(Bio)Statistics Research and Career Day

September 21st, 2018
McGill University

Bronfman Building
Underground, Room 001
1001 rue Sherbrooke Ouest

9:00–9:30 Registration

9:30–10:15 **Dr. Luke Bornn: Possession Sketches: Mapping NBA Strategies**

10:15–10:30 Coffee Break

10:30–10:50 Chien-Lin Mark Su: Doubly Robust Estimation and Causal Inference for Recurrent Event Data

10:50–11:10 Menglan Pang: Flexible Accelerated Failure Time Model for Right Censored Data in Survival Analysis

11:10–11:30 Asad Haris: Nonparametric regression with adaptive truncation via a convex hierarchical penalty

11:30–11:50 Tyrel Stokes: Unmeasured Confounding, Bias Amplification, and Model Selection

11:50–12:00 Break

12:00–13:00 **Dr. Joelle Pineau: Improving health-care: challenges and opportunities for reinforcement learning**

13:00–2:20 Lunch and poster presentations

Anthony Coache: The Significance of the Adjusted R Squared Coefficient

Rajib Dey: A Non-parametric Estimator of Time-dependent AUC in the Presence of Competing Risks

Asad Haris: Convex Modeling Interactions with Strong Heredity

Gabrielle Simoneau: Non-regular Inference for Dynamic Weighted Ordinary Least Squares: Understanding the Impact of Solid Food Intake in Infancy on Childhood Weight

Hao Zhang: A novel mixed method approach for Bayesian prior elicitation and its application in factor analysis

Ting Zhang: On Simulation Design for Evaluating Type I Error: What is the 'correct' null model?

Kaiqiong Zhao: Smooth modelling of covariate effects in bisulfite sequencing-derived measures of DNA methylation

14:20–14:40 Nicholas Beck: Estimation of extreme storm-surges using a spatial GEV model

14:40–15:00 James Hugh McVittie: Parametric modelling of combined incident and length-biased failure time data

15:00–15:20 Mohamad Elmasri: Tree-Dependent Decomposable Graphs

15:20–15:35 Coffee Break

15:35–16:35 **CANSSI distinguished visitor Dr. James Robins: Causal Inference and Machine Learning: Improved Inference under No Assumptions**

16:35–17:00 Transition to Purvis Hall

Thomson House, Underground
3650 McTavish St

17:00–19:30 Career panel and Cocktail.

Panelists:

Sahir Bhatnagar, McGill University

Félix Boyer, Normandin Beaudry

Geneviève Lefebvre, Université du Québec à Montréal

Azadeh Shohudi Mojdehi, Analysis Group

Jonathan Moscovici, QuintilesIMS

Luc Perreault, Hydro-Québec Research Institute (IREQ)

Members of the organizing committee: Simon Chatelain, Steve Ferreira, Yu Luo and Guanbo Wang.

Thanks to our sponsors: CANSSI, Department of Mathematics, Statistics, Department of Epidemiology, Biostatistics and Occupational Health and Post-Graduate Students' Society of McGill University.

We would like to thank the people who helped in organizing the event: André-Yves Gagnon, Katherine Hayden, Marie Boncoeur Erica Moodie and Gill Paradis from the EBOH Department, and Jason Stillman, Angela White, Russell Steele and David A. Stephens from the Department of Mathematics and Statistics.

Abstracts

Keynote Speakers

Possession Sketches: Mapping NBA Strategies

Luke Bornn, Assistant Professor of Statistics at Simon Fraser University, Vice President of Strategy and Analytics at Sacramento Kings

We present Possession Sketches, a new machine learning method for organizing and exploring a database of basketball player-tracks. Our method organizes basketball possessions by offensive structure. We first develop a model for populating a dictionary of short, repeated, and spatially registered actions. Each action corresponds to an interpretable type of player movement. We examine statistical patterns in these actions, and show how they can be used to describe individual player behavior. Leveraging this vocabulary of actions, we develop a hierarchical model that describes interactions between players. Our approach draws on the topic-modeling literature, extending Latent Dirichlet Allocation (LDA) through a novel representation of player movement data which uses techniques common in animation and video game design. We show that our model is able to group together possessions with similar offensive structure, allowing for efficient search and exploration of the entire database of player-tracking data. We show that our model finds repeated offensive structure in teams (e.g. strategy), providing a much more sophisticated, yet interpretable lens into basketball player-tracking data. This is joint work with Andrew Miller.

