrative analysis focus on a common set of samples for which all individual data types are available. However, in practice, certain genomic features are measured in only a small fraction of the entire study population. Eliminating samples that do not overlap across data types can result in substantial information loss. Therefore, we propose a multi-channel fusion framework to efficiently integrate multi-omics data while maximizing information gain. We build upon distributional robust optimization theory and summarize the information from those non-overlapping samples using estimating equations. Our framework makes a relaxed assumption for the underlying outcome generating mechanism and explicitly allows for heterogeneity in the covariate distributions. We anticipate our integrative framework will yield better-calibrated association estimates and improve inference and prediction results.

De multiples types de données génomiques sont désormais de plus en plus disponibles. Les méthodes statistiques existantes pour l'analyse intégrative se concentrent sur un ensemble commun d'échantillons pour lesquels tous les types de données individuelles sont disponibles. Cependant, dans la pratique, certaines caractéristiques génomiques ne sont mesurées que pour une petite fraction de l'ensemble de la population étudiée. Or l'élimination des échantillons qui ne se chevauchent pas entre types de données peut entraîner une perte d'information substantielle. Par conséquent, nous proposons un cadre de fusion multicanal qui permet d'intégrer efficacement les données multi-omiques tout en maximisant le gain d'information. Nous nous appuyons sur la théorie de l'optimisation robuste distributive et résumons l'information des échantillons qui ne se chevauchent pas via des équations estimantes. Notre cadre fait une hypothèse détendue pour le mécanisme sous-jacent de génération de résultats et permet explicitement l'hétérogénéité dans les distributions de covariables. Nous prévoyons que notre cadre intégratif produira des estimations d'association mieux calibrées et améliorera l'inférence les résultats de prédiction.

[14:00-14:30]

Caitlin Ward (University of Calgary) **Rob Deardon** (University of Calgary) **Alexandra M. Schmidt** (McGill University)

Sound the alarm: modeling behavioral changes in response to epidemic intensity

Alarme ! Modélisation des changements de comportement en réponse à l'intensité d'une épidémie

For many infectious disease outbreaks, the at-risk population changes their behavior in response to the outbreak severity, changing the transmission dynamics to change in real-time. Various approaches to behavioral change modeling have been proposed, but work assessing the statistical properties of these models is limited. We propose a model formulation where time-varying transmis-

Dans le cas de nombreuses éclosions de maladies infectieuses, la population à risque modifie son comportement en fonction de la gravité de l'épidémie, ce qui modifie la dynamique de transmission en temps réel. On a proposé diverses approches de modélisation des changements de comportement, mais les travaux évaluant les propriétés statistiques de ces modèles sont limités. Nous proposons un modèle dans lequel la transmission variant

sion is captured by the level of "alarm" in the population and specified as a function of the past epidemic trajectory. The model is set in a data-augmented Bayesian framework as epidemic data are often only partially observed, and we can utilize prior information to help with parameter identifiability. We investigate the estimability of the population alarm across a wide range of scenarios, using both parametric functions and non-parametric Gaussian process and splines. The benefit and utility of the proposed approach is illustrated through an application to COVID-19 data from various cities.

dans le temps est représentée par le niveau « d'alarme » de la population et est définie comme une fonction de la trajectoire épidémique passée. Nous créons ce modèle dans un cadre bayésien de données augmentées, puisque les données épidémiques ne sont souvent que partiellement observées. Ainsi, nous pouvons utiliser de l'information a priori pour faciliter l'identifiabilité des paramètres. Nous analysons l'estimabilité de l'alarme de population dans un grand nombre de scénarios, en utilisant à la fois des fonctions paramétriques, des processus gaussiens non paramétriques et des splines. Nous illustrons les avantages et l'utilité de notre approche en l'appliquant aux données sur la COVID-19 de plusieurs villes.

[14:30-15:00]

Cédric Beaulac (Simon Fraser University/University of Victoria)

Neural Network Classifiers for Features Extraction in Neuroimaging Genetics

Utiliser un réseau de neurones de classification pour extraire des variables d'imagerie cérébrale.

A major issue in the association of genes to imaging phenotypes is the high dimension of both the genetic data and imaging data. In this talk, we discuss our recent work addressing the latter problem. A key concept of our work is a neuroimaging genetic pipeline where we separate the neuroimaging genetic association in three distinct steps: first is image processing, then, neuroimaging feature extraction and finally the genetic association study. The novelty of our approach is using a neural network classifier in order to accomplish the second step; extracting neuroimaging features that are related with AD. One could think of our model as an AutoEncoder where the decoder is replaced by a prediction function. To conclude, we compared the predictive power of the features automatically extracted by our approach to expert-selected features.

Un problème majeur lors de la régression des phénotypes d'imagerie cérébrale sur le génotype est la grande dimension à la fois des images et des gènes. Dans cette présentation, nous discutons de notre récent projet où nous nous attaquons au premier problème. Un concept central de notre approche est de séparer l'inférence génétique de l'extraction des variables d'imagerie cérébrale. La nouveauté principale est dans l'utilisation d'un modèle construit à l'aide d'un réseau de neurones entrainé pour la classification à des fins de réduction de la dimension. Dans ce projet, on trace un parallèle entre notre modèle et un autoencodeur où l'on remplace le décodeur par une fonction de classification. Pour conclure, on compare la précision de prédiction des variables automatiques extraites par notre modèle à celles sélectionnées par des experts.

Recent Advances in Mixture Models: Theory and Application
Avancées récentes en modèles de mélange : théorie et application

Chair/Président: Juxin Liu

Organizer/Responsable: Juxin Liu

Date: Friday June 3 / vendredi 3 juin

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-14:00]

Zeny Feng (University of Guelph) **Sanjeena Subedi** (Carleton University) **Stephen Bak** (University of Guelph) **Drew Neish** (University of Guelph)

Mixture of Dirichlet Multinomial (DM) Models in Microbiome Data Analysis

Mélange de modèles multinomiaux de Dirichlet (MD) dans l'analyse de données de microbiome

The next generation sequencing technique enables to generate a massive among of data such as metagenomic microbiome data for the exploration and detection of relationships among biological/environmental traits, microbiome composition, and their impacts on human health. Human gut microbial communities vary greatly among individuals, clustering human gut to different types (enterotypes) is an excited research area. The observed bacterial operational taxonomic unit (OUT) counts at a given rank is modelled by a Dirichlet Multinomial (DM) distribution. We propose to cluster the microbial profiles into enterotypes via mixture of DM regression where influencing factors on clustering can be incorporated in the model through regression. In the M-step of the EM algorithm that is used to estimate model parameters, a minorization-maximization method is used. Simulation studies and real data analysis are included to evaluate the performance and demonstrate the application of the proposed method.

La technique de séquençage de deuxième génération permet de générer une quantité massive de données comme des données de microbiomes métagénomiques servant à découvrir et repérer les liens entre les traits environnementaux et biologiques, la composition de microbiome et leurs impacts sur la santé. Les communautés bactériennes intestinales chez l'humain varient grandement selon la personne. C'est pourquoi le regroupement d'intestins humains en différents types (entérotypes) est un domaine de recherche qui suscite de l'intérêt. Le compte observé d'une unité taxonomique opérationnelle (UTO) bactérienne à un rang donné est modélisé par une distribution multinomiale de Dirichlet. Nous proposons de regrouper les profils bactériens en entérotypes par mélange de régression MD où les facteurs influant le regroupement peuvent être intégrés dans le modèle par une régression. Nous employons une méthode de minimisation-maximisation dans l'étape M de l'algorithme EM utilisé pour estimer les paramètres du modèle. Des études en simulation et des analyses de données réelles sont comprises pour évaluer la performance de la méthode proposée et pour démontrer son application.

[14:00-14:30]

Abbas Khalili (McGill University) **Tudor A. Manole** (Carnegie Mellon University)

A Group-Sort-Fuse Procedure for Estimating the Number of Components in Finite Mixture Models

Une procédure Group-Sort-Fuse pour l'estimation du nombre de composants dans des modèles de mélanges finis

Estimation of the number of components (or order) of a finite mixture model is a long-standing and challenging problem in statistics. We propose the Group-Sort-Fuse (GSF) procedure, a new penalized likelihood approach for simultaneous estimation of the order and mixing measure in multidimensional finite mixture models. Unlike methods that fit and compare mixtures with varying orders using criteria involving model complexity, our

En statistique, l'estimation du nombre de composants (ou de l'ordre) d'un modèle de mélanges finis est depuis longtemps un problème difficile. Nous proposons la procédure Group-Sort-Fuse (GSF), une nouvelle approche par vraisemblance pénalisée pour l'estimation simultanée de l'ordre et de la mesure de mélange dans des modèles de mélanges finis multidimensionnels. Contrairement aux méthodes qui ajustent et comparent des mélanges d'ordres différents à l'aide de critères qui font intervenir la com-

approach directly penalizes a continuous function of the model parameters. More specifically, given a conservative upper bound on the order, the GSF groups and sorts mixture component parameters to fuse those which are redundant. For a wide range of finite mixture models, we show that the GSF is consistent in estimating the true mixture order. The GSF is implemented for several univariate and multivariate mixture models in the R package GroupSortFuse. We also discuss its finite sample performance via simulations and a real data example.

plexité du modèle, notre approche pénalise directement une fonction continue des paramètres du modèle. Plus précisément, étant donné une limite supérieure conservatrice sur l'ordre, la GSF regroupe et trie les paramètres des composants du mélange, afin de fusionner ceux qui sont redondants. Pour une vaste gamme de modèles de mélanges finis, nous montrons que la procédure GSF est cohérente dans l'estimation de l'ordre véritable du mélange. La GSF est implémentée dans plusieurs modèles de mélanges univariés et multivariés dans le paquet GroupSortFuse de R. La GSF est implémentée pour plusieurs modèles de mélange univariés et multivariés dans le package R GroupSortFuse. Sa performance sur des échantillons de taille finie est illustrée par des études par simulations et un exemple avec des données réelles.

[14:30-15:00]

Jiahua Chen (The University of British Columbia) **Qiong Zhang** (University of British Columbia)

Gaussian Mixture Reduction based on Composite Transportation Divergence

Réduction de mélange gaussien en fonction de la divergence de transport composite

In many applications, researchers wish to approximate a finite Gaussian mixture distribution with a high order by one with a lower order. Examples include density estimation, recursive tracking in hidden Markov model, and belief propagation. A direct solution to such a Gaussian Mixture Reduction problem is computationally challenging due to the non-convexity of commonly employed optimality targets. One popular line of approach is to employ some clustering-based iterative algorithms. Neither their convergence nor destination, however, are thoroughly discussed. In this paper, we propose a new GMR method by minimizing some novel composite transportation divergence (CTD). This divergence permits an easy to implement Majorization-Minimization (MM) algorithm. We prove that the MM algorithms converge under general conditions, and many existing clustering-based algorithms are special cases of our approach. We further investigate the property of this approach with various choices of cost functions and demonstrate its effectiveness and computational costs.

Dans de nombreuses applications, les chercheurs souhaitent approximer une distribution d'ordre supérieur de mélange gaussien fini par une distribution d'ordre inférieur. Les exemples portent sur l'estimation de la densité, le suivi récursif dans un modèle de Markov caché et la propagation des croyances. Il est difficile de trouver une solution directe à ce problème de réduction du mélange gaussien en raison de la non-convexité des objectifs d'optimalité couramment utilisés. Une approche populaire consiste à utiliser certains algorithmes itératifs basés sur le groupement. Cependant, ni la convergence ni la destination de ces algorithmes ne sont examinées en détail. Dans cette étude, nous proposons une nouvelle méthode de réduction de mélange gaussien par la réduction au minimum d'une nouvelle divergence de transport composite. Cette divergence est à l'origine d'un algorithme majorisation-minimisation facile à mettre en œuvre. Nous démontrons que l'algorithme de majorisation-minimisation converge dans des conditions générales, et que de nombreux algorithmes actuels reposant sur le regroupement sont des cas particuliers de notre approche. Nous examinons plus en détail les propriétés de cette approche selon différents choix de fonctions de coût, et nous démontrons son efficacité et ses coûts de calcul.

2022 CJS Award Address
Allocution du récipiendaire du prix de la RCS 2022

Chair/Président: Andrei Volodin

Date: Friday June 3 / vendredi 3 juin

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-14:30]

Li Xing (University of Saskatchewan) **Xuekui Zhang** (University of Victoria) **Igor Burstyn** (Drexel University) **Paul Gustafson** (University of British Columbia)

The shape of the relationship between a continuous exposure variable and a binary disease variable is often central to epidemiologic investigations. This article investigates a number of issues surrounding inference and the shape of the relationship. Presuming that the relationship can be expressed in terms of regression coefficients and a shape parameter, we investigate how well the shape can be inferred in settings which might typify epidemiologic investigations and risk assessment. We also consider a suitable definition of the median effect of exposure, and investigate how precisely this can be inferred. This is done both in the case of using a model acknowledging uncertainty about the shape parameter and in the case of ignoring this uncertainty and using a two-step method, where in step one we transform the predictor and in step two we fit a simple logistic model with transformed predictor. All these investigations require a family of exposure-disease relationships indexed by a shape parameter. For this purpose, we employ a family based on the Box–Cox transformation.

La forme de la relation entre une variable continue d'exposition et une variable binaire de maladie est souvent centrale dans les enquêtes e'pide'miologiques. Nous explorons certains proble'mes lie's a' l'infe'rence et a' la forme de cette relation. En supposant que la relation peut s'exprimer par des coefficients de re'gression, nous e'tudions a' quel point les formes typiquement attendues dans les enquêtes e'pide'miologiques et l'e'valuation des risques peuvent être infe're'es. Nous considéront e'galement une de'finition approprie'e pour l'effet me'dian d'exposition, et investiguont a' quel point elle peut être infe're'e. Nous considéront deux cas, l'un en reconnaissant l'incertitude par rapport au parame'tre de forme, et l'autre en ignorant cette incertitude dans le cadre d'une me'thode en deux e'tapes consistant en une transformation des pre'dicteurs, suivie de l'ajustement d'une re'gression logistique sur ces variables transforme'es. Ces investigations requie'rent une famille de relations entre l'exposition et la maladie indexe'e par un parame'tre de forme. Nous utilisont une famille base'e sur la transforme'e de Box-Cox.

Perspectives on Open Education Resources
Perspectives sur les ressources éducatives libres

Chair/Président: Sotirios Damouras

Organizer/Responsable: Sotirios Damouras

Date: Friday June 3 / vendredi 3 juin

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-14:00]

Surita Jhangiani (The University of British Columbia)

Leveraging Open Educational Resources in Higher Education

Ressources éducatives libres dans l'enseignement supérieur

Open educational resources (OER) enable faculty to leverage digital technologies to reuse, revise, remix, retain, and redistribute educational resources, making education more affordable and accessible to students while shaping instructional materials to reflect course learning outcomes. From openly-licensed textbooks and interactive simulations to open homework systems, OER enable faculty to engage students and colleagues in new, exciting, and innovative ways. This presentation will introduce and demystify OER and discuss their multifaceted benefits for both faculty and students. Participants will leave with an understanding of open licensing, know how to discover relevant and high quality resources for statistics, and how to leverage the affordances of OER in service of their pedagogical goals.

Les ressources éducatives libres (REL) permettent au corps enseignant d'exploiter les technologies numériques pour réutiliser, réviser, remixer, conserver et redistribuer des ressources éducatives, rendant ainsi l'éducation plus abordable et plus accessible aux étudiants tout en façonnant le matériel pédagogique de manière à refléter les résultats d'apprentissage des cours. Qu'il s'agisse de manuels sous licence ouverte, de simulations interactives ou de systèmes de devoirs ouverts, les REL permettent aux professeurs d'impliquer les étudiants et leurs collègues de manière nouvelle, passionnante et innovante. Cette présentation introduira et démystifiera les REL et discutera de leurs multiples avantages tant pour le corps enseignant que pour les étudiants. Les participants repartiront avec une compréhension des licences ouvertes, sauront comment découvrir des ressources pertinentes et de haute qualité en statistique, et comment tirer parti des possibilités des REL au service de leurs objectifs pédagogiques.

[14:00-14:30]

Trevor Campbell (The University of British Columbia) **Melissa Lee** (University of British Columbia) **Tiffany A. Timbers** (University of British Columbia)

Creating Open Resources for an Introductory Data Science Course

Création de ressources ouvertes pour un cours d'introduction à la science des données

Open-source educational resources have many benefits for both instructors and learners alike, such as cost-savings for learners, quick iteration on materials for instructors, and increased quality through auditable collaboration. And given that the field of data science has embraced open-source development practices in such a big way, we believe that data science educators should "practice what they preach" by teaching with, and sharing, open-source tools and resources. However, data science is a very new field; educational resources at the truly introductory level are

Les ressources éducatives open-source présentent de nombreux avantages tant pour les instructeurs que pour les apprenants, tels que des économies pour les apprenants, une itération rapide du matériel pour les instructeurs et une qualité accrue grâce à une collaboration vérifiable. Étant donné que le domaine de la science des données n'a pas hésité à adopter les pratiques de développement open-source, nous pensons que les éducateurs en science des données devraient « pratiquer ce qu'ils prêchent » en enseignant avec des outils et des ressources open-source et en les partageant. Cependant, la science des données est un domaine très récent, ce qui fait que les ressources éducatives au niveau vrai-

sparse and generally not based on real classroom experience. In this talk, we will discuss an open-source textbook (https://datasciencebook.ca) and complementary worksheets for introducing students to data science that we have created at the University of British Columbia. These resources were built using rigorous open-source workflows and tools (e.g., R, bookdown, Python, Jupyter, Git, and GitHub) in a public repository (https://github.com/ubc-dsci/introduction-to-datascience), and have been tested in a first-year undergraduate course that has served thousands of students over three years. Specifically, we will comment on the initial objectives of the project and how they evolved, tooling choices we made and their rationale, our collaboration strategy, and hard-won lessons we learned both in-class and offline about effective data science pedagogy in a large-scale classroom.

ment introductif sont rares et ne sont généralement pas basées sur une expérience réelle en classe. Dans cet exposé, nous discuterons d'un manuel open-source (https ://datasciencebook.ca) et de feuilles de travail complémentaires pour introduire la science des données aux étudiants que nous avons créés à l'Université de Colombie-Britannique. Ces ressources ont été créées avec des flux de travail et outils open-source rigoureux (par exemple, R, bookdown, Python, Jupyter, Git et GitHub) dans un dépôt public (https ://github.com/ubc-dsci/introduction-to-datascience) et testées dans un cours de première année du premier cycle proposé à des milliers d'étudiants sur trois ans. Plus précisément, nous commenterons les objectifs initiaux du projet et la façon dont ils ont évolué, nos choix d'outils et leur justification, notre stratégie de collaboration et les leçons durement acquises en classe et hors ligne sur la pédagogie efficace de la science des données dans une classe à grande échelle.

[14:30-15:00]

Toby Hodges (The Carpentries)

Perspectives on Development and Use of Open Educational Resources at Scale

Perspectives sur le développement et l'utilisation de ressources éducatives libre à l'échelle

The Carpentries teaches essential software and data skills to researchers and librarians around the world, using lessons that are all open source, open licensed, and freely available online. Our Instructors are trained and certified to teach these skills in short Software, Data, and Library Carpentry workshops. In addition to the 27 lessons in our official Lesson Programs, the community reuses and adapts the systems and approaches behind those lessons to develop and teach many more. The Carpentries Incubator is home to more than one hundred of these community-developed lessons, at various levels of maturity. This development is supported by dedicated training, a system of open peer review, and bespoke infrastructure for accessible lesson websites. In this talk, I will outline past and present efforts to foster community around creating, maintaining, and teaching open lessons, and steps we are taking to ensure that these initiatives can scale as the community continues to grow.

Le projet ≪The Carpentries≫ enseigne un savoir-faire relatif aux données et l'utilisation de logiciels à des chercheurs et des bibliothécaires partout dans le monde au moyen de cours ouverts et à licence libre, le tout offert gratuitement en ligne. Nos enseignants sont formés et certifiés pour enseigner ces compétences dans de courts ateliers Carpentry en matière de logiciels, données et bibliothèque. En plus des 27 cours de notre programme de cours officiel, la communauté réutilise et adapte les systèmes et les approches derrière ces cours pour élaborer et enseigner davantage. Le site ≪Carpentries Incubator≫ comprend plus d'une centaine de ces cours conçus par la communauté, à plusieurs niveaux de maturité. Ce développement est soutenu par une formation assidue, un système de rétroaction ouvert et une infrastructure sur mesure pour des sites de cours accessibles. Lors de cet exposé, je décrirai les efforts mis en œuvre jusqu'à maintenant pour encourager la communauté à créer, entretenir et offrir des cours ouverts, ainsi que les étapes que nous suivons pour nous assurer que ces initiatives puissent s'élargir parallèlement à la croissance de la communauté.

Chair/Président: Aerambamoorthy A Thavaneswaran

Organizer/Responsable: Aerambamoorthy A Thavaneswaran

Date: Friday June 3 / vendredi 3 juin

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-14:00]

Shelton Peiris (The University of Sydney) **David Dowe** (Monash University, Victoria, Australia) **Zheng Fang** (Monash University, Victoria, Australia) **Dedi Rosadi** (University of Gadjah Mada, Indonesia) **Aerambamoorhty Thavaneswaran** (University of Manitoba)

A Novel ARFIMA-ANN Hybrid Model for Financial Time Series Forecasting

Nouveau modèle hybride de ARFIMA-RNA pour la prévision des séries chronologiques financières

Autoregressive Fractionally Integrated Moving Average (ARFIMA) has been successfully applied in modelling and forecasting economic time series with long memory. It is known that the Artificial Neural Network (ANN) approach can be used to capture additional complex nonlinear economic relationships with many unknown patterns. This paper proposes a hybrid model, which is distinctive in integrating the advantages of ARFIMA and ANN in modelling and the analysis of linear and nonlinear components of a financial time series with long memory. This builds upon earlier work in combining (i) deep learning ANN LSTM modelling with (ii) the Bayesian information-theoretic minimum message length (MML) principle applied to ARMA time series. A simulation study is carried out to investigate the properties of this ARFIMA-ANN hybrid modelling and forecasting. We justify the usefulness of the proposed hybrid model in practice and compare the accuracy of forecasting with existing models.

Le modèle de moyenne mobile autorégressive partiellement intégrée et de réseau de neurones artificiels (ARFIMA) a été appliqué avec succès à la modélisation et à la prévision de séries chronologiques économiques à mémoire longue. On sait que l'approche du réseau de neurones artificiels (RNA) peut être utilisée pour capturer des relations économiques non linéaires complexes supplémentaires avec de nombreux modèles inconnus. Dans cette présentation, nous proposons un modèle hybride, qui se distingue par l'intégration des avantages de l'ARFIMA et du RNA dans la modélisation et l'analyse des composantes linéaires et non linéaires d'une série chronologique financière à mémoire longue. Ce modèle s'appuie sur des travaux antérieurs combinant i) la modélisation d'un RNA de la mémoire à long et court terme (apprentissage profond) et ii) le principe de la longueur minimale des messages de la théorie de l'information bayésienne appliqué aux séries chronologiques de modèles ARMA (modèles autorégressifs et moyenne mobile). Nous réalisons une étude de simulation pour examiner les propriétés de cette modélisation et de cette prévision hybrides ARFIMA-RNA. Nous démontrons concrètement l'utilité du modèle hybride proposé et comparons la précision de la prévision aux modèles actuels.

[14:00-14:30]

You Liang (Ryerson University)

Long Term Interval Forecasts of Demand using Data-Driven Dynamic Regression Models

Prévisions à intervalles à long terme de la demande à l'aide de modèles de régression dynamiques axés sur les données

Long-term electricity load forecasts are the main inputs of production planning at different horizons and load forecasting plays an important role in balancing the electricity grid. Forecast (prediction) intervals provide the measure of uncertainty of the point forecasts (predic-

Les prévisions de la charge électrique à long terme sont les principaux éléments de la planification de la production sur différents horizons et la prévision de la charge joue un rôle important dans le maintien de l'équilibre du réseau électrique. Les intervalles de prévision (prédiction) fournissent la mesure de l'in-

tions). However, a data-driven innovation distribution approach is not available to calculate long-term prediction intervals when using deep learning neural networks dynamic regression (NNDR) models. In this talk, a feedforward NNDR model and an LSTM NNDR model for long-term electricity demand forecasting are introduced and the corresponding data-driven prediction intervals are obtained. It is shown that NNDR models are capable of modelling seasonality as well as nonlinearity directly for co-integrated demand and other features. NNDR models are evaluated through numerical experiments.

certitude des prévisions (prédictions) ponctuelles. Cependant, il n'existe pas d'approche de la distribution de l'innovation axée sur les données pour calculer les intervalles de prédiction à long terme lors de l'utilisation de modèles de régression dynamique de réseaux neuronaux d'apprentissage profond. Dans cette présentation, nous présentons un modèle prédictif de régression dynamique de réseaux neuronaux d'apprentissage profond et un modèle de régression dynamique de réseaux neuronaux d'apprentissage profond à mémoire à long court terme pour la prévision de la demande d'électricité à long terme. Ensuite, nous obtenons des intervalles de prédiction correspondants axés sur les données. Nous montrons que les modèles de régression dynamique de réseaux neuronaux d'apprentissage profond permettent de modéliser la saisonnalité ainsi que la non-linéarité directement pour la demande cointégrée et d'autres caractéristiques. Nous analysons les modèles de régression dynamique de réseaux neuronaux d'apprentissage profond avec des expériences numériques.

[14:30-15:00]

Ruppa K Thulasiram (University of Manitoba) **Japjeet Singh** (University of Manitoba) **Sulalitha Bowala** (University of Manitoba) **Aerambamoorhty Thavaneswaran** (University of Manitoba) **Saumen Mandal** (University of Manitoba)

Hybrid Data-Driven Fuzzy Risk Forecasts for Cryptocurrencies

Prévisions de risque floues reposant sur des données hybrides pour les cryptomonnaies

Volatility forecasting is an integral component of commonly used risk management models, and hence the quality of volatility forecasts greatly impacts the robustness of such models. Recently the interest in cryptocurrencies as an alternative financial asset has exploded. This research describes fuzzy set theory with data-driven volatility and data-driven neuro-volatility forecasts to compute fuzzy risk forecasts. The key underlying idea, unlike the existing risk forecasting models, is the use of hybrid nonlinear adaptive fuzzy model for volatility to account for uncertainty that many existing risk forecasting models do not consider. We used both the traditional as well as recently proposed data-driven EWMA model, and neuro-volatility models to compute volatility forecasts. We report the forecasts of Value at Risk (VaR) and Expected Shortfall (ES) for the top six Cryptocurrencies by market capitalization. The narrower fuzzy intervals imply a better forecast quality. Real data examples show that the data-driven models produce better forecasts for cryptocurrencies, while for the traditional stocks and indices, neuro-volatility model gave better forecasts.

La prévision de la volatilité fait partie intégrante des modèles de gestion des risques souvent utilisés. La qualité des prévisions de la volatilité a donc une grande incidence sur la robustesse de ces modèles. Récemment, l'intérêt pour les cryptomonnaies comme autre actif financier a explosé. Ces travaux décrivent la théorie des ensembles flous et les prévisions de volatilité et de neurovolatilité issues des données pour calculer les prévisions de risque floues. Contrairement aux modèles de prévision des risques actuels, l'idée principale est d'utiliser un modèle flou adaptatif non linéaire hybride de volatilité pour tenir compte de l'incertitude que de nombreux modèles de prévision des risques actuels ne prennent pas en considération. Nous avons utilisé le modèle classique et le modèle proposé de moyenne mobile à pondération exponentielle reposant sur des données, ainsi que les modèles de neurovolatilité pour calculer les prévisions de volatilité. Nous présentons les prévisions de la valeur à risque (VaR) et du déficit attendu (ES) des six premières cryptomonnaies par capitalisation boursière. Les intervalles flous plus restreints entraînent une meilleure qualité de prévision. Des exemples de données réelles montrent que les modèles reposant sur des données produisent de meilleures prévisions pour les cryptomonnaies, tandis que pour les actions et les indices classiques, le modèle de neurovolatilité donne de meilleures prévisions.

Chair/Président: Yanbo Tang

Date: Friday June 3 / vendredi 3 juin

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-13:45]

Mufan Li (University of Toronto) **Sinho Chewi** (Massachusetts Institute of Technology) **Murat A. Erdogdu** (University of Toronto) **Ruoqi Shen** (University of Washington) **Matthew Zhang** (University of Toronto)

Analysis of Langevin Monte Carlo from Poincaré to Log-Sobolev

Analyse de l'algorithme Monte-Carlo de Langevin, de l'inégalité de Poincaré et de l'inégalité de Sobolev logarithmique

Classically, the continuous-time Langevin diffusion converges exponentially fast to its stationary distribution Π under the sole assumption that Π satisfies a Poincaré inequality. Using this fact to provide guarantees for the discrete-time Langevin Monte Carlo (LMC) algorithm, however, is considerably more challenging due to the need for working with chi-squared or Rényi divergences, and prior works have largely focused on strongly log-concave targets. In this work, we provide the first convergence guarantees for LMC assuming that Π satisfies either a Latała–Oleszkiewicz or modified log-Sobolev inequality, which interpolates between the Poincaré and log-Sobolev settings. Unlike prior works, our results allow for weak smoothness and do not require convexity or dissipativity conditions.

Traditionnellement, la diffusion de Langevin en temps continu converge à une vitesse exponentielle vers sa distribution stationnaire Π, si Π satisfait à une inégalité de Poincaré. Cependant, se baser sur ce constat pour garantir l'algorithme de Langevin Monte-Carlo en temps discret est beaucoup plus difficile en raison de la nécessité de travailler avec des divergences chi carré ou de Rényi. De plus, les travaux antérieurs étaient surtout axés sur des cibles à forte concavité logarithmique. Dans le cadre de ces travaux, nous apportons les premières garanties de convergence pour l'algorithme Monte-Carlo de Langevin dans l'hypothèse que Π satisfasse soit une inégalité de Latała-Oleszkiewicz ou une inégalité de Sobolev logarithmique modifiée, qui se situe entre les paramètres de Poincaré et du logarithme de Sobolev. Contrairement aux travaux antérieurs, nos résultats prévoient un faible lissage et ne nécessitent pas de conditions de convexité ou de dissipativité.

[13:45-14:00]

Golara Zafari (Simon Fraser University) **Jean-François Bégin** (Simon Fraser University)

Parametric Inference of Multifactor Stochastic Volatility Models with Variance-Dependent Pricing Kernel

Inférence paramétrique des modèles de volatilité stochastique multifactorielle avec noyau de tarification dépendant de la variance

This study introduces a general class of discrete-time multifactor stochastic volatility dynamics with jumps for which the weak jump-diffusion limit resembles the well-established three-factor model of Andersen, Fusari, and Todorov (2015). Using a general pricing kernel that captures the equity risk premium as well as the risk associated with each volatility factor and jump component, we obtain closed-form model cumulants for the framework. The latter quantities are used to estimate the model, which relies on the combination of efficient filters (i.e., the discrete nonlinear filter and the unscented Kalman filter). A simulation study is conducted to assess the performance of the proposed

Cette étude porte sur une classe générale de dynamique de volatilité stochastique multifactorielle à temps discret et avec sauts pour laquelle la limite faible de diffusion avec saut ressemble au modèle à trois facteurs bien établi de Andersen, Fusari et Todorov (2015). À l'aide d'un noyau d'évaluation général qui capture la prime de risque sur capitaux propres ainsi que le risque associé à chaque facteur de volatilité et composante de saut, nous obtenons des cumulants de modèle sous forme explicite pour cette étude. Les quantités de ces cumulants sont utilisées pour estimer le modèle, qui repose sur la combinaison de filtres efficaces (filtre non linéaire discret et filtre de Kalman sans parfum). Nous réalisons une étude de simulation pour évaluer l'efficacité de la méthode d'estimation proposée. Enfin, nous ajustons le modèle à l'aide d'informations

estimation methodology. Finally, we fit the model using information from the S&P 500 index level and option prices.

provenant du niveau de l'indice S&P 500 et du cours de l'option.

[14:00-14:15]

Roberto Casarin (Ca' Foscari University of Venice) **Mauro Costantini** (University of Aquila, Italy) **Anthony Osuntuyi** (Ca' Foscari University of Venice)

Bayesian nonparametric panel Markov-switching GARCH models

Modèles GARCH bayésiens non paramétriques à changements de régimes markovien

This paper introduces a new model for panel data with Markov-switching GARCH effects to deal with regime changes and temporal clustering of conditional volatility and expected return in a large group of US financial assets. The model incorporates a series-specific hidden Markov chain process that drives the GARCH parameters. To cope with the high-dimensionality of the parameter space, the paper exploits the cross-sectional clustering of the series by Bayesian nonparametric prior distribution. A numerical analysis is carried out to evaluate the performance of the new model. More specifically, some simulation experiments are run along with an empirical application to financial returns data in the US.

Dans cette présentation, nous présentons un nouveau modèle de données de panel avec des effets de GARCH à changements de régimes markovien. L'objectif de cette étude est de traiter les changements de régime et le regroupement temporel de la volatilité conditionnelle et du rendement attendu dans un grand groupe d'actifs financiers américains. Le modèle intègre un processus de chaîne de Markov caché spécifique à la série qui commande les paramètres de l'hétéroscédasticité conditionnelle autorégressive généralisée. Pour faire face à la grande dimensionnalité de l'espace des paramètres, l'étude exploite le regroupement transversal des séries par une distribution a priori non paramétrique bayésienne. Nous effectuons une analyse numérique pour évaluer l'efficacité du nouveau modèle. Plus précisément, nous réalisons des expériences de simulation avec une application empirique aux données de rendements financiers aux États-Unis.

[14:15-14:30]

Mohsen Bahremani (Wilfrid Laurier University) **Xu (Sunny) Wang** (Wilfrid Laurier University)

Modeling Multivariate Hopfield-Transformer Hawkes Process: Application to Sovereign Credit Default Swaps

Modélisation de processus transformateur-Hopfield multivarié de Hawkes : application au contrat d'échange sur défaillance de crédit souverain

In Hawkes Process (HP), the arrival time of a sequence of events relies on the occurrence time of earlier events. Most recent events have the most significant contribution to the intensity, while the effect of previous ones decays over time. However, in multi-dimensional HP, the intensity of future events does not depend only on the history of their process; it also relies on the behavior of other types of events, named a contagious effect. Moreover, many real-world data do not follow HP's assumptions and become more complex to be modeled, so the neural-HP was developed to tackle the challenges. However, they fail to capture long-term dependencies among multiple point processes, and Transformer Hawkes processes only address temporal characteristics of HP. This research aims to develop a more accurate and efficient approach for modeling HP to deal with the challenges by combining HP, Transformer NN, and Hopfield NN, resulting in better accu-

Dans les processus de Hawkes (PH), le temps d'arrivée d'une suite d'événements dépend de quand les événements précédents se produisent. Les événements les plus récents contribuent le plus pertinemment à l'intensité, tandis que l'effet des événements précédents se détériore au fil du temps. Toutefois, dans les PH multidimensionnels, l'intensité des événements à venir ne dépend pas seulement de l'historique de leurs processus, elle dépend aussi du comportement d'autres types d'événements ; c'est ce qu'on appelle un effet contagieux. En outre, de nombreuses données réelles ne suivent pas l'hypothèse du PH et deviennent plus complexes à modéliser. C'est d'ailleurs pour aborder ces défis que le PH neuronal a été conçu. Cependant, ils ont du mal à rendre compte des dépendances à long terme dans des processus à point multiple, et les processus transformateurs de Hawkes ne résolvent que les caractéristiques temporelles du PH. Cette recherche a comme objectif la conception d'une approche précise et efficace pour modéliser le PH et qui résoudra les problèmes en combinant le PH, le transformateur NN et le NN de Hopfield, ce qui procurera

racy, log-likelihood, and RMSE.

une précision, une vraisemblance et un écart-type supérieurs.

[14:30-14:45]

Adam B. Kashlak (University of Alberta) **Giseon Heo** (University of Alberta) **Prachi Loliencar** (University of Alberta)

Topological Hidden Markov Models

Modèles de Markov cachés topologiques

Hidden Markov Models (HMMs) have been applied to many areas of data analysis since their inception almost 60 years ago. However, their reliance on probability density functions fit parameter re-estimation makes them ill-suited for handling data in infinite dimensional spaces. In this talk, we introduce a new approach to modelling data using HMMs where the observed data are realizations of a Gaussian measure on a locally convex topological vector space. This allows for modelling of sequential functional observations or stochastic processes. Our topological HMM is shown to successfully model biological signals data such as electrical responses within muscles via electromyography (EMG) and brain waves via electroencephalogram (EEG).

Les modèles de Markov cachés ont été appliqués à de nombreux domaines de l'analyse des données depuis leur apparition il y a près de 60 ans. Cependant, leur dépendance à l'égard de la réestimation des paramètres des fonctions de densité de probabilité les rend peu adaptés au traitement des données dans des espaces de dimension infinie. Dans cette présentation, nous présentons une nouvelle approche de modélisation des données utilisant des modèles de Markov cachés, dans le contexte où les données observées sont des réalisations d'une mesure gaussienne sur un espace vectoriel topologique localement convexe. Cette méthode permet de modéliser des observations fonctionnelles séquentielles ou des processus stochastiques. Nous montrons que notre modèle de Markov caché topologique modélise avec succès les données de signaux biologiques, tels que l'activité électrique des muscles (à l'aide d'une électromyographie) et les ondes cérébrales (à l'aide d'un électroencéphalogramme).

[14:45-15:00]

Giulia Carallo (Università Ca' Foscari di Venezia) **Roberto Casarin** (Ca' Foscari University of Venice) **Christian P. Robert** (Université Paris-Dauphine)

Generalized Poisson Difference Autoregressive Processes

Processus autorégressifs de la différence du modèle généralisé de Poisson

This paper introduces a new stochastic process with values in the set Z of integers with sign. The increments of the process are Generalized Poisson differences and the dynamics has an autoregressive structure. We study the properties of the process and exploit the thinning representation to derive stationarity conditions, the stationary distribution of the process and its conditional and unconditional moments. We provide a Bayesian inference framework and an efficient posterior approximation procedure based on Markov Chain Monte Carlo. Numerical illustrations on simulated data show the effectiveness of the proposed inference. The applications to accidents data and cyber threats data show that the proposed model is well suited for capturing persistence in the conditional moments and in the over-dispersion feature of the data.

Dans cette présentation, nous présentons un nouveau processus stochastique dont les valeurs se situent dans l'ensemble Z des nombres entiers comportant un signe. Les incréments du processus sont des différences du modèle généralisé de Poisson, et la dynamique a une structure autorégressive. Nous analysons les propriétés du processus et exploitons la représentation d'amincissement pour obtenir des conditions de stationnarité, la distribution stationnaire du processus et ses moments conditionnels et inconditionnels. Nous proposons un cadre d'inférence bayésienne et une procédure efficace d'approximation a posteriori reposant sur la méthode de Monte-Carlo par chaînes de Markov. Des illustrations numériques sur des données simulées montrent l'efficacité de l'inférence proposée. Les applications aux données d'accidents et de cybermenaces montrent que le modèle proposé est bien adapté pour saisir la persistance dans les moments conditionnels et dans la caractéristique de surdispersion des données.

Chair/Président: Candemir Cigsar

Date: Friday June 3 / vendredi 3 juin

Time/Heure: 13:30-15:00

Abstract/Résumé

[13:30-13:45]

James A. Hanley (McGill University) **Maryse Kochoedo** (McGill University) **Rajib Dey** (McGill University) **Wilber Deck** (Direction de Santé Publique, Gaspé)

Measuring the Numbers of Lung Cancer (LC) Deaths Averted by Screening

Mesure du nombre de décès dus au cancer du poumon évités par le dépistage

Across follow-up (FU) time, screening-induced mortality reductions are zero initially, reach a maximum after a few years, and then subside. A 2021 meta-analysis calculated a 16% reduction (%R) in LC mortality, while noting that including deaths up to 10y after the last screen biases the %R towards 0. In the largest trial, screens ended at 2y; by 6.5y the screening arm had 87 fewer LC deaths (20%R). By 12.5 y, it was 89; but the overall %R was now only 8%. Including FU where no benefit is expected dilutes the %R, but doesn't change the difference in the numbers of deaths. Over all 9 trials (with 1-10 rounds), 162 deaths were averted by 154K screens (1/950). The difference in LC deaths is unaffected by the no.s of screens and the amount of FU after the effect of the last screen has subsided. Thus, calculations based on it may provide a more robust and more appropriate summary. We address the use of this measure and implications for the size & FU duration of cancer screening trials.

Dans le temps de suivi, la réduction de mortalité entrainée par le dépistage est initialement nulle, atteint un pic après quelques années, et puis décline. Une méta-analyse en 2021 a démontré une réduction (%R) de 16% de la mortalité due au cancer du poumon (CP), tout en notant que l'inclusion des décès jusqu'à 10 ans après le dernier dépistage biaise le %R vers 0. Dans le plus grand essai, le dépistage cessait après 2 ans ; après 6,5 ans le groupe dépisté avait 87 décès de moins dus au CP (20%R). Après 12,5 ans, c'était 89 ; mais alors le %R était de 8%. L'inclusion du temps de suivi où aucun bénéfice n'est attendu dilue le %R, mais ne change pas la différence du nombre de décès. Sur les 9 essais (ayant 1-10 rondes), 162 décès ont été avertis par 154k dépistages (1/950). La différence en décès dus au CP n'est pas influencée par le nombre de dépistages ni par la durée du suivi après que l'effet du dernier dépistage soit redescendu. Ainsi, des calculs basés sur la différence peuvent fournir un résumé plus robuste et plus approprié. Nous examinons l'usage de cette mesure et ses implications pour la taille des essais de dépistage du cancer et leur durée de suivi.

[13:45-14:00]

Jennifer McNichol (University of Victoria) **Connie Stewart** (University of New Brunswick Saint John)

Simultaneous Maximum Unified Fatty Acid Signature Analysis

Analyse simultanée de la signature maximale unifiée des acides gras

Quantitative fatty acid signature analysis (QFASA) has been the cornerstone of dietary estimation for marine predators since its introduction in 2004. However, QFASA relies upon calibration coefficients (CCs) to account for the differences in fatty acids between a predator and its prey. CCs are determined by way of captive feeding studies and must be uniquely determined for each species of predator. This is a major limitation for QFASA since CCs have not been determined for all predators. There has been a growing interest

L'analyse quantitative de la signature des acides gras (AQSAG) est la pierre angulaire de l'estimation du régime alimentaire des prédateurs marins depuis son introduction en 2004. Cependant, l'AQSAG repose sur des coefficients d'étalonnage (CE) pour tenir compte des différences d'acides gras entre un prédateur et sa proie. Les CE sont déterminés par le biais d'études d'alimentation en captivité et doivent être déterminés de manière unique pour chaque espèce de prédateur. Ceci constitue une limitation majeure pour l'AQSAG puisque les CE n'ont pas été déterminés pour tous les prédateurs. On s'intéresse de plus en plus

in developing a QFASA alternative that does not rely upon predetermined CCs. Some evidence suggests that CCs may be estimated alongside diets with little trade off in accuracy as compared to QFASA with feeding study derived CCs. In this work we propose a new QFASA-type method for diet (and CC) estimation via a maximum likelihood approach which does not rely on feeding study derived CCs. The results of a simulation study are presented, along with a real-life example.

au développement d'une alternative à l'AQSAG qui ne repose pas sur des CE prédéterminés. Certaines preuves suggèrent que les CE peuvent être estimés en même temps que les régimes alimentaires avec peu de perte de précision par rapport à l'AQSAG avec des CE dérivés d'études d'alimentation. Dans ce travail, nous proposons une nouvelle méthode de type AQSAG pour l'estimation des régimes (et des CE) via une approche du maximum de vraisemblance qui ne repose pas sur les CE dérivés de l'étude d'alimentation. Nous présentons les résultats d'une étude de simulation, ainsi qu'un exemple concret.

[14:00-14:15]

Xiaoqing Zhang (University of Regina) **Dianliang Deng** (University of Regina)

Lindley Binomial Model: A Flexible Approach for Modelling the Proportions with Sparseness and Excessive zeros

Modèle binomial de Lindley : une approche souple pour la modélisation des proportions lors de dispersion et de surreprésentation des zéros

In this paper, we present a new modeling approach for the proportions with sparseness and excessive zeros. The distribution of proportional data typically exhibits overdispersion, zero-inflation and sparseness, and heavy tails. We propose a new Lindley binomial distribution, by compounding the two-parameter Lindley family of distributions with the binomial distribution. This distribution can flexibly handle each of the aforementioned features of proportional data. We study the probabilistic properties of this distribution such as moment, moment generating function, and develop a computational approach to accurately evaluate the likelihood of the proposed model and to perform the penalized maximum likelihood estimation via the EM algorithm. We assess the performance of our developed algorithm for the estimation of parameters in the proposed model with/without covariates and demonstrate the application to Incidence of Hepatitis A and Yellow Fever data.

Notre exposé propose une nouvelle approche de modélisation des proportions lors de dispersion et de surreprésentation des zéros. La distribution de données proportionnelles présente généralement une surdispersion, une surreprésentation de zéros et une rareté des données ainsi que des queues lourdes. Nous proposons une nouvelle distribution binomiale de Lindley, en combinant la famille de distributions Lindley à deux paramètres avec la distribution binomiale. Cette distribution peut traiter avec souplesse chacune des caractéristiques précitées des données proportionnelles. Nous étudions les propriétés probabilistes de cette distribution, comme ses moments, la fonction génératrice de moments et nous développons une approche computationnelle pour évaluer avec exactitude la vraisemblance du modèle proposé et procéder à une estimation par le maximum de vraisemblance pénalisée en utilisant l'algorithme espérance-maximisation. Nous évaluons la performance de l'algorithme que nous avons développé pour l'estimation des paramètres dans le modèle proposé avec et sans covariables et en illustrons l'application avec des données sur l'incidence de l'hépatite A et de la fièvre jaune.

[14:15-14:30]

Philip J. Schmidt (University of Waterloo) **Ellen Cameron** (University of Waterloo) **Kirsten Muller** (University of Waterloo) **Monica Emelko** (University of Waterloo)

Amplicon Sequencing Diversity Analysis: Multinomial Models and Variants You Don't Know You Didn't See

Analyse de la diversité du séquençage d'amplicons : modèles multinomiaux et variantes que vous ne savez pas que vous n'avez pas vues

Diversity analysis of amplicon sequencing data is mainly limited to a pipeline of deterministic calculations that does not reflect the probabilistic process generating observed data or a century of fundamentals of quantitative microbiology that intrigued Student, Fisher, and others. The random errors in the process generating

L'analyse de la diversité des données de séquençage d'amplicons se limite principalement à un pipeline de calculs déterministes qui ne reflète ni le processus probabiliste générant les données observées ni le siècle de principes fondamentaux en microbiologie quantitative qui ont intrigué Student, Fisher et d'autres. Nous discutons des erreurs aléatoires dans le processus générant les

amplicon sequencing data and their probabilistic representation are discussed, noting several errors that compromise the basic multinomial model for arrays of sequence variant counts. Presuming a multinomial model, simulation experiments show the bias in probabilistic estimation of the Shannon index resulting from a type of zeros omitted from previous studies: unobserved zeros. Unbiased estimation of source diversity is concluded to be impossible unless the number of sequence variants in the source is known a priori. Performance of repeated rarefying to provide sample-level diversity analysis conditional on a particular library size is evaluated via simulation.

données de séquençage d'amplicons et leur représentation probabiliste, en notant plusieurs erreurs qui compromettent le modèle multinomial de base pour les réseaux de comptage des variantes de séquences. En supposant un modèle multinomial, nous montrons par des expériences de simulation les biais de l'estimation probabiliste de l'indice de Shannon qui résultent d'un type de zéros omis dans les études précédentes : les zéros non observés. On conclut que l'estimation non biaisée de la diversité de la source est impossible à moins que le nombre de variantes de séquences dans la source soit connu a priori. Nous évaluons par simulation la performance de la raréfaction répétée pour fournir une analyse de la diversité au niveau de l'échantillon conditionnelle à une taille de bibliothèque particulière.

[14:30-14:45]

Jonathan Jalbert (Polytechnique Montreal) **Gamet Philémon** (Polytechnique Montréal)

A flexible extended generalized Pareto distribution for tail estimation

Loi de Pareto généralisée étendue pour la modélisation des valeurs extrêmes

In environmental applications, tail distributions often correspond to extreme risks and an accurate modelling is mandatory. The peaks-over-threshold model is a classic way to model the exceedances over a high threshold with the generalized Pareto distribution. However, for some applications, the choice of a high threshold is challenging and the asymptotic conditions for using this model are not always satisfied. The class of extended generalized Pareto models can be used for lower thresholds when the asymptotic conditions are not met. We propose new extensions of the generalized Pareto distribution suitable for low threshold. The proposed extensions provide better estimate of the tail index for low thresholds than existing models. They are also appropriate for high thresholds because in that case, the extended models simplifies to the generalize Pareto model. Finally, this new distributions is used to model extreme temperatures and precipitation in Montreal.

Dans les applications environnementales, les valeurs extrêmes constituent souvent des événements qui menacent la sécurité des du public et leur modélisation précise est primordiale. L'approche classique pour étudier les valeurs extrêmes consiste à modéliser les excédents au-dessus d'un seuil élevé avec la loi de Pareto généralisée. Cependant, pour certaines applications, le choix d'un seuil est difficile et les conditions asymptotiques d'utilisation de ce modèle ne sont pas toujours satisfaites. Dans ce cas, la loi de Pareto généralisée étendue peut être utilisée pour des seuils bas. Dans cette présentation, nous proposons de nouvelles extensions à la loi de Pareto généralisée adaptées à des seuils bas. Les extensions proposées fournissent notamment une meilleure estimation de l'indice de queue pour les seuils bas que les modèles existants. Enfin, ces nouvelles extensions sont utilisées pour modéliser les températures et les précipitations extrêmes à Montréal.

[14:45-15:00]

Michaël Lalancette (University of Toronto) **Sebastian Engelke** (Université de Genève) **Stanislav Volgushev** (University of Toronto)

Inference for Extremal Graphical Models

Inférence pour les modèles graphiques extrémaux

Multiple characterizations and models exist for extremal dependence, the dependence structure of multivariate data in unobserved tail regions. However, statistical inference for extremal dependence uses merely a fraction of the available data, drastically reducing the effective sample size and creating challenges even in mod-

Plusieurs modèles existent pour la dépendance extrémale, c'est-à-dire la structure de dépendance de données multivariées dans les queues. Cependant, l'inférence statistique pour la dépendance extrémale n'utilise qu'une fraction des données, réduisant ainsi la taille échantillonnale et compliquant l'estimation même en dimension modérée. Récemment introduits, les modèles graphiques

erate dimension. Recently introduced graphical models for multivariate extremes allow for enforced sparsity in moderate- to high-dimensional settings, reducing the effective dimension. In this work, we propose a novel, scalable method for selection of extremal graphical models that makes no assumption on the underlying graph structure, as opposed to existing approaches. It exploits existing tools for Gaussian graphical model selection such as the graphical lasso and neighborhood selection. Model selection consistency is established in sparse regimes where the dimension is allowed to be exponentially larger than the effective sample size.

extrémaux permettent d'imposer une parcimonie lorsque la dimension est élevée, réduisant ainsi la dimension effective. Nous proposons une méthode novatrice pour la sélection de modèles graphiques extrémaux qui ne requiert aucune hypothèse sur le graphe sous-jacent, contrairement aux approches existantes. Elle exploite des outils existants pour la sélection de modèles graphiques gaussiens tels que le lasso graphique et la sélection de voisinage. L'exactitude asymptotique de la méthode est établie dans des régimes parcimonieux où la dimension peut être exponentiellement plus grande que la taille échantillonnale.

Methodological Advances in Classification Models for Complex Longitudinal Data
Avancées méthodologiques des modèles de classification pour données longitudinales complexes

Chair/Président: Tolu Sajobi

Organizer/Responsable: Tolu Sajobi

Date: Friday June 3 / vendredi 3 juin

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-15:52]

Anuradha Roy (The University of Texas at San Antonio) **Ricardo Leiva** (Universidad Nacional de Cuyo, Argentina)

Linear discrimination for three-level multivariate data

Discrimination linéaire des données multivariées à trois niveaux

We study a linear discriminant function for three-level m-variate observations under the assumption of multivariate normality. We assume that the m-variate observations have a doubly exchangeable covariance structure. The new discriminant function is very efficient in discriminating individuals in a small sample scenario, which is the case in many medical and biomedical studies. An iterative algorithm is proposed to calculate the MLEs of the unknown population parameters. The new discriminant function is applied to a real data as well as to simulated data. We compare our findings with other linear discriminant functions for three-level multivariate data including the traditional linear discriminant function. Error rates of the proposed classification rule is found to be much less than the error rates of the traditional classification rule, when in fact the traditional classification rule fails most of the time owing to the small sample sizes. This is joint work with Ricardo Leiva.

Nous analysons une fonction discriminante linéaire pour les observations de variables aléatoires m à trois niveaux dans l'hypothèse d'une normalité multivariée. Nous partons du principe que les observations des variables aléatoires m ont une structure de covariance doublement échangeable. La nouvelle fonction discriminante est très efficace pour discriminer les individus dans un cadre de petit échantillon, ce qui est le cas dans de nombreuses études médicales et biomédicales. On propose un algorithme itératif pour calculer les estimations du maximum de vraisemblance des paramètres inconnus de la population. On applique la nouvelle fonction discriminante à des données réelles et simulées. Nous comparons nos résultats à d'autres fonctions discriminantes linéaires de données multivariées à trois niveaux, ainsi qu'à la fonction discriminante linéaire traditionnelle. Les taux d'erreurs de la règle de classification proposée s'avèrent bien inférieurs aux taux d'erreurs de la règle de classification classique. En effet, la règle de classification classique échoue la plupart du temps en raison de la petite taille des échantillons. Il s'agit de travaux conjoints avec Ricardo Leiva.

[15:52-16:14]

Anita Brobbey (University of Calgary) **Lisa M. Lix** (University of Manitoba) **Alberto Nettel-Aguirre** (University of Wollongong) **Tyler Williamson** (University of Calgary) **Samuel Wiebe** (University of Calgary) **Tolu Sajobi** (University of Calgary)

Repeated Measures Discriminant Analysis using Generalized Estimating Equations

Analyse discriminante des mesures répétées utilisant des équations d'estimation généralisées

Discriminant analysis (DA) procedures have been developed for classification in multivariate repeated measures data. However, these models rely on the assumption of multivariate normality and are less accurate in non-normal data. This study developed discriminant analysis based on generalized estimating equations (DA-GEE) and examined its accuracy using Monte Carlo methods. The DA-GEE resulted in at least 5% higher

Les procédures d'analyse discriminante (DA) ont été développées pour la classification des données multivariées de mesures répétées. Ces modèles dépendent toutefois de l'hypothèse de normalité multivariée et sont moins exacts que pour des données non normales. Cette étude a permis de développer une analyse discriminante basée sur des équations d'estimation généralisées (DA-GEE) et d'examiner son exactitude à l'aide de méthodes de Monte-Carlo. La DA-GEE a donné comme résultat une exactitude de la clas-

mean overall classification accuracy than DA based on MLE in multivariate non-normal data. Impact of correlation misspecification was largely influenced by population distribution. Misspecification of the correlation structure resulted in at least 1% decrease in classification accuracy under multivariate normal data and some non-normal data. DA-GEE models are useful for prediction/classification in multivariate repeated measures designs.

sification générale d'au moins 5 % supérieure à celle d'une DA basée sur un estimateur du maximum de vraisemblance de données non normales multivariées. La distribution de la population a largement influé sur l'impact de l'erreur de spécification des corrélations. Le résultat de cette erreur de spécification de la structure de la corrélation était une exactitude de la classification réduite d'au moins 1 % sous des données normales multivariées et quelques données non normales. Les modèles de DA-GEE sont utiles pour la prédiction ou classification dans les concepts de mesures répétées multivariées.

[16:14-16:36]

David Hughes (University of Liverpool)

Dynamic Longitudinal Discriminant Analysis Using Multiple Longitudinal Markers of Different Types

Analyse discriminante longitudinale et dynamique au moyen de marqueurs longitudinaux multiples de différents types

There is an emerging need in clinical research to accurately predict patients' disease status and disease progression by optimally integrating multivariate clinical information. Clinical data are often collected over time for multiple biomarkers of different types (e.g. continuous, binary and counts). In this talk I will describe some recently developed longitudinal discriminant analysis methods utilising multiple longitudinal biomarkers. We utilize multivariate generalized linear mixed models with flexible random effects distributions. These longitudinal models are subsequently used in a multivariate time-dependent discriminant scheme to predict, at any time point, the probability of belonging to a particular risk group. I will describe the methods and briefly show their use in a variety of clinical applications including epilepsy, diabetes, and liver cancer.

Il y a un besoin émergent dans la recherche clinique de prédire avec précision l'état et la progression d'une maladie d'un patient en intégrant de façon optimale les renseignements cliniques multivariés. Les données cliniques sont souvent recueillies au fil du temps par des biomarqueurs de différents types (p. ex. continus, binaires et de dénombrement). Dans cet exposé, je décrirai certaines méthodes d'analyse discriminante longitudinale conçues récemment se servant de biomarqueurs longitudinaux multiples. Nous utilisons des modèles linéaires mixtes généralisés multivariés avec des distributions d'effets aléatoires flexibles. Ces modèles longitudinaux sont subséquemment employés dans un schéma discriminant multivarié dépendant dans le temps pour prédire la probabilité d'appartenance à un certain groupe risque à n'importe quel point dans le temps. Je décrirai ces méthodes et montrerai brièvement leur utilisation dans une variété d'applications cliniques comme l'épilepsie, le diabète et le cancer du foie.

[16:36-16:58]

Jeffrey L. Andrews (University of British Columbia, Okanagan) **Ryan P. Browne** (University of Waterloo) **Liam Welsh** (University of Toronto)

Finite mixture models for longitudinal data with dynamic group membership

Modèles de mélanges finis pour les données longitudinales présentant une appartenance dynamique à un groupe

We introduce a compositional approach for the building and fitting of a finite Gaussian mixture model, permitting highly constrained components to be added to the model at very low cost with respect to growth in free parameters. The explicit goal of this approach is to enable both the detection and modelling of small numbers of observations which change groups over time in longitudinal data — all under a fully unsupervised paradigm. The proposed approach can be considered an alternative to others in the literature which rely on hidden Markov models to achieve a similar effect. We provide both sim-

Nous présentons une approche compositionnelle pour l'élaboration et l'ajustement d'un modèle de mélange gaussien fini afin de rajouter à ce modèle des composantes soumises à de fortes contraintes, à un coût très faible par rapport à la croissance des paramètres libres. L'objectif de cette approche est à la fois de détecter et de modéliser un petit nombre d'observations qui changent de groupe dans des données longitudinales au fil du temps, le tout dans un paradigme entièrement dépourvu de supervision. L'approche que nous proposons peut être considérée comme une solution de rechange à d'autres approches de la littérature qui reposent sur des modèles de Markov cachés pour obtenir un effet semblable. Nous

ulations and real data applications for illustrative purposes.

illustrons notre démarche par des simulations et des applications de données réelles.

Fairness in Data-driven Research
Recherche fondée sur les données et équité

Chair/Président: Sanjeena Dang

Organizer/Responsable: Sanjeena Dang

Date: Friday June 3 / vendredi 3 juin

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-16:00]

Veronique Tremblay (Beneva / HEC Montréal)

Responsible use of algorithms in decision making: ethical principles and recommendations

Utilisation responsable des modèles dans la prise de décision : principes éthiques et recommandations

In recent years, potentially discriminatory uses of predictive models have made headlines. These widely publicized cases have led several researchers to examine the issue of algorithmic fairness. Much of the available scientific literature on algorithmic fairness comes from computer science, law, and ethics. The scarcity of statisticians in the ensuing search for solutions is quite surprising since, as we will see during the talk, many of the problems raised by algorithmic fairness are well known to statisticians. The first part of the talk is an interdisciplinary (law, philosophy, computer science, statistics) review on algorithmic fairness. The second part of the presentation will focus on a list of recommendations for practitioners and researchers. The goal of the talk is to highlight potential contributions of statisticians to solving the problem. The work done differs from the current literature in that it is interdisciplinary and focused on industry practice.

Au cours des dernières années, des cas largement médiatisés d'algorithmes discriminatoires ont fait la une et ont conduit plusieurs chercheurs à se pencher sur la notion d'équité algorithmique. Une grande partie de la littérature scientifique disponible sur le thème provient de l'informatique, du droit et de l'éthique. La rareté des statisticiens dans la recherche sur le sujet est assez surprenante car, comme nous le verrons dans l'exposé, de nombreux problèmes soulevés par l'équité algorithmique sont bien connus des statisticiens. La première partie de l'exposé est une revue interdisciplinaire (droit, philosophie, informatique, statistique) sur le thème. La deuxième partie de la présentation se concentrera sur une liste de recommandations pour les praticiens et les chercheurs. L'objectif de l'exposé est de souligner la contribution potentielle des statisticiens sur le sujet. Le travail effectué diffère de la littérature actuelle par son caractère interdisciplinaire et très appliqué.

[16:00-16:30]

David R. Hunter (Pennsylvania State University)

Gratz v. Bollinger and Statistical Machine Learning

Gratz contre Bollinger et l'apprentissage automatique statistique

The 2003 United States Supreme Court case known as Gratz v. Bollinger addresses a formula for college admissions that was created by a statistics graduate student (the presenter) for a specific purpose. The method used to create this formula is considered simplistic by modern machine learning standards; yet the debate that ensued, which could not have happened if a more modern approach had been used, illustrates that science and society do not always benefit from machine learning models that achieve the best possible predictive performance.

L'affaire Gratz contre Bollinger, jugée par la Cour suprême des États-Unis en 2003, porte sur une formule d'admission dans les universités. Cette formule a été créée par un étudiant diplômé en statistiques (le présentateur) dans un but précis. La méthode utilisée pour créer cette formule est considérée comme simpliste par rapport aux normes modernes d'apprentissage automatique. Pourtant, le débat qui a suivi (qui n'aurait pas pu avoir lieu si une approche plus moderne avait été utilisée) montre que la science et la société ne tirent pas toujours profit des modèles d'apprentissage automatique qui obtiennent les meilleurs résultats possible

This talk discusses the history of the legal case, the admissions formula and how it was created, and the implications of the debate for how we build predictive models.

en matière de prédiction. Cette présentation traite de l'histoire de l'affaire judiciaire, de la formule d'admission et de la façon dont cette formule a été créée, ainsi que des répercussions du débat sur la façon dont nous créons les modèles prédictifs.

[16:30-17:00]

Warut Khern-am-nuai (McGill University)

Addressing Fairness in Machine Learning Predictions: Strategic Best-Response Fair Discriminant Removed Algorithm

Aborder la justesse dans les prédictions d'apprentissage automatique : Algorithme stratégique de meilleure réponse juste discriminant éliminé

Discrimination in machine learning (ML) has become a prominent issue. Although many algorithms are designed to address such discrimination issues, virtually all of them focus on alleviating the disparity in prediction results. However, the algorithms do not consider behavioral responses of prediction subjects. So, even if disparity in prediction results can be removed, disparity in behavior may persists across different subpopulations of prediction subjects. To study this issue, we define a new notion called strategic best-response fair (SBR-fair). Even if an algorithm is trained on biased data, will it lead to identical equilibrium behaviors of subpopulations? If yes, we define the ML as SBR-fair. We then demonstrate that many fair ML algorithms in the literature are not SBR-fair. As a result, implementing these algorithms may impose fairness at prediction results, but actually induce disparity between privileged and unprivileged individuals in the long run.

La discrimination dans l'apprentissage automatique (AA) est devenue un problème important. Bien que plusieurs algorithmes soient conçus pour résoudre de tels problèmes de discrimination, ils se concentrent pratiquement tous à réduire l'écart entre les résultats de prédiction. Toutefois, les algorithmes ne tiennent pas compte des réponses comportementales des sujets de prédictions. Donc même si l'on peut éliminer l'écart dans les résultats de prédiction, les écarts de comportement peuvent persister à travers différentes sous-populations de sujets de prédiction. Afin d'étudier le problème, nous définissons une nouvelle notion que l'on surnomme ≪meilleure réponse stratégique juste(MRS-juste)≫. Même si un algorithme est formé à partir de données biaisées, mènera-t-il à des comportements d'équilibre identiques dans les sous-populations ? Si oui, nous définissons l'AA par MRS-juste. Nous démontrons ensuite que plusieurs algorithmes d'AA juste documentés ne sont pas ROS-juste. Conséquemment, l'intégration de ces algorithmes peut imposer une justesse dans les résultats de prédiction, mais aussi causer un écart entre les individus privilégiés et non privilégiés à long terme.

Chair/Président: Francois Brisebois

Organizer/Responsable: Francois Brisebois

Date: Friday June 3 / vendredi 3 juin

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-16:00]

Brady West (University of Michigan)

New measures for assessing non-ignorable selection bias in non-probability samples and low response rate probability samples

Nouvelles mesures pour évaluer le biais de sélection important des échantillons non probabilistes et probabilistes à faible taux de réponse

Recent developments in survey statistics have yielded simple, novel measures of the non-ignorable selection bias in estimates of means, proportions, and regression coefficients that may arise due to deviations from ignorable sample selection, where these deviations might be introduced by the sampling mechanism (e.g., non-probability sampling) or survey nonresponse. Responsive survey designs rely on active monitoring of sound indicators of survey errors to inform real-time design decisions, and these new measures, which are easy to compute at any point in time during a data collection, have the potential to serve as useful indicators of the possible selection bias in estimates of interest. This presentation will review the computation of these indicators, the data required to compute them, software tools for computing them, and how they might be actively monitored in real time to inform design decisions in responsive survey designs.

Dans le domaine des statistiques d'enquête, des progrès récents ont permis de découvrir de nouvelles mesures simples du biais de sélection important dans les estimations des moyennes, des proportions et des coefficients de régression, qui peuvent être causées par des variations de la sélection des échantillons dont il ne faut pas tenir compte, et dans lesquelles ces variations peuvent être introduites par le mécanisme d'échantillonnage (par exemple, l'échantillonnage non probabiliste) ou la non-réponse à l'enquête. Les plans d'enquête réactifs reposent sur la surveillance active d'indicateurs solides d'erreurs d'enquête afin d'éclairer les décisions de conception en temps réel. Ces nouvelles mesures, qui sont faciles à calculer à tout moment au cours d'une collecte de données, peuvent servir d'indicateurs utiles du biais de sélection possible dans les estimations en question. Au cours de cette présentation, nous passerons en revue le calcul de ces indicateurs, les données requises et les outils logiciels permettant de les calculer, ainsi que la manière dont ils pourraient être surveillés activement en temps réel pour éclairer les décisions relatives aux plans d'enquête réactifs.

[16:00-16:30]

Yajuan Si (University of Michigan)

A Case Study of Nonresponse Bias Analysis in Educational Assessment Surveys

Étude de cas de l'analyse du biais de non-réponse dans les enquêtes d'évaluation de l'éducation

Nonresponse bias is a widely prevalent problem for data on education. We develop a ten-step exemplar to guide nonresponse bias analysis (NRBA) in cross-sectional studies and apply these steps to the Early Childhood Longitudinal Study, Kindergarten Class of 2010-11. A key step is the construction of indices of nonresponse bias based on proxy pattern-mixture models for survey variables of interest. A novel feature is to character-

Le biais de non-réponse est un problème largement répandu pour les données sur l'éducation. Nous élaborons un modèle en dix étapes pour guider l'analyse du biais de non-réponse dans les études transversales, puis nous appliquons ces étapes à une étude longitudinale sur la petite enfance (classe maternelle de 2010 à 2011). La création d'indices de biais de non-réponse basés sur des modèles de mélange de profils par procuration pour les variables d'enquête concernées représente une étape clé. Une nouveauté

ize the strength of evidence about nonresponse bias contained in these indices, based on the strength of the relationship between the characteristics in the nonresponse adjustment and the key survey variables. Our NRBA improves existing methods by incorporating both missing at random and missing not at random mechanisms, and all analyses can be done straightforwardly with standard statistical software.

consiste à caractériser la force de l'évidence du biais de non-réponse contenue dans ces indices, sur la base de la force de la relation entre les caractéristiques des variables d'ajustement de non-réponse et les variables clés de l'enquête. Notre analyse du biais de non-réponse améliore les méthodes actuelles par l'intégration de mécanismes de données manquantes aléatoirement et non aléatoirement. Par ailleurs, toutes les analyses peuvent être effectuées directement à l'aide de logiciels de statistiques standards.

[16:30-17:00]

Peter G. Wright (Statistics Canada) **Patrice Martineau** (Statistics Canada) **François Brisebois** (Statistics Canada)

Improving Response by Studying Citizen Participation in Social Surveys

Amélioration de la réponse en examinant la participation citoyenne aux enquêtes sociales

Like other national statistical agencies Statistics Canada faces many challenges, including a downward trend in response rates to social surveys. Over the years several strategies, tools and methods have been implemented to address the issue of reduced response rates. Statistics Canada plans to explore avenues that go beyond the traditional framework of prioritization during collection and methodological adjustments using auxiliary data during estimation. The research initiative, labelled as the Project to study citizen participation, encompasses qualitative studies to improve our understanding of the motivations to respond, quantitative studies to improve our estimation and to reduce error, and quantitative studies aimed at improving the estimation methods in place, using auxiliary data or follow-up surveys. This presentation will outline Statistics Canada's multi-year plan to seek solutions that prevent nonresponse, to manage nonresponse and to correct for nonresponse.

Comme d'autres agences statistiques nationales, Statistique Canada fait face à de nombreux défis, y compris des taux de réponse décroissants aux enquêtes sociales. Au fil des années, plusieurs stratégies, outils et méthodes ont été mis en place pour remédier à cette situation. Statistique Canada veut explorer des pistes qui vont au-delà du cadre traditionnel de priorisation lors de la collecte et de l'exploitation de données auxiliaires lors de l'estimation. Cette initiative, Le projet d'étude de la participation citoyenne, comprend des études qualitatives permettant de mieux comprendre les motivations à répondre, des études quantitatives servant à améliorer nos estimations et à réduire l'erreur, et des études quantitatives visant à améliorer les méthodes d'estimation en place, en utilisant des données auxiliaires ou des enquêtes de suivi. Cette présentation exposera le plan pluriannuel de Statistique Canada qui cherche des solutions pour prévenir, gérer et corriger la non-réponse.

Maintaining Relevancy Through New Tools, Data Science and Data Visualizations
Maintien de la pertinence grâce à de nouveaux outils, à la science des données et aux visualisations de données.

Chair/Président: Beatrice D. Baribeau

Organizer/Responsable: Beatrice D. Baribeau

Date: Friday June 3 / vendredi 3 juin

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-16:00]

Peter Solymos (E Source) **Khalid Lemzouji** (Analythium Solutions)

Best Practices for Delivering Applied Statistics from Concept to Production

Les meilleures pratiques de livraison de statistiques appliquées, de la conception à la production

Modern applied statistics involve communicating the results to various audiences. This communication increasingly takes place in interactive media rather than status reports. Traditional education for statisticians does not adequately prepare applied scientists for effectively handling such requirements. However, healthy exposure to software engineering skills and practices can greatly facilitate the timely delivery of results. This is due to the shorter time to working prototypes, shorter feedback loops involving stakeholders, and easier communication with IT/engineering when it comes to scale and performance. We will use real-world examples to show how to organize analysis code that also lays the foundation for quickly building prototypes and user interfaces.

La statistique appliquée moderne comprend la communication de résultats à plusieurs publics. Cette communication se fait de plus en plus par l'entremise de médias interactifs plutôt que des rapports de situation. L'enseignement traditionnel en statistiques ne prépare pas adéquatement les scientifiques à gérer ces exigences efficacement. Toutefois, une exposition à la pratique et à l'apprentissage du génie logiciel peut grandement faciliter la distribution de résultats. Les raisons principales étant : arriver plus rapidement à des prototypes fonctionnels, raccourcir la boucle de rétroaction pour les intervenants, et faciliter la communication avec l'ingénieur en TI en ce qui concerne l'échelle et la performance. Nous utiliserons des exemples réels pour montrer comment organiser le code d'analyse servant de base pour construire rapidement des prototypes et des interfaces d'utilisateur.

[16:00-16:30]

Kenneth C.K. Chu (Statistics Canada)

Spaceborne Radar Earth Observation (Big) Data, Emerging Opportunities for Statisticians and Data Scientists

(Méga)données d'observation terrestre par radar spatioporté, occasions émergentes pour les statisticiens et les scientifiques des données

In this presentation, I will give an account of a recent and very fruitful collaboration of mine as an applied statistician with research geographers applying functional principal component analysis to extract temporal trends from radar satellite time series data for land cover classification. This is an example of the emerging collaboration opportunities between statisticians and natural scientists to develop suitable analytical techniques for spaceborne Synthetic Aperture Radar (SAR) Earth Observation data, which are spatiotemporal Big Data. I will close with an overview of some of the challenges/opportunities for exciting and potentially impactful collaboration in this relatively new area for statisti-

Dans cette présentation, je rendrai compte d'une collaboration récente et très fructueuse en tant que statisticien appliqué avec des chercheurs en géographie, dans le cadre de laquelle l'analyse en composantes principales fonctionnelles a été appliquée pour extraire les tendances temporelles des données de séries chronologiques de satellites radar pour la classification de couverture terrestre. Il s'agit d'un exemple des nouvelles opportunités de collaboration entre les statisticiens et les spécialistes des sciences naturelles afin de développer des techniques analytiques appropriées pour les données d'observation terrestre par radar à synthèse d'ouverture (RSO) spatioporté, qui sont des mégadonnées spatiotemporelles. Je terminerai par un aperçu de certains défis et occasions de collaboration intéressants et potentiellement importants

cians and data scientists.

dans ce domaine relativement nouveau pour les statisticiens et les scientifiques des données.

[16:30-17:00]

Martin Monkman (BC Stats, Province of British Columbia)

Continuous Learning in Times of Continuous Change

L'apprentissage continu en période de changement permanent

It's a universally accepted truism that we live in a time of unprecedented change, both in terms of the magnitude and the pace of those changes. In the world of statistics, there have been enormous changes at the intersection of tools, methodologies and techniques, and transparency. New software packages, new ways to analyze data, and new open data arrive every day, each of which have implications for how we approach our professional development. In this talk, I will present observations drawn from applied statistical practice and teaching mid-career professionals that suggest possible responses that will allow us to keep pace with a perpetually shifting ecosystem.

C'est un truisme universellement accepté que nous vivons une époque de changements sans précédent, tant en termes d'ampleur que de rythme de ces changements. Dans le monde des statistiques, les outils, les méthodologies, les techniques et la transparence ont connu d'énormes changements. De nouveaux logiciels, de nouvelles méthodes d'analyse des données et de nouvelles données ouvertes arrivent chaque jour, chacun ayant des implications sur la manière dont nous abordons notre développement professionnel. Dans cet exposé, je présenterai des observations tirées de la pratique de la statistique appliquée et de l'enseignement à des professionnels en milieu de carrière qui suggèrent des réponses possibles qui nous permettront de suivre le rythme d'un écosystème en perpétuelle évolution.

Chair/Président: Kathryn Morrison

Date: Friday June 3 / vendredi 3 juin

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-15:45]

Jeffrey W. Peitsch (University of Winnipeg)

Classification-Based Inference for Spatially Stratified Infectious Disease Systems

Inférence basée sur la classification pour systèmes de maladies infectieuses spatialement stratifiées

Infectious disease transmission dynamics are a function of complex interactions between susceptible and infectious individuals. When modelling such transmission dynamics, we must account for complex heterogeneities within the population, making the model fitting process computationally inefficient in traditional methods such as Bayesian Markov chain Monte Carlo framework. To address these challenges, we propose to use a supervised learning-based approach where the epidemic generating models are classified using epidemic summary statistics as predictors. We consider a spatially heterogenous population and implement natural stratification to better predict the epidemic generating models and compare the performance of ensemble learning techniques such as random forest and XG boost. The methods are applied to simulated data, and the 2001 UK foot and mouth disease epidemic data. The results show that natural stratification can help to predict the model more accurately than global data.

La dynamique de transmission des maladies infectieuses est fonction d'interactions complexes entre individus sensibles et infectieux. Lors de la modélisation d'une telle dynamique de transmission, nous devons tenir compte d'hétérogénéités complexes au sein de la population, ce qui rend le processus d'ajustement du modèle inefficace sur le plan informatique dans les méthodes traditionnelles telles que le cadre bayésien de Monte Carlo à chaîne de Markov. Pour relever ces défis, nous proposons d'utiliser une approche basée sur l'apprentissage supervisé où les modèles générateurs d'épidémies sont classés en utilisant les statistiques sommaires des épidémies comme prédicteurs. Nous considérons une population spatialement hétérogène et employons une stratification naturelle pour mieux prédire les modèles générateurs d'épidémies, puis nous comparons les performances des techniques d'apprentissage d'ensemble telles que la forêt aléatoire et XG boost. Nous appliquons ces méthodes à des données simulées et aux données de l'épidémie de fièvre aphteuse de 2001 au Royaume-Uni. Les résultats montrent que la stratification naturelle aide à prédire le modèle avec plus de précision que les données globales.

[15:45-16:00]

Rick E Danielson (Fisheries and Oceans Canada) **Hui Shen** (Pêches et Océans Canada) **Jing Tao** (Pêches et Océans Canada) **Will Perrie** (Pêches et Océans Canada)

Towards a Characterization of North Atlantic Right Whale Habitat from Space: Dependence of Ocean Current Features on Wind

Vers une caractérisation de l'habitat des baleines noires de l'Atlantique Nord depuis l'espace : dépendance des caractéristiques des courants océaniques par rapport au vent

Measures of dependence - distance correlation and a proposed linear and nonlinear decomposition of Pearson correlation - are examined to identify a broad peak in the relationship between ocean watermass boundaries, as contrasts seen from space, and surface wind speed, where a variable wind speed exponent is employed to maximize these measures. Locations of recent North

Nous examinons des mesures de dépendance - la corrélation de distance et une proposition de décomposition linéaire et non linéaire de la corrélation de Pearson - pour identifier un large pic dans la relation entre les limites de la masse d'eau océanique, en tant que contrastes vus de l'espace, et la vitesse du vent de surface, où on emploie un exposant variable de la vitesse du vent pour maximiser ces mesures. Nous échantillonnons les emplacements

Atlantic right whale (Eubalaena glacialis) sightings in the Gulf of St. Lawrence are sampled between 2008 and 2020 by 324 Radarsat-2 scenes, with marine wind speed taken from a global reanalysis. The relationship between Radarsat-2 contrast (a proxy of ocean current convergence) and wind speed is quantified, and correlation following Radarsat-2 contrast adjustment is obtained. In addition to facilitating a search for hotspots of biological activity in the water column, we take the opportunity to compare distance correlation and Pearson components with their proposed interpretation.

des observations récentes de baleines noires de l'Atlantique Nord (Eubalaena glacialis) dans le golfe du Saint-Laurent entre 2008 et 2020 par 324 scènes Radarsat-2, la vitesse du vent marin étant tirée d'une réanalyse mondiale. Nous quantifions la relation entre le contraste Radarsat-2 (une approximation de la convergence des courants océaniques) et la vitesse du vent, et obtenons une corrélation après ajustement du contraste Radarsat-2. En plus de faciliter la recherche de points chauds d'activité biologique dans la colonne d'eau, nous en profitons pour comparer la corrélation de distance et les composantes de Pearson avec leur interprétation proposée.

[16:00-16:15]

Madeline Ward (University of Calgary) **Lorna E. Deeth** (University of Guelph) **Rob Deardon** (University of Calgary)

Incorporating Behavioural Change into Spatial Individual-Level Models for Infectious Disease Transmission

Incorporer le changement de comportement dans les modèles spatiaux de transmission de maladies infectieuses au niveau individuel

Individual-level models can flexibly incorporate information on individual risk factors, including spatial location. This can account for the high degree of heterogeneity that is characteristic of population mixing, and, thus, infection transmission. However, these models have typically assumed stable population behaviour over time. As we have observed throughout the COVID-19 pandemic, behaviour often changes based on the current perceived risk of contracting the disease. In turn, this behaviour change can have a large impact on the transmission dynamics of the disease. We will present a new class of behavioural-change individual-level models where various functions of infection prevalence affect susceptibility and/or population mixing and illustrate their use through simulated and real data on foot and mouth disease. We will demonstrate the use of spike and slab priors to determine whether behaviour change is present in an epidemic, along with the effects of model misspecification.

Les modèles au niveau individuel peuvent incorporer de manière flexible des informations sur les facteurs de risque individuels, y compris la localisation spatiale. Cela permet tenir compte du taux considérable d'hétérogénéité qui caractérise le mixage de la population, et par conséquent la transmission des infections. Cependant, on suppose en général avec ces modèles que le comportement de la population reste stable dans le temps. Comme nous avons pu le constater avec la pandémie de COVID-19, le comportement change souvent en fonction des risques perçus. Réciproquement, ce changement de comportement peut influencer énormément la dynamique de la transmission de la maladie. Nous présenterons une nouvelle classe de modèles de changement de comportement au niveau individuel dans laquelle des fonctions de la prévalence affectent la susceptibilité et/ou le mixage de la population et nous illustrons leur utilisation à l'aide d'une étude de simulation et d'un ensemble de données provenant d'une épidémie de fièvre aphteuse. Nous démonterons l'usage des a priori de type spike-and-slab pour déterminer si le changement de comportement est présent dans une épidémie, et aussi les effets d'une erreur de spécification du modèle.

[16:15-16:30]

Selvakkadunko Selvaratnam (University of Toronto)

Applications of Robust Methods in Modern Spatial Analysis

Applications de méthodes robustes en analyses spatiales modernes

Spatial data analysis provides a valuable information to government as well as companies. A rapid improvement of modern technology with a geographic information system (GIS) can lead to collect and store more spatial data. We develop algorithms to choose optimal locations from locations that are permanently in a space

L'analyse de données spatiales procure de précieux renseignements aux gouvernements tout comme aux entreprises. Une grande avancée de la technologie moderne avec un système d'information géographique (SIG) peut contribuer à obtenir et entreposer davantage de données spatiales. Nous élaborons des algorithmes pour choisir l'emplacement idéal parmi des emplacements

for an efficient spatial data analysis. Distances between neighbouring permanent locations are not necessary to be equispaced distances. Robust and sequential methods are used to develop algorithms for design constructions. The constructed designs are robust against misspecified regression responses and variance/covariance structures of responses.

en permanence dans un espace afin de favoriser l'efficacité de l'analyse de données spatiales. Les distances entre les emplacements permanents voisins ne sont pas nécessaires pour être des distances en phase. Les méthodes séquentielles et robustes sont employées pour le développement d'algorithmes de conception de plan. Les plans conçus sont robustes contre les réponses de régression mal spécifiées et les structures variance-covariance de réponses.

[16:30-16:45]

Kyran Cupido (St Francis Xavier University) **Petar Jevtic** (Arizona State University) **Tim Boonen** (University of Amsterdam)

Space, Mortality, and Economic Growth

Espace, mortalité et croissance économique

At present, academic actuarial research involving the mortality modeling of multiple populations mainly focuses on factor-based approaches. This comes with little attention to interpretable models of mortality that take patterns across space into consideration. To address this, we propose a family of models that extend the seminal factor-based stochastic mortality modeling framework of Li and Lee (2005) to include spatial patterns. Specifically, in this paper, we study the relationship between economic growth, as represented by the real gross domestic product (GDP), and mortality of the contiguous United States. The proposed spatial lag of GDP with GDP (SLGG) model was used to produce forecasts of mortality rates and annuity pricing for each of the states of the United States and demonstrated the effects which economic growth has on mortality. A comparison of annuity pricing across space revealed that the SLGG model preserves more regional differences when it comes to pricing compared to the Li and Lee (2005) model. In a larger context, this research provides a blueprint for the inclusion of spatial components and economic growth into mortality modeling.

À l'heure actuelle, la recherche actuarielle académique impliquant la modélisation de la mortalité de populations multiples se concentre principalement sur des approches basées sur des facteurs. On n'accorde que peu d'attention aux modèles de mortalité interprétables qui prennent en compte les schémas dans l'espace. Pour remédier à cela, nous proposons une famille de modèles qui étendent le cadre séminal de modélisation stochastique de la mortalité basé sur des facteurs de Li et Lee (2005) pour y inclure des modèles spatiaux. Plus précisément, dans cet article, nous étudions la relation entre la croissance économique, représentée par le produit intérieur brut (PIB) réel, et la mortalité aux États-Unis contigus. Le modèle proposé de décalage spatial du PIB avec le PIB (SLGG) a été utilisé pour produire des prévisions des taux de mortalité et de la tarification des rentes pour chacun des États des États-Unis et a démontré les effets de la croissance économique sur la mortalité. Une comparaison de la tarification des rentes dans l'espace a révélé que le modèle SLGG préserve davantage les différences régionales en matière de tarification par rapport au modèle de Li et Lee (2005). Dans un contexte plus large, cette recherche fournit un plan pour inclure des composantes spatiales et la croissance économique dans la modélisation de la mortalité.

[16:45-17:00]

Sara Zapata-Marin (McGill University) **Alexandra M. Schmidt** (McGill University) **Scott Weichenthal** (McGill University) **Eric Lavigne** (Health Canada)

Modelling Temporally Misaligned Data Across Space

Modélisation de données temporellement désalignées dans l'espace

Due to the high costs of monitoring environmental processes, these studies commonly involve temporally misaligned data. Temporal misalignment refers to measurements taken at different temporal scales across different spatial locations, e.g., some sites provide weekly measurements while others provide daily ones. We

En raison des coûts élevés liés à la surveillance des processus environnementaux, des données temporellement désalignées sont couramment présentes dans les études. Le désalignement temporel fait référence à des mesures prises à diverses échelles temporelles dans différents emplacements spatiaux, par exemple, certains sites fournissent des mesures hebdomadaires et d'autres des mesures quo-

propose a spatiotemporal model that accounts for this temporal misalignment. Inference is performed under the Bayesian framework, and uncertainty about unknown quantities is naturally accounted for. The motivating example consists of temporally misaligned measurements of total pollen concentration across Toronto, Canada. The proposed model provides estimates of the daily measurements at sites where only weekly data was available. We also show how temporal aggregation impacts the associations with the available covariates.

tidiennes. Nous proposons un modèle spatio-temporel qui prend en compte ce désalignement temporel. L'inférence est faite dans un cadre bayésien et l'incertitude relative aux quantités inconnues est prise en compte naturellement. Nous illustrons notre modèle avec des mesures temporellement désalignées de la concentration totale de pollen à Toronto, au Canada. Le modèle proposé fournit une estimation de mesures quotidiennes dans des sites où seules des données hebdomadaires étaient disponibles. Nous montrons aussi l'impact de l'agrégation temporelle sur les associations avec les covariables disponibles.

Chair/Président: Meng Yuan

Date: Friday June 3 / vendredi 3 juin

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-15:45]

Yanglei Song (Queen's University) **Meng Zhou** (Queen's University)

Truncated LinUCB for Stochastic Linear Bandits

Algorithme LinUCB tronqué pour des bandits linéaires stochastiques

This paper considers contextual bandits with a finite number of arms, where the contexts are independent and identically distributed d-dimensional random vectors, and the expected rewards are linear in both the arm parameters and contexts. The LinUCB algorithm, which is nearly minimax and asymptotically optimal for related linear bandits, is shown to have a cumulative regret that is suboptimal in both the dimension d and the time horizon T, due to its over-exploration. A truncated version of LinUCB is proposed and termed "Tr-LinUCB", which follows LinUCB up to a truncation time S and performs pure exploitation afterwards. The Tr-LinUCB algorithm is shown to achieve $O(d\log(T))$ regret if $S = Cd\log(T)$ for a sufficiently large constant C, and a matching lower bound is established, which shows the rate optimality of Tr-LinUCB in a low dimension setup. Further, if $S = d\log^{\kappa}(T)$ for some $\kappa > 1$, the loss compared to the optimal is a multiplicative $\log\log(T)$ factor, which does not depend on d. This insensitivity to overshooting in choosing the truncation time of Tr-LinUCB is of practical importance.

Cet article présente des bandits contextuels avec un nombre de bras fini lorsque les contextes sont des vecteurs d-dimensionnels indépendants et identiquement distribués et les récompenses attendues sont linéaires à la fois dans les paramètres des bras et les contextes. En raison de son exploration excessive, l'algorithme LinUCB, qui est presque minimax et asymptotiquement optimal pour les bandits linéaires liés, a montré un regret cumulatif sous-optimal à la fois dans la dimension d et l'horizon temporel T. Une version tronquée du LinUCB proposée sous le nom « Tr-LinUCB » suit LinUCB jusqu'au temps de troncature S et exécute une exploitation pure par après. On a montré que l'algorithme Tr-LinUCB peut atteindre un regret $O(d\log(T))$ si $S = Cd\log(T)$ pour une constante C suffisamment grande et une borne inférieure correspondante est établie, ce qui montre l'optimalité du taux de Tr-LinUCB dans une configuration à faible dimension. De plus, lorsque $S = d\log^{\kappa}(T)$ pour certains $\kappa > 1$, la perte comparée à l'optimum est un facteur $\log\log(T)$ multiplicateur qui ne dépend pas de d. Cette insensibilité au surajustement dans le choix du temps de troncature du Tr-LinUCB est importante sur le plan pratique.

[15:45-16:00]

Archer Gong Zhang (University of British Columbia) **Jiahua Chen** (University of British Columbia)

Estimation Efficiency under a Two-Sample Density Ratio Model

Efficacité de l'estimation sous un modèle de rapport de densité à deux échantillons

In many applications, we collect independent samples from interconnected populations. These population distributions share some latent structure, so it is advantageous to jointly analyze the multiple samples. Recently, many researchers have advocated the use of the semiparametric density ratio model (DRM) to account for the latent structure the multiple populations share and have developed more efficient data analysis procedures based on pooled data. In this talk, we investigate

Dans bon nombre d'applications, on procède à la collecte d'échantillons indépendants de populations interconnectées. Comme ces distributions de populations partagent une certaine structure latente, une analyse conjointe de multiples échantillons est avantageuse. Des chercheurs ont récemment préconisé l'utilisation d'un modèle de rapport de densité (DRM) semi-paramétrique pour prendre en compte la structure latente partagée par les multiples populations et ils ont développé des procédures plus efficaces d'analyse de données basées sur des données totalisées. Notre exposé

the efficiency of some estimators under a two-sample DRM. We consider the scenario where we have two samples whose sizes grow to infinity at different rates, and study the DRM-based inferences for the population corresponding to the smaller-sized sample. We theoretically prove that some DRM-based estimators achieve the same asymptotic efficiency as the parametric estimates derived under a specific parametric model. Our simulation studies on quantile estimation help to confirm our theoretical results.

porte sur l'efficacité de certains estimateurs sous un DRM à deux échantillons. Dans notre scénario, nous prenons deux échantillons dont la taille s'accroît à l'infini à différentes cadences et nous étudions des inférences basées sur le DRM pour la population correspondant à l'échantillon de plus petite taille. Nous apportons la preuve théorique que certains estimateurs basés sur le DRM ont la même efficacité asymptotique que les estimations paramétriques dérivées sous un modèle paramétrique spécifique. Nos études en simulation sur une estimation quantile contribuent à confirmer nos résultats théoriques.

[16:00-16:15]

Jervis Gallanosa (University of Manitoba) **Yuliya V. Martsynyuk** (University of Manitoba)

Nonparametric Asymptotic Tests for Change in the Mean with Better Balanced Power Functions

Tests asymptotiques non paramétriques de changements de la moyenne avec fonctions de puissance équilibrée supérieure

Let n independent chronologically ordered real-valued observables either form a random sample with a finite positive variance, or be such that the first k observables have a common mean that is different from the common mean of the rest $n-k$ observables, $1 \leq k < n$, where the time k of the change in the mean is usually unknown. We conduct a simulation study of finite-sample power functions of known nonparametric tests for detecting such a change in the mean that are based on convergence in distribution of sup- and integral-functionals of certain weighted tied-down partial sums processes. Accordingly, none of these tests are seen to be uniformly more powerful than the rest in hand, as their type I errors and powers depend on heavy-tailedness of the underlying sample and on if the change in the mean occurs on the tails or in the middle of the sample. To obtain tests with better balanced power functions, we consider and study numerically combinations of the sup- and integral-functional tests.

Soit n observables à valeur réelle chronologiquement ordonnés indépendants tirés d'un échantillon aléatoire avec une variance positive finie, telle que les k premiers observables ont une moyenne commune différente de la moyenne commune des $n-k$ autres observables, $1 \leq k < n$, où le temps k de changement de la moyenne est généralement inconnu. Nous menons une étude de simulation sur les fonctions de puissance en échantillon fini de tests non paramétriques connus pour détecter un tel changement de la moyenne qui sont basés sur la convergence dans la distribution des sup-fonctionnelles et intégrales de certains processus de sommes partielles liées pondérées. En conséquence, aucun de ces tests n'est uniformément plus puissant que les autres, vu que leurs erreurs de type I et leur puissance dépendent de l'épaisseur de la queue de l'échantillon sous-jacent et de la possibilité de changement de la moyenne pouvant se produire dans les queues ou dans le centre de l'échantillon. Afin d'obtenir des tests dotés de fonctions de puissance équilibrée supérieures, nous examinons et étudions numériquement des combinaisons de tests sup fonctionnels et intégrals.

[16:15-16:30]

Marc Angelo Parsons (McGill University) **Jingjun (Victor) Chen** (McGill University) **Andrea Benedetti** (McGill University)

Modelling Non-linear Exposure-outcome Relationships in Quantitative Systematic Reviews: A Meta-epidemiological Review of Current Practice

Modélisation des liens exposition-effet non-linéaires dans les revues systématiques quantitatives : une revue méta-épidémiologique de la pratique courante

Traditional statistical methods used in quantitative meta-analyses often assume linear exposure-outcome relationships. However, in many clinical contexts this type of relationship may not hold. In order to better model these relationships in the presence of non-linearity, one may employ statistical smoothing meth-

Les méthodes méta-analytiques utilisées pour l'analyse des données provenant des revues systématiques présupposent souvent un lien exposition-effet linéaire. Cependant, ce n'est pas valable dans tous les contextes cliniques. Les méthodes de lissage peuvent être utilisées afin de mieux modéliser les liens non-linéaires. Des exemples de celles-ci incluent les splines et la régression locale. Il

ods. Examples of these include spline models and local regression. The extent to which these methods are appropriately employed in meta-analyses is not clear. The goal of this meta-review is to summarise the use of these methods in the systematic review literature. To this end, a search of medical literature databases was conducted to identify systematic reviews which employed statistical smoothing methods. Publication characteristics, types of smoothing methods used, and reporting of the methods were extracted from included studies. It is anticipated that this review will identify gaps in the current use of such methods and provide suggestions for their future reporting.

n'est pas clair dans quelle mesure ces méthodes sont utilisées de façon appropriée dans les études méta-analytiques systématiques. Cette revue a comme but de décrire l'utilisation de ces méthodes dans la littérature courante. Une recherche bibliographique a été effectuée pour identifier les revues systématiques qui emploient ces méthodes de lissage. Des données sur les caractéristiques de publication, type de méthodes utilisées et la façon dont ces méthodes ont été communiquées ont été extraites des revues incluses. On s'attend à ce que cette revue identifie les lacunes dans l'utilisation courante des méthodes de lissage et mette de l'avant des suggestions quant à la façon de les communiquer.

[16:30-16:45]

Xiaoting Li (The University of British Columbia) **Harry Joe** (University of British Columbia)

Nonparametric Estimation of Multivariate Tail Probabilities

Estimation non paramétrique des probabilités de queue multivariées

For d¿=2 risk variables, we propose two methods to estimate the joint multivariate tail probabilities and extreme quantile curves. The methods are based on weak assumptions on the joint tails of the copulas of the d variables. The first method is developed based on the tail expansion of copula along different directions to the joint upper or lower corner. The second method is based on the asymptotic expansion of a family of tail-weighted functions defined from the copula. The methods can distinguish the tail properties of the copula, such as reflection asymmetry, permutation asymmetry, and heterogeneous tail dependence. Extensive simulation studies are conducted to compare the estimation methods. Data examples are presented to illustrate the applicability of the proposed methods as inference and diagnostic tools.

Pour d¿=2 variables de risque, nous proposons deux méthodes pour estimer les probabilités de queue multivariées conjointes et les courbes de quantiles extrêmes. Les méthodes sont basées sur des hypothèses faibles sur les queues conjointes des copules des d variables. La première méthode est développée en allongeant la queue de la copule dans différentes directions vers le coin supérieur ou inférieur conjoint. La deuxième méthode est basée sur l'expansion asymptotique d'une famille de fonctions pondérées par la queue définies à partir de la copule. Ces méthodes permettent de distinguer les propriétés de queue de la copule, telles que l'asymétrie de réflexion, l'asymétrie de permutation et la dépendance de queue hétérogène. Nous menons des études de simulation approfondies pour comparer les méthodes d'estimation. Nous présentons ensuite des exemples de données pour illustrer l'applicabilité des méthodes proposées comme outils d'inférence et de diagnostic.

[16:45-17:00]

Deli Li (Lakehead University) **Yu Miao** (Henan Normal University, China) **George Stoica** (University of New Brunswick, Canada)

A General Large Deviation Result for Partial Sums of Super-Heavy Tailed Random Variables

Résultat général de grand écart pour les sommes partielles de variables aléatoires à queue super lourde

In this work, a general large deviation result for partial sums of independent and identically distributed random variables with super-heavy tailed distribution is established. Our main result extends in particular the results of Stoica [*Large gains in the St. Petersburg game.* C. R. Math. Acad. Sci. Paris **346, no. 9-10**, 563-566 (2008)] and Nakata [*Large deviations for super-heavy tailed random walks.* Stat. Prob. Lett.**180**, Arti-

Dans ce travail, nous établissons un résultat général de grand écart pour les sommes partielles de variables aléatoires indépendantes et identiquement distribuées avec une distribution à queue superlourde. Notre résultat principal étend en particulier les résultats de Stoica [*Large gains in the St. Petersburg game.* C. R. Math. Acad. Sci. Paris **346, no. 9-10**, 563-566 (2008)] et Nakata [*Large deviations for super-heavy tailed random walks.* Stat. Prob. Lett.**180**, Article 109240 (2022)]. La technique de symétrisation, une des

cle 109240 (2022)].The symmetrization technique, one of Lévy's inequalities, and two preliminary results on slowly and regularly varying functions are paramount in the proof of our main result. This work has been published in Statistics and Probability Letters **184**, Article 109371 (2022).

inégalités de Lévy, et deux résultats préliminaires sur les fonctions à variation lente et régulière sont primordiaux pour la preuve de notre résultat principal. Ce travail a été publié dans Statistics and Probability Letters **184**, Article 109371 (2022).

New Sampling Techniques and High-dimensional Data Analysis
Nouvelles techniques d'échantillonnage et analyse des données à haute dimension

Chair/Président: Yixiu Liu

Date: Friday June 3 / vendredi 3 juin

Time/Heure: 15:30-17:00

Abstract/Résumé

[15:30-15:45]

Johanna de Haan-Ward (University of Western Ontario) **Simon Bonner** (University of Western Ontario) **Douglas G. Woolford** (University of Western Ontario)

Comparison of Subsampling Methods for Prediction of Rare Events, with Application to Human-Caused Wildland Fire Prediction

Comparaison de méthodes de sous-échantillonnage pour la prédiction d'événements rares, avec application à la prédiction des incendies de forêt d'origine humaine

Datasets used for prediction of rare events are often large and severely imbalanced in the outcome. This presents challenges for predictive modelling. Subsampling of the data is often employed to create data that is balanced in the response. This study compares three methods for using subsampled data in a logistic regression model with smooth functions of covariates: case-control sampling with a deterministic offset, local case-control sampling, and case-control sampling with sampling weights. We illustrate these methods using wildland fire data, comparing models using metrics suitable for rare events, such as area under the precision-recall curve. Local case-control sampling favours observations that reflect conditions in which fires are more probable and shows improved performance over the other methods. This suggests that sampling methods which use domain knowledge may be advantageous in the prediction of rare events, especially for large and complex spatio-temporal data structures.

Les ensembles de données utilisés pour la prédiction d'événements rares sont souvent volumineux et fortement déséquilibrés au niveau des résultats. Cela présente des défis pour la modélisation prédictive. On a souvent recours au sous-échantillonnage des données pour créer des données équilibrées dans la réponse. Cette étude compare trois méthodes de sous-échantillonnage dans un modèle de régression logistique avec fonctions lisses de covariables : l'échantillonnage de cas-témoins avec un décalage déterministe, l'échantillonnage local de cas-témoins et l'échantillonnage de cas-témoins avec des poids d'échantillonnage. Nous illustrons ces méthodes à l'aide de données sur les incendies de forêt, en comparant les modèles à l'aide de métriques adaptées aux événements rares, telles que l'aire sous la courbe de précision-rappel. L'échantillonnage local de cas-témoins favorise les observations qui reflètent les conditions dans lesquelles les incendies sont plus probables et présente de meilleures performances que les autres méthodes. Cela suggère que les méthodes d'échantillonnage qui utilisent la connaissance du domaine peuvent être avantageuses dans la prédiction d'événements rares, en particulier pour des structures de données spatio-temporelles vastes et complexes.

[15:45-16:00]

Lorenzo Frattarolo (European Commission Joint Research Centre) **Roberto Casarin** (University Ca' Foscari of Venice) **Radu V. Craiu** (University of Toronto) **Christian P. Robert** (CEREMADE, University Paris-Dauphine PSL and University of Warwick)

Living on the Edge: An Unified Approach to Antithetic Sampling

Vivre à la limite : une approche unifiée de l'échantillonnage antithétique

We identify recurrent ingredients in the antithetic sampling literature leading to a unified sampling framework. We introduce a new class of antithetic schemes that includes the most used antithetic proposals. This perspective enables the derivation of new properties of

Nous identifions les composants récurrents de l'échantillonnage antithétique dans la littérature, ce qui nous conduit à la définition d'un cadre d'échantillonnage unifié. Nous présentons une nouvelle classe de schémas antithétiques qui comprend les propositions antithétiques les plus utilisées. Cette approche permet de d'obtenir

the sampling schemes: i) optimality in the Kullback-Leibler sense; ii) closed-form multivariate Kendall's τ and Spearman's ρ; iii) ranking in concordance order and iv) a central limit theorem that characterizes stochastic behavior of Monte Carlo estimators when the sample size tends to infinity. The proposed simulation framework inherits the simplicity of the standard antithetic sampling method, requiring the definition of a set of reference points in the sampling space and the generation of uniform numbers on the segments joining the points. We provide applications to Monte Carlo integration and Markov Chain Monte Carlo Bayesian estimation.

de nouvelles propriétés des schémas d'échantillonnage : i) l'optimalité au sens de Kullback-Leibler ; ii) les formes fermées des valeurs multivariées du tau de Kendall et du rho de Spearman ; iii) le classement par ordre de concordance et iv) un théorème central limite qui caractérise le comportement stochastique des estimateurs de Monte-Carlo lorsque la taille de l'échantillon tend vers l'infini. Le cadre de simulation proposé hérite de la simplicité de la méthode d'échantillonnage antithétique standard, pour laquelle il suffit de définir un ensemble de points de référence dans l'espace d'échantillonnage et de générer des nombres uniformément sur les segments reliant les points. Nous fournissons des applications en intégration de Monte-Carlo et en estimation bayésienne de Monte-Carlo par chaînes de Markov.

[16:00-16:15]

Xiaotong Liu **Zihang Lu** (Queen's University) **Myrtha Reyna** (The Hospital for Sick Children)

Defining Lifestyle Patterns Using High Dimensional Questionnaire Data

Définition des modes de vie à l'aide de données de questionnaire à haute dimension

Multi-dimensional mixed-type data are prevalent in a large cohort study. Our study is motivated by a Canadian birth cohort study. We aim to define lifestyle patterns using multi-dimensional data from questionnaires and associate them with common diseases such as obesity and asthma in children. In an attempt to describe lifestyle patterns using longitudinal continuous and categorical variables, we adopt and compare several dimension reduction methods, such as principal component analysis (PCA), multiple correspondence analysis (MCA), and PCA of a mixture of numerical and categorical data (PCAmix). Our analysis results from a real dataset will be discussed and presented.

Les données multidimensionnelles de type mixte sont courantes dans les grandes études de cohortes. Notre étude est motivée par une étude de cohorte de naissances canadienne. Nous visons à définir les modes de vie à l'aide de données multidimensionnelles issues de questionnaires et à les associer à des maladies courantes telles que l'obésité et l'asthme chez les enfants. Pour tenter de décrire les habitudes de vie à l'aide de variables longitudinales continues et catégorielles, nous adoptons et comparons plusieurs méthodes de réduction des dimensions, telles que l'analyse en composantes principales (ACP), l'analyse des correspondances multiples (ACM) et l'ACP d'un mélange de données numériques et catégorielles (ACPmix). Nous discuterons et présenterons les résultats de nos analyses à partir d'un jeu de données réel.

[16:15-16:30]

Derek Latremouille (University of Toronto) **Dehan Kong** (University of Toronto) **Linglong Kong** (University of Alberta)

High-Dimensional, Low-Sample Tests of Normality Based on Concentration

Tests de normalité en haute dimension avec petite taille d'échantillon basés sur la concentration

Methods routinely used to analyze high-dimension, low-sample-size (HDLSS) data are often based on the assumption of multivariate Normality in conjunction with restrictions on the degree of dependence amongst variables. However, while results can be quite sensitive to departures from these models, there is a dearth of Normality tests valid in high dimensions. To address this, we exploit concentration-type effects to develop a class of Normality tests valid in HDLSS settings. Corroborating asymptotic validity of the tests, simulation analysis indicates that Type I error is well-controlled in both HDLSS settings and contexts in which sample

Les méthodes couramment utilisées pour analyser des données de haute dimension avec petite taille d'échantillon (HDLSS) sont souvent basées sur l'hypothèse de normalité multivariée en conjonction avec des restrictions sur le degré de dépendance entre les variables. Cependant, alors que les résultats peuvent être très sensibles aux écarts par rapport à ces modèles, il existe une pénurie de tests de normalité valables en haute dimension. Pour remédier à cela, nous exploitons les effets de type concentration pour développer une classe de tests de normalité valables dans des contextes HDLSS. Corroborant la validité asymptotique des tests, l'analyse de simulation indique que l'erreur de type 1 est bien contrôlée à la fois dans les contextes HDLSS et dans

size is proportional to the number of variables. Relevant alternatives for which our tests achieve high power are identified via asymptotic properties as well as simulation analysis. Our methodology is then applied to gene expression and brain imaging data frequently analyzed using procedures based on the multivariate Normal model.

les contextes où la taille de l'échantillon est proportionnelle au nombre de variables. Nous identifions les alternatives pertinentes pour lesquelles nos tests atteignent une puissance élevée via les propriétés asymptotiques ainsi qu'une analyse de simulation. Nous appliquons ensuite notre méthodologie à des données d'expression génétique et d'imagerie cérébrale fréquemment analysées à l'aide de procédures basées sur le modèle normal multivarié.

[16:30-16:45]

Richard Le Blanc (CHUS)

Noncentral Distributions' Bayesian Inference in terms of Orthogonal Polynomials

Inférence bayésienne concernant les distributions noncentrales en termes de polynômes orthogonaux

The noncentral distributions can be derived by multiplying their central distributions by translation factors T. If they are constructed in terms of a translated hypersphere, the T factors become generating functions for families of orthogonal polynomials, with the central distributions standing for the family-defining weights. The polynomials are Gegenbauer, Hermite, Jacobi and Laguerre for the noncentral t, normal, F and chi square distributions, respectively. Gibbs prior can be constructed in terms of these noncentral distributions and the entropic convex duals of the empirical densities to be modeled. Expanding the duals on the orthogonal polynomials allow for expedient computation of priors using very few polynomial coefficients. Genomic scale empirical distributions can be modeled using less than 10 coefficients. The powerful formalism even allows for computation of Bayes Factors without having to compute priors or posteriors. Genomics and geophysics example will be provided.

Les distributions non centrales peuvent être construites en multipliant les distributions centrales par des facteurs de translation T. Lorsque construits en termes de translation d'hypersphère, les facteurs T deviennent pour les distributions non centrales t, normale, F et chi carré des fonctions génératrices pour les polynômes orthogonaux de Gegenbauer, Hermite, Jacobi et Laguerre respectivement, avec les distributions centrales représentant les fonctions de poids des familles correspondantes. Les a priori de Gibbs peuvent être construits à l'aide de ces distributions et des fonctions duales convexes entropiques des densités à modéliser. L'expansion des fonctions duales sur les polynômes orthogonaux permet une détermination aisée des a priori nécessitant qu'un petit nombre de coefficients. Des distributions d'échelle génomique pouvant être modélisées avec moins de 10 coefficients. Ce puissant formalisme permet calcul des facteurs de Bayes sans avoir à calculer les a priori ou a posteriori. Des exemples en génomique et géophysique seront donnés.

[16:45-17:00]

Jiarui Zhang (Simon Fraser University)

An Annealed Sequential Monte Carlo Method for Generalized Bayesian Multidimensional Scaling

Méthode de Monte-Carlo séquentielle recuite pour l'échelonnement multidimensionnel bayésien généralisé

Multidimensional scaling is widely used to reconstruct a map with the points' coordinates in a low-dimensional space from the original high-dimensional space while preserving the pairwise distances. Within a Bayesian framework, the available approach with Markov chain Monte Carlo algorithms has some possible limitations in model generalization and performance comparison. To overcome these limitations, we first developed a general framework that incorporates non-Gaussian measurement errors and robustness to fit different observed dissimilarities. Then, we proposed an annealed Sequential Monte Carlo algorithm for Bayesian multidimen-

L'échelonnement multidimensionnel est largement utilisé pour reconstruire une carte avec les coordonnées des points dans un espace à faible dimension à partir de l'espace original à haute dimension tout en préservant les distances par paire. Dans un cadre bayésien, l'approche disponible avec les algorithmes de Monte Carlo à chaîne de Markov peut présenter certaines limites dans la généralisation des modèles et la comparaison des performances. Pour surmonter ces limites, nous avons d'abord développé un cadre général qui intègre les erreurs de mesure non gaussiennes et la robustesse pour s'adapter aux différentes dissimilarités observées. Ensuite, nous avons proposé un algorithme de Monte Carlo séquentiel recuit pour l'inférence bayésienne

sional scaling inference. This algorithm provides an approximate posterior distribution over the points' coordinates in a low-dimensional space and an unbiased estimator for the marginal likelihood. We have found that the proposed algorithm outperforms other benchmark algorithms under the same computational budget according to some commonly used metrics.

d'échelonnement multidimensionnel. Cet algorithme fournit une distribution postérieure approximative sur les coordonnées des points dans un espace à faible dimension et un estimateur sans biais pour la vraisemblance marginale. Nous avons constaté que l'algorithme proposé surpasse les autres algorithmes de référence sous le même budget de calcul pour certaines métriques couramment utilisées.

Chair/Président: Jean-François Beaumont

Organizer/Responsable: Jean-François Beaumont

Date: Saturday June 4 / samedi 4 juin

Time/Heure: 12:30-17:00

Abstract/Résumé

[12:30-17:00]

Changbao Wu (University of Waterloo)

From Sample Surveys to Missing Data and Causal Inference

Des enquêtes par sondage aux données manquantes et à l'inférence causale

We provide an introduction to analysis of complex probability survey samples with a major focus on weighting and calibration methods. We show that the techniques, combined with the empirical likelihood methods, can be used as general inferential tools for missing data analysis and causal inference with observational data. Some recent developments on doubly robust and multiply robust inference and methods for analyzing non-probability survey samples are discussed to illustrate the use of weighting and calibration for these seemingly different but intrinsically connected topics. The workshop is designed for senior undergraduate or graduate students in statistics, biostatistics or related fields and young researchers interested in the topics.

Nous proposons une introduction à l'analyse d'échantillons d'enquêtes probabilistes complexes, en mettant l'accent sur les méthodes de pondération et de calibration. Nous montrons que ces techniques, combinées aux méthodes de vraisemblance empirique, peuvent être utilisées comme outils inférentiels généraux pour l'analyse des données manquantes et l'inférence causale avec des données d'observation. Nous discutons de certains développements récents en matière d'inférence doublement robuste et multiplement robuste et de méthodes d'analyse des échantillons d'enquête non probabilistes pour illustrer l'utilisation de la pondération et de la calibration dans ces domaines apparemment différents mais intrinsèquement liés. L'atelier est conçu pour les étudiants de premier cycle ou diplômés en statistique, biostatistique ou domaines connexes et les jeunes chercheurs intéressés par ces sujets.

Chair/Président: Nathaniel Tyler Stevens

Organizer/Responsable: Nathaniel Tyler Stevens

Date: Saturday June 4 / samedi 4 juin

Time/Heure: 13:00-16:30

Abstract/Résumé

[13:00-16:30]

Rodolfo Lourenzutti (University of British Columbia) **Arman Seyed-Ahmadi** (University of British Columbia) **Diego Ardila** (Shopify)

Intro to Databases in Industry: Data Cleaning, Querying, and Modeling at Scale

Introduction aux bases de données en industrie : nettoyage des données, interrogation et modélisation à grande échelle

This workshop is intended to walk the participants through the journey of data from the "raw" to an "analysis-ready" state. Using R, we will explore the basic flow of data cleaning and organizing raw data such that the outcome is error-free, consistent and accurate. The participants will then be introduced to relational databases—the most widely used option for storing clean, well-structured data. We will explore how to interact with and efficiently retrieve data from relational databases using their well-known, powerful query language of SQL. Finally, we will show how to connect R to SQL databases for reading and writing purposes.

Cet atelier a pour but de guider les participants dans le parcours des données, de l'état « brut » à un état « prêt à l'analyse ». À l'aide de R, nous explorerons le flux de base du nettoyage des données et de l'organisation des données brutes de sorte que le résultat soit exempt d'erreurs, cohérent et précis. Les participants seront ensuite initiés aux bases de données relationnelles - option la plus largement utilisée pour stocker des données propres et bien structurées. Nous explorerons comment interagir avec les bases de données relationnelles et en extraire efficacement des données à l'aide du langage d'interrogation puissant et bien connu SQL. Enfin, nous montrerons comment connecter R aux bases de données SQL à des fins de lecture et d'écriture.

Author List • Liste des auteurs

Abdallah, Anas, 41, 218
Abdi, Hervé, 15, 91
Adamic, Peter, 30, 170
Adelzadeh, Masoud, 14, 90
Afriyie, Gabriel Oppong, 21, 127
Aghababaei Jazi, Omidali, 49, 263
Aghahosseinalishirazi, Zahra, 51, 278
Aguirre, Alberto Nettel, 21, 127
Ahmed, S. Ejaz, 20, 119
Ahmed, Suborna Shekhor, 31, 172
Akande, Olanrewaju Michael, 13, 82
Alakus, Cansu, 49, 266
Alexander, Rohan, 48, 258
Andrews, Jeffrey L., 56, 304
Arbour, David, 18, 112
Ardila, Diego, 10, 66, 325
Arsenault-Mahjoubi, Louis, 30, 169
Asgharian, Masoud, 43, 231
Assa, Hirbod, 27, 151
Auger-Mèthè, Marie, 46, 249
Augustyniak, Maciej, 13, 84
Avey, Marc T., 15, 95
Aviña-Zubieta, J. Antonio, 41, 220

Babyn, Jonathan, 50, 273
Badescu, Alexandru, 13, 84
Badescu, Andrei L., 27, 41, 150, 218
Bae, Gunho, 52, 284
Bahremani, Mohsen, 55, 297
Bahroui, Tarik, 17, 105
Bai, Kailun, 43, 230
Bai, Wei, 25, 140
Bak, Stephen, 53, 289
Bansal, Ayuish, 23, 135
Bao, Yanchun, 53, 287
Barajas, Vianey Leos, 48, 259
Basnayake, Shanika, 18, 114
Bassim, Carol, 45, 239
Beaulac, Cédric, 11, 53, 72, 288
Bédard, Mylène, 32, 175
Bégin, Jean-François, 13, 30, 55, 84, 169, 296
Belalia, Mohamed, 48, 259
Béliveau, Audrey, 35, 189, 191
Bellhouse, David R., 50, 272
Benedetti, Andrea, 11, 52, 58, 71, 73, 286, 317
Bennett, David A., 22, 128
Benrimoh, David, 27, 155
Berger, Jeffrey, 22, 131
Berke, Olaf, 49, 267
Bhagwat, Nikhil, 45, 243
Bhagwat, Pankaj Uttam, 22, 131
Bhardwaj, Shivani, 45, 245
Bhatnagar, Sahir R., 18, 21, 112, 123
Bian, Mengjie, 11, 74
Bian, Yuan, 23, 24, 137
Bian, Zeyu, 18, 112
Bicker, Caroline, 42, 226
Bilodeau, Blair, 41, 221
Bingham, Derek, 19, 116
Biziaev, Timofei, 11, 74
Blier-Wong, Christopher, 17, 109
Bolton, Liza, 15, 94
Bonner, Simon, 58, 320
Boonen, Tim, 58, 314
Bornn, Luke C., 30, 166
Bouchard-Côté, Alexandre, 23, 50, 133, 270
Boudreault, Mathieu, 21, 122
Boukili-Makhoukhi, A., 48, 261
Bourget, Mathilde, 26, 149
Bowala, Sulalitha, 54, 295
Boychuk, Den, 28, 159
Briollais, Laurent, 29, 34, 164, 186

Brisebois, François, 57, 309
Brobbey, Anita, 21, 56, 127, 303
Brooks-Wilson, Angela, 45, 241
Brossard, Myriam, 28, 158
Brown, Patrick E., 23, 135
Browne, Ryan P., 14, 52, 56, 89, 282, 304
Bull, Shelley B., 28, 41, 158, 223
Burak, Katherine, 34, 187
Bureau, Alexandre, 35, 193
Buro, Karen, 40, 214
Burr, Wesley, 9, 33, 64, 181
Burstyn, Igor, 53, 291
Bushby, Alexandra S., 44, 238

Cadigan, Noel, 50, 274
Caetano, Samantha-Jo, 48, 258
Cai, Hengrui, 33, 179
Calcetero, Sebastian F, 17, 107
Cameron, Ellen, 55, 300
Campbell, Dave, 33, 182
Campbell, Harlan, 23, 135
Campbell, Trevor, 23, 54, 133, 292
Cannings, Timothy I., 20, 119
Canty, Angelo J., 11, 74
Cao, Jiguo, 11, 25, 34, 45, 46, 72, 142, 185, 244, 245, 249
Carallo, Giulia, 55, 298
Casarin, Roberto, 22, 55, 59, 128, 297, 298, 320
Castellucci, Lana, 22, 131
Caubet Fernandez, Miguel, 42, 227
Ceka, Amarildo, 24
Chan, Ian Weng, 27, 150
Charlin, Laurent, 28, 158
Charvadeh, Yasin Khadem, 23, 137
Chen, Chyong-Mei, 46, 251
Chen, Hsin-Jen, 46, 251
Chen, Jiahua, 52, 53, 58, 284, 290, 316
Chen, Jingjun (Victor), 11, 58, 73, 317
Chen, Meixi, 34, 187
Chen, Panxi, 23
Chen, Rong, 44, 234
Chen, Si, 30, 170
Chen, Sixia, 13, 81
Chen, Tingting, 30, 170
Chen, Tom, 12, 76
Chen, Xia, 44, 235
Chen, Zeyu, 52, 283
Chen, Ziming, 22, 131
Chenouri, Shojaeddin, 40, 215
Chevalier, Fanny, 26, 145
Chewi, Sinho, 54, 296
Chiaromonte, Francesca, 49, 264
Chipman, Hugh, 19, 116
Choi, Wonje, 24
Christensen, Rebecca, 15, 94
Chu, Kenneth C.K., 57, 310
Cigsar, Candemir, 36, 42, 198, 223
Cohen Freue, Gabriela, 21, 33, 182
Colditz, Graham, 25, 142
Collier, Chris, 15, 95
Cook, Richard J., 20, 41, 47, 118, 222, 256
Cossette, Hélène, 17, 109
Costantini, Mauro, 55, 297
Coulombe, Janie, 51, 275
Cowen, Laura L.E., 34, 37, 185, 200
Craiu, Radu V., 48, 59, 259, 320
Cremona, Marzia Angela, 31, 49, 173, 264
Csik, Samantha, 36, 200
Cui, Jingyu, 24, 44, 237
Culbert, Patrick, 31, 172
Cupido, Kyran, 58, 314

Dadié, Éric, 31, 173
Dagdoug, Mehdi, 13, 81
Daly-Grafstein, Daniel, 34, 184
Dang (Subedi), Sanjeena, 51, 278
Dang, Huy, 49, 264
Dang, Sanjeena, 14, 90
Danielson, Rick E, 57, 312
Dasylva, Abel C., 35, 190
Davis, Jack, 31, 172
Davison, Anthony, 11, 70
de Haan-Ward, Johanna, 58, 320
De Jager, Philip L., 22, 128
De Souza, Dr Camila, 51, 278
De Vera, Mary A., 41, 220
Dean, Charmaine B., 16, 25, 28, 102, 144, 159
Dean, Nema, 18, 111
Deardon, Rob, 53, 57, 287, 313
Deck, Wilber, 55, 299
Deeth, Lorna E., 49, 57, 267, 313
Deng, Dianliang, 55, 300
Deng, Gansen, 21, 24, 127
Denis, Marie, 12, 76
Denzil, Rachel, 31
Desmond, Anthony F., 30, 170
Dey, Rajib, 55, 299

Di Bernardino, Elena, 20, 121
Diao, Liqun, 20, 118
Dinh, Sonny, 31
Dioni, Abdoulaye, 35, 193
Diop, Awa, 42, 224
Dockes, Jerome, 45, 243
Doig, Renny, 31, 35, 193
Domaratzki, Mike, 31, 174
Dong, Chun, 32
Dong, Mei, 25, 44, 140, 238
Doroshenko, Lyubov, 38, 207
Dossa, H. Roland G., 48, 259
Dowe, David, 54, 294
Du, Pang, 46, 248
Dubin, Joel A., 18, 38, 46, 110, 210, 246
Duguay, Sébastien, 63
Duguay, Sébastien, 9, 47, 254
Duong, Mylinh, 45, 239

Eghbalzadeh, Ramin, 26, 149
El Adlouni, Salah, 48, 261
El Hannoun, W., 48, 261
Elliott, Lloyd T., 26, 34, 45, 148, 185, 241
Elmasri, Mohamad, 28, 158
Emelko, Monica, 55, 300
Eng, Jeremy, 39, 211
Engelke, Sebastian, 32, 55, 176, 301
Erdogdu, Murat A., 54, 296
Eslami, Aida, 15, 35, 91, 193
Espin-Garcia, Osvaldo, 38, 52, 207, 283
Estep, Donald, 52, 286
Etienne, Marceau, 17, 109

Falk, Carl F., 52, 280
Fallone, Andrew, 32
Fan, Jun, 12, 77
Fan, Xinyao, 48, 260
Fang, Zheng, 54, 294
Farkouh, Michael, 22, 131
Faroughi, Pouya, 17, 109
Fell, Leslie G., 49, 267
Fellows, Ian E., 21, 126
Felsky, Daniel, 22, 128
Feng, Christina, 24
Feng, Cindy Xin, 47, 252
Feng, Jingxue, 23, 137
Feng, Shui, 51, 276
Feng, Zeny, 52, 53, 284, 289
Fick, Gordon H., 38, 208
Field, Chris, 50, 273
Fissler, Tobias, 37, 202
Fortune, Sarah M.E., 46, 249
Forzley, Quinn, 14, 88
Foster, Kathleen Lois, 49, 268
Fournier, Patrick, 45, 241
Francis, Sylvester Ranjith, 31
Frattarolo, Lorenzo, 59, 320
Friesen, Lucas, 40, 217
Frydman, Halina, 42, 225
Fung, Tsz Chai, 27, 41, 150, 218

Gallanosa, Jervis, 58, 317
Gan, Chong, 52, 284
Gao, Dechen, 30, 169
Gao, Xin, 29, 164
Garrett, Rose, 48, 262
Garrett-Mayer, Elizabeth, 16, 100
Genest, Christian, 29, 32, 50, 272
Ghannam, Mai, 34, 186
Ghazzali, Nadia, 16, 102
Ghodsi, Ali, 43, 231
Gibbons, Melanie C. H., 15, 95
Gibbs, Alison L., 33, 181
Gile, Krista J., 21, 126
Gillis, Darren, 35, 191
Gjika, Eralda, 24
Goga, Camelia, 13, 81
Golbeck, Amanda, 16, 102
Golchi, Shirin, 27, 155
Goligher, Ewan, 22, 131
Goussanou, Arthur, 35, 190
Graham, Jinko, 11, 26, 72, 147
Granville, Kevin, 28, 159
Greenwood, Celia, 45, 243
Griffith, Lauren, 45, 239
Griffith, Skye Paphora, 46, 246
Gronsbell, Jessica, 10, 68
Gu, Yue, 24
Gu, Yuwen, 12, 77
Guan, Tianyu, 25, 142
Guo, Beibei, 16, 100
Guo, Hui, 24, 52, 282
Guo, Jing, 24
Guo, Lulu, 49, 264
Gustafson, Paul, 23, 34, 53, 135, 184, 291
Gutierrez, Eduardo, 23

Hagar, Luke, 19, 115
Hamidi, Hamid, 32
Han, Tian, 10, 67
Han, Yansan, 52, 283
Handcock, Mark S, 126
Handcock, Mark S., 21
Hanley, James A., 55, 299
Haque, Md Ashiqul, 24
Harris, Les N., 35, 191
Hassan, Samah, 21, 127
Haziza, David, 13, 81
He, Wenqing, 21, 23, 38, 50, 127, 137, 211, 270
Heath, Anna, 22, 131
Heckman, Nancy, 46, 249
Hector, Emily, 33, 178
Hellingman, Sean, 18, 114
Heo, Giseon, 55, 298
Herrmann, Klaus, 37, 203
Hinton, Alexander, 40, 216
Ho, Lam, 51, 276
Ho, Nhat, 36, 196
Hodges, Toby, 54, 293
Hofert, Marius, 37, 203
Hoque, Md Rashedul, 41, 220
Hoque, S.M. Nehanul, 32
Horst, Allison, 36, 200
Hossain, Belal, 24
Hossain, Shakhawat, 14, 38, 42, 88, 206, 224
Hou-Liu, Jason, 52, 282
Hu, Boyi, 46, 245
Hu, Jianhua, 42, 227
Hu, Pingbo, 34, 185
Hu, Pingzhao, 25, 141
Hu, X. Joan, 43, 233
Huang, Jingyue, 18, 111
Huang, Mei Ling, 52, 283
Huang, Shiheng, 24
Huang, Zhenzhen, 27, 151
Hughes, David, 21, 56, 127, 304
Hunt, Beverley, 22, 131
Hunter, David R., 56, 306
Hunter, William, 15, 95

Ilagan, Michael John, 52, 280

Jaimungal, Sebastian, 26, 150
Jalbert, Jonathan, 55, 301
Jayaraman, Sarath Kumar, 13, 84
Jayasooriya, Apsara Pathum, 27, 154
Jessup, Sébastien, 17, 24, 107
Jevtic, Petar, 58, 314
Jhangiani, Surita, 54, 292
Jiang, Bei, 12, 80
Jiang, Cong, 38, 210
Jiang, Depeng, 16, 101
Jiang, Shu, 25, 142
Jiang, Wenjun, 30, 168
Joe, Harry, 58, 318
Ju, Hanwen, 24
Jung, Hyejung, 52, 279
Jung, Michael, 30, 166

Kalaria, Shreena Nisha, 24
Kalbfleisch, Jack, 16, 97
Kamath, Ravish, 31
Kang, Sohee, 33, 181
Kanters, Steve, 47, 253
Karunanayaka, Ruwan C., 11, 72
Kashlak, Adam B., 34, 55, 187, 298
Katz, Alan, 31, 174
Kelly, Mary, 30, 170
Kennedy, Edward H., 18, 112
Khadem Charvadeh, Yasin, 23, 136
Khalili, Abbas, 36, 53, 196, 289
Khan, Gabriel, 10, 67
Khan, Mohammad Kaviul Anam, 22, 129
Khan, Shahedul, 42, 47, 224, 251
Kharaghani, Amin, 22, 128
Khern-am-nuai, Warut, 56, 307
Klein, Lauren, 26, 145
Kluger, Dan, 24, 139
Kochoedo, Maryse, 55, 299
Koh, Ryan, 21, 127
Kong, Dehan, 42, 59, 228, 321
Kong, Linglong, 12, 33, 53, 59, 80, 180, 287, 321
Konstantinidis, Menelaos, 49, 263
Kornblith, Lucy, 22, 131
Kouritzin, Michael A., 13, 85
Krahn, Jody, 42, 224
Kroell, Emma, 26, 150
Kulperger, Reg, 37, 204
Kumbhare, Dinesh, 21, 127
Kustra, Rafal, 22, 129

Labbe, Aurélie, 22, 28, 42, 44, 49, 129, 158, 226, 238, 266
Lakmali, Muditha, 24

Lalancette, Michaël, 32, 55, 176, 301
Landsman, Victoria, 38, 207
Larocque, Denis, 28, 42, 49, 158, 225, 266
Latremouille, Derek, 59, 321
Lavigne, Eric, 58, 314
Lawler, Ethan, 50, 273
Lawler, Patrick, 22, 131
Lawless, Jerald F., 17, 47, 104, 256
Le Blanc, Richard, 59, 322
Lee, Chel Hee, 21, 127
Lee, Hongzhe, 47, 255
Lee, Melissa, 54, 292
Lefebvre, Geneviève, 41, 42, 220, 227
Legendre Bilodeau, Sarah, 9, 47, 63, 254
Lei, Mengying, 22, 129
Leifer, Eric, 22, 131
Leiva, Ricardo, 56, 303
Lemieux, Marie-Pier, 34, 189
Lemzouji, Khalid, 9, 57, 61, 310
Letac, Gerard, 29, 165
Li, Bosheng, 16, 101
Li, Deli, 58, 318
Li, Fan, 13, 82
Li, Harrison, 24, 139
Li, Jingyu, 31
Li, Longhai, 25, 47, 140, 252
Li, Mufan, 54, 296
Li, Na, 21, 127
Li, Pengfei, 40, 213
Li, Shu, 17, 109
Li, Wuchen, 10, 67
Li, Xiaoting, 58, 318
Li, Yifan, 37, 204
Liang, You, 54, 294
Lim, Lily S. H., 49, 263
Lin, Liyuan, 27, 151
Lin, Lynn, 43, 229
Lin, X. Sheldon, 27, 41, 150, 218
Lin, Zhenhua, 42, 228
Liseo, Brunero, 38, 207
Liu, Dan, 38, 211
Liu, Dongmeng, 11, 72
Liu, Fang, 12, 79
Liu, Hua, 46, 245
Liu, Juxin, 47, 255
Liu, Suyu, 16, 100
Liu, Xiaohua, 19, 115
Liu, Xiaotong, 59, 321
Liu, Xiaoyan, 32
Liu, Yi, 11, 71
Liu, Zhihui (Amy), 33, 45, 183, 244
Lix, Lisa M., 56, 303
Locas, Félix, 37, 204
Lockhart, Richard, 25, 31, 144, 173
Loek, Melvin, 10, 67
Loewen, Dan, 51, 279
Loliencar, Prachi, 55, 298
Long, Quan, 45, 242
Lopera, Pablo Andres, 23
Lou, Wendy, 15, 91
Loughin, Thomas, 33, 182
Lourenzutti, Rodolfo, 10, 66, 325
Lovblom, Leif Erik, 34, 186
Lu, Xuewen, 27, 153
Lu, Yi, 17, 108
Lu, Zhe, 20, 117
Lu, Zihang, 15, 59, 91, 321
Luo, Kexin, 28, 158
Luo, Lan, 33, 178
Luu, Son, 31
Lysy, Martin, 34, 35, 187, 191, 194, 195
Lyu, Guanjie, 48, 259
Lyu, Qi, 27, 154
Lyu, Xiangyu, 32
Lyu, Yunhong, 37, 205

Ma, Jinhui, 38, 45, 208, 239
Ma, Junling, 34, 185
Ma, Renjun, 30, 168
Macdonald, Brian, 40, 216
Mackay, Anne, 13, 85
Mahmoudi, Fatemeh, 27, 153
Mailhot, Mélina, 17, 20, 37, 107, 121, 202
Malloy, Shane, 30, 166
Mandal, Saumen, 52, 54, 284, 295
Manole, Tudor A., 36, 53, 196, 289
Mao, Daniel, 24
Marchand, Eric, 22, 131
Marriott, Paul, 10, 14, 67, 87
Marshall, François A., 28, 159
Marshall, William, 52, 283
Martineau, Patrice, 57, 309
Martsynyuk, Yuliya V., 45, 58, 245, 317
Massam, Helene, 29, 164
Matharaarachchi, Surani, 31, 174
Mayhew, Alexandra, 45, 239

Mazza-Anthony, Cody, 36, 196
McCandless, Lawrence, 41, 220
McFayden, Colin B., 28, 159
McGee, Glen, 28, 157
McGregor, Kevin, 48, 257
McNealis, Vanessa, 18, 111
McNeney, Brad, 26, 147
McNichol, Jennifer, 55, 299
McNicholas, Paul D., 11, 51, 71, 278
Meng, Di, 37, 204
Merz, Michael, 37, 202
Metzler, Adam, 37, 204
Miao, Yu, 58, 318
Michal, Victoire, 22, 133
Milic, Milos, 22, 128
Mills Flemming, Joanna, 50, 52, 273, 286
Min, Joosung, 31
Mitani, Aya A., 38, 44, 207, 238
Mitrache, Christian, 31
Molladavoudi, Saeid, 29, 162
Monkman, Martin, 57, 311
Montgomery, Jamie, 36, 200
Montufar, Guido, 10, 67
Moodie, Erica E.M., 18, 27, 48, 110, 111, 112, 153, 155, 257
Moon, Nathalie, 15, 94
Moore, Lynne, 35, 193
Moreau, Clara, 45, 243
Morenz, Eric, 41, 221
Morrison, Conor, 34, 184
Morrison, Tim, 24, 139
Mortensen, Jacob, 30, 166
Mossman, Alexandra, 24
Mousazadeh, Bahar, 32
Muller, Kirsten, 55, 300
Munaweera Arachchilage, Inesh Prabuddha, 35, 191
Murphy, Ian, 46, 249
Murphy, Orla, 11, 71
Murua, Alejandro, 36, 197
Mussavi Rizi, Marzieh, 38, 210
Muthukumarana, Saman, 31, 35, 51, 174, 191, 279

Nadarajah, Tharshanna, 15, 93
Nadeem, Hira, 27, 155
Nan, Bin, 16, 97
Nasri, Bouchra, 14, 17, 86, 105
Neal, Matthew, 22, 131
Neal, Radford M., 18, 114
Negeri, Zelalem F., 11, 73
Neish, Drew, 53, 289
Nešlehová, Johanna G., 18, 25, 37, 110, 144, 203
Nettel-Aguirre, Alberto, 56, 303
Ngi-Song, Adele, 42, 226
Nguyen, Dinh-Toan, 14, 86
Nguyen, Nga, 32
Nguyen, Vi, 24
Nia, Vahid, 43, 232
Nkurunziza, Sévérien, 17, 34, 37, 105, 186, 205
Nolde, Natalia, 20, 32, 121, 176
Nosyk, Bohdan, 43, 233
Nyein, Thet Htet Chan, 32

Oganisian, Arman, 34, 183
Okaeme, Nneka, 48, 257
Osgood, Nathaniel David, 39, 211
Osuntuyi, Anthony, 55, 297
Oualkacha, Karim, 44, 238
Owen, Art, 24, 139

Palacios Rodriguez, Fatima, 20, 121
Pan, Yongwen, 24
Pandher, Sharandeep Singh, 38, 206
Panju, Maysum, 43, 231
Park, Jonghoon, 31
Park, Sungki, 24
Parker, Matthew R.P., 34, 185
Parsons, Marc Angelo, 58, 317
Paterson, Phyllis G., 15, 95
Payne, Andrea, 51, 278
Peiris, Shelton, 54, 294
Peitsch, Jeffrey W., 57, 312
Peng, Yingwei (Paul), 46, 251
Perkins, Bruce A., 34, 186
Perreault, Andrea, 50, 274
Perreault, Samuel, 31, 173
Perrie, Will, 57, 312
Peruzzi, Antonio, 22, 128
Pesenti, Silvana Manuela, 26, 37, 150, 202
Philémon, Gamet, 55, 301
Picinini Freitas, Laís, 22, 133
Pigeon, Mathieu, 17, 107
Platt, Robert, 12, 76
Pokharel, Gyanendra, 14, 89
Poline, Jean-Baptiste, 45, 243
Poulos, Jason, 13, 82
Pramij, Shenita, 42, 223

Provost, Serge B., 48, 260
Pullenayegum, Eleanor M., 36, 44, 48, 49, 198, 238, 262, 263

Qi, Weinan, 14, 87
Qi, Yangqian, 31, 172
Qian, Yanzhao, 32
Qian, Yi, 12, 41, 79, 220
Qin, Xiaoke, 14, 90
Qu, Annie, 16, 98
Quan, Samuel, 24

Rahman, Azizur, 24
Raina, Parminder, 38, 45, 208, 239
Ramdas, Aaditya, 18, 112
Ramezan, Reza, 34, 187
Ramkissoon, Jonathan, 35, 194
Rang, Guanglin, 44, 235
Ratnasekera, Pulindu, 26, 147
Raymond, Henry F., 21, 126
Reesor, Mark, 37, 204
Reid, Nancy, 32, 175
Reiter, Jerome, 29, 163
Rémillard, Bruno, 14, 17, 25, 86, 105, 144
Ren, Jiandong, 17, 30, 109, 169
Reyna, Myrtha, 59, 321
Ribaud, Melina, 44, 238
Rice, Gregory, 46, 246
Robert, Christian P., 55, 59, 298, 320
Robinson, Deniza, 32
Romanovska, Sofiia, 37, 200
Rosadi, Dedi, 54, 294
Rosenthal, Jeffrey S., 23, 135
Rosner, Bernard, 25, 142
Rothstein, Steven, 51, 278
Rotondi, Michael, 15, 95
Rotondi, Nooshin Khobzi, 15, 95
Roy, Anuradha, 56, 303
Roy, Daniel M., 41, 221
Rudoler, David, 15, 95
Ruth, William, 31, 173

Saarela, Olli, 33, 45, 183, 244
Saini, Jessica, 24
Sajobi, Tolu, 56, 303
Sajobi, Tolulope, 21, 127
Samoilenko, Mariia, 41, 42, 220, 227
Samworth, Richard J., 20, 119
Sang, Peijun, 46, 248
Sanusi, Olayinka, 15, 95
Sari, Eyyüb, 43, 232
Savy, Nicolas, 27, 153
Scassa, Teresa, 29, 162
Schaubroeck, Matt, 51, 279
Schmidt, Alexandra M., 22, 53, 58, 133, 287, 314
Schmidt, Philip J., 55, 300
Schneider, Julie A., 22, 128
Schnitzer, Mireille, 12, 33, 76, 182
Schroeppel, Philipp, 31
Schulz, Juliana, 48, 257
Selvaratnam, Selvakkadunko, 58, 313
Selvitella, Alessandro Maria Maria, 49, 268
Severino, Federico, 31, 173
Seyed-Ahmadi, Arman, 10, 66, 325
Shang, Zuofeng, 46, 248
Shen, Hua, 20, 36, 117
Shen, Hui, 57, 312
Shen, Junwei, 27, 155
Shen, Pao-sheng, 46, 251
Shen, Ruoqi, 54, 296
Shen, Ye, 33, 179
Shen, Yi, 14, 87
Shestopaloff, Alexander, 18, 114
Shi, Haolun, 46, 249
Shi, Yu, 24
Shi, Yuliang, 18, 110
Shortreed, Susan, 18, 112
Si, Yajuan, 57, 308
Sidrow, Evan, 46, 249
Silva, Anjali, 51, 278
Simmons, Susan, 37, 201
Simonoff, Jeffrey S., 42, 225
Singh, Gurbakhshash, 38, 208
Singh, Japjeet, 54, 295
Sinha, Ritwik, 18, 112
Sirois, Caroline, 42, 224
Sixta, Sabrina, 14, 86
Skrzydlo, Diana Katherine, 15, 94
Slater, Justin James Ian, 23, 135
Slessor, Jordan A., 40, 214
So, Hon-Yiu, 38, 45, 208, 239
Soave, David, 30, 170
Sobhan, Shamsia, 22, 132
Solymos, Peter, 9, 57, 61, 310
Song, Jian, 44, 235
Song, Peter X., 33, 178
Song, Rui, 33, 179

Song, Yanglei, 58, 316
Soufiani, Elham, 37, 205
Spicker, Dylan Z., 44, 237
Steeves, Holly N., 37, 200
Stephens, David A., 27, 153
Stevens, Nathaniel T., 19, 115
Stewart, Connie, 55, 299
Stoica, George, 58, 318
Stringer, Alex, 28, 157
Strug, Lisa, 16, 102
Stryhn, Henrik, 49, 267
Stukel, Thérèse A., 25, 29, 144, 161
Su, Wanhua, 52, 280
Subedi, Sanjeena, 14, 53, 89, 289
Subramani, Pranav, 35, 194
Subramaniam, Shoba, 21, 127
Sun, Ke, 12, 80
Sun, Lei, 22, 25, 128, 144
Sun, Lijun, 22, 129
Sun, Meng, 17, 108
Sun, Shuo, 18, 110
Sun, Yifan, 38, 209
Suresh, Parvathy, 31
Surjanovic, Nikola, 23, 133
Syed, Saifuddin, 23, 133

Taback, Nathan A., 26, 40, 214
Tadesse, Mahlet G., 12, 76
Talbot, Denis, 42, 224
Tan, Yiren, 24
Tang, Boxin, 11, 72
Tang, Rebecca, 24
Tang, Thai-Son, 33, 45, 183, 244
Tang, Yanbo, 14, 87
Tao, Jing, 57, 312
Tchetgen Tchetgen, Eric, 21, 124
Tellez, Martha, 23
Thabane, Lehana, 45, 239
Thakur, Rohit, 23
Thavaneswaran, Aerambamoorhty, 294, 295
Thavaneswaran, Aerambamoorhty A., 54
Thioub, Mamadou Yamar, 14, 86
Thombs, Brett D., 11, 73
Thompson, Mary E., 18, 38, 50, 114, 210
Thomson, Trevor, 43, 233
Thulasiram, Ruppa K, 54, 295
Tian, Jizhou, 11, 71
Tim, Dockhorn, 43, 232
Timbers, Tiffany A., 9, 54, 64, 292
Tindel, Samy, 44, 235
Tio, Earvin, 22, 128
Tomal, Jabed, 20, 120
Tomlinson, George, 34, 186
Torabi, Mahmoud, 22, 132
Tremblay, Veronique, 56, 306
Trites, Andrew W., 46, 249
Trotz-Williams, Lise A., 49, 267
Tseung, Spark, 27, 150
Tsybakin, Aleksandr, 24
Tu, Wangshu, 14, 89
Turchetta, Armando, 27, 153
Turgeon, Max, 38, 209
Tyuryaev, Vadim, 24

Valipour, Mojtaba, 43, 231
Varian, Hal, 24, 139
Variyath, Asokan Mulayath, 27, 154
Verschoor, Chris, 45, 239
Vishnyakova, Olga, 31, 45, 241
Volgushev, Stanislav, 32, 55, 176, 301
Volodin, Andrei, 38, 206

Wüthrich, Mario V., 37, 202
Wallace, Michael, 38, 44, 210, 237
Wang, Guan, 45, 242
Wang, Guanbo, 12, 76
Wang, Haixu Alex, 45, 244
Wang, Jianan, 32
Wang, Jie, 23, 137
Wang, Liangliang, 23, 35, 50, 137, 193, 270
Wang, Lijia, 41, 222
Wang, Linbo, 41, 42, 221, 228
Wang, Meng, 21, 127
Wang, Nanwei, 29, 164
Wang, Rui, 12, 76
Wang, Ruodu, 27, 151
Wang, Shijia, 50, 270
Wang, Xu (Sunny), 55, 297
Wang, Yafei, 46, 248
Wang, Ye, 30, 168
Wang, Yiran, 35, 191
Wang, Yue, 16, 97
Wang, Zhenhua, 13, 82
Wang, Zilin, 18, 30, 114, 170
Ward, Caitlin, 53, 287
Ward, Madeline, 57, 313

Waudby-Smith, Ian E., 18, 112
Wei, Pengyu, 27, 151
Weichenthal, Scott, 58, 314
Welch, William J., 20, 120
Weldon, Kenneth Laurence, 16, 103
Welsh, Liam, 56, 304
Weng, Chengguo, 27, 151
Weng, Yijia, 23, 24, 136
West, Brady, 57, 308
Whitaker, Douglas, 15, 93
White, Bethany J.G., 17, 103
Wickramasinghe, Ashani N., 51, 279
Wiebe, Samuel, 56, 303
Willard, James, 22, 132
Williamson, Tyler, 56, 303
Wong, Octavia, 24
Wong, Ting Kam Leonard, 10, 67
Woolford, Douglas G., 28, 58, 159, 320
Wright, Emily, 24, 37, 202
Wright, Peter G., 57, 309
Wu, Changbao, 10, 18, 65, 111, 324
Wu, Haoyu, 23, 135
Wu, Jeff, 13, 83
Wu, Michelle, 52, 279
Wu, Mohan, 35, 194
Wu, Sidi, 11, 72
Wu, Tingxuan, 47, 252
Wu, Weichi, 17, 106
Wu, Yanyan, 29, 164

Xi, Dexen D.Z., 14, 90
Xian, Chenqian, 24
Xie, Hui, 12, 41, 49, 79, 220, 264
Xie, Jinhan, 25, 143
Xing, Li, 53, 291
Xiong, Juan, 50, 270
Xu, Changchang, 41, 223
Xu, Chao, 13, 81
Xu, Hainan, 24
Xu, Wei, 25, 140

Yan, Fangrong, 16, 101
Yan, Guohua, 30, 168
Yanez, Juan-Sebastian, 41, 219
Yang, Ce, 20, 118
Yang, Chengxin, 29, 163
Yang, Po, 18, 19, 114, 115
Yang, Xiao, 23
Yang, Yi, 12, 77
Yang, Yunfeng, 49, 268
Yao, Weichi, 42, 225
Yeh, Chi-Kuang, 46, 246
Yi, Grace Y., 23, 44, 136, 137, 237
Yi, Li, 49, 266
Yi, Liu, 12, 80
Yi, Yanqing, 27, 154
Yin, Mingren, 17, 108
Yoon, Thomas, 35, 190
You, Bowen, 43, 231
You, Jinhong, 46, 245
Yu, Daisy, 31
Yu, Hao, 37, 204
Yu, Jianping, 24
Yu, Tingting, 12, 77
Yu, Yaoliang, 43, 232
Yuan, Yan, 20, 117
Yuan, Ying, 16, 100

Zafari, Golara, 55, 296
Zamar, Ruben H., 20, 120
Zang, Yishan, 48, 260
Zapata-Marin, Sara, 58, 314
Zarychanski, Ryan, 22, 131
Zeng, Leilei, 18, 43, 111, 233
Zeng, Michelle, 31, 172
Zerguini, Ghislene, 47, 253
Zhang, Archer Gong, 58, 316
Zhang, Jiarui, 23, 59, 137, 322
Zhang, Jun, 10, 67
Zhang, Matthew, 54, 296
Zhang, Qingrun, 26, 147
Zhang, Qiong, 53, 290
Zhang, Shi, 30, 168
Zhang, Ting, 45, 243
Zhang, Wensha, 31
Zhang, Xi, 11, 71
Zhang, Xiaoqing, 55, 300
Zhang, Xiawen, 48, 262
Zhang, Xinyi, 21, 123
Zhang, Xuekui, 50, 53, 270, 291
Zhang, Yuexia, 42, 227
Zhao, Kaiqiong, 53, 287
Zhao, Lihui, 43, 229
Zhao, Yingqi, 32, 178
Zhao, Yue, 12, 77
Zhou, Meng, 58, 316

Zhou, Menglin, 20, 121
Zhou, Xiaowen, 51, 276
Zhou, Zhou, 17, 106
Zhu, Ji, 15, 97
Zhu, Yeying, 18, 41, 110, 222
Zimmerman, Robert, 48, 259
Zoglat, A., 48, 261
Zolnouri, Mahdi, 43, 232