Improving health-care: challenges and opportunities for reinforcement learning

Joelle Pineau, Associate Professor and co-director of the Reasoning and Learning Lab in the School of Computer Science at McGill University, Head of the Facebook AI Research lab

Reinforcement learning offers a powerful paradigm for automatically discovering and optimizing sequential treatments for chronic and life-threatening diseases. In particular, we will focus on how data collected in multi-stage sequential trials can be used to automatically generate treatment strategies that are tailored to patient characteristics and time-dependent outcomes. We will also examine promising methods to improve the efficiency of clinical trials through adaptation. Examples will be drawn from several ongoing research projects on developing new treatment strategies for epilepsy, mental illness, diabetes, and cancer.

Causal Inference and Machine Learning: Improved Inference under No Assumptions

James Robins, Mitchell L. and Robin LaFoley Dong Professor of Epidemiology, Department of Epidemiology and Biostatistics at Harvard School of Public Health

I describe new methods for improving confidence intervals for causal effects when the outcome regression and propensity score function have been fit using machine learning algorithms with completely unknown statistical properties. Higher order influence function based tests and estimators are the basis of the methodology.

Students & Post doc fellows

Doubly Robust Estimation and Causal Inference for Recurrent Event Data

Chien-Lin Mark Su

Recurrent events are frequently observed in many biomedical longitudinal studies. The interest of this paper is to estimate the average causal effects for recurrent event data in the presence of confounders. We propose a doubly robust estimator which combines the weighted Nelson-Aalen estimator and regression estimator based on an assumed semiparametic multiplicative rate model for recurrent event data. The proposed estimators are shown to be consistent and asymptotically normal. In addition, a model diagnostic plot of residuals is presented to assess the adequacy of the semiparametric model. The finite sample behavior of the proposed estimators is evaluated through simulation studies. The proposed methodologies are illustrated via an injury database for circus artists.

Flexible Accelerated Failure Time Model for Right Censored Data in Survival Analysis

Menglan Pang

Background: The accelerated failure time (AFT) model has been suggested as an interesting alternative to the Cox proportional hazards model. However, a parametric AFT model requires the specification of an appropriate distribution for the event time, which is often difficult to identify in real-life studies and, thus, may limit applications. Methods that are more robust with respect to event time distribution specification are desirable. Recently, a semiparametric AFT model was developed by Komarek et al. based on smoothed error

distribution. This method allows for estimating covariates effects and predicting the hazard and survival probabilities for a given covariate pattern while leaving the distribution unspecified.

Methods: We develop a new flexible AFT model that also does not need the specification of the parametric family of event time distribution. The baseline hazard function is modeled by regression B-splines and thus allows for the estimation of arbitrary shapes. In comprehensive simulations, we validate the performance of our approach in terms of effect estimates, baseline hazard and survival probabilities, and compare with the results from parametric AFT models and the approach of Komarek.

Results: The survival probabilities estimated by parametric AFT models with misspecified event time distribution deviated from the truth. Both the proposed flexible AFT model and the approach of Komarek provided unbiased effects estimates and unbiased survival curves for a variety of scenarios in which the event time follow different distributions, including conventional parametric models and more complex mixture distributions. However, the proposed flexible AFT model always yielded more stable estimates of the hazard function.

Conclusion: Our flexible AFT model provides a useful approach to analyze survival data, and can provide insights regarding how a prognostic factor affects survival.

Nonparametric regression with adaptive truncation via a convex hierarchical penalty

Asad Haris

We consider the problem of nonparametric regression with a potentially large number of covariates. We propose a convex, penalized estimation framework that is particularly well-suited for high-dimensional sparse additive models and combines appealing features of finite basis representation and smoothing penalties. In the case of additive models, a finite basis representation provides a parsimonious representation for fitted functions but is not adaptive when component functions posses different levels of complexity. In contrast, a smoothing spline type penalty on the component functions is adaptive but does not provide a parsimonious representation. Our proposal simultaneously achieves parsimony and adaptivity in a computationally efficient way. We establish minimax rates for a large class of sparse additive models. We also develop an efficient algorithm that scales similarly to the lasso with the number of covariates and sample size.

Unmeasured Confounding, Bias Amplification, and Model Selection

Tyrel Stokes

In ignorability approaches to identifying Average Causal effects, we seek to condition on a set of variables such that the potential outcomes are conditionally independent of the treatment [Rubin (1974),Rubin and Rosenbaum (1983), Wooldridge (2010)] . In non-experimental

settings, however, the entire set of variables required to satisfy the ingorability condition is rarely available and confounding paths remain, biasing estimation of the causal effects. Given recent theoretical work in bias amplification [Wooldridge (2013),Pearl (2012),Pearl (2011),Ding et al (2017)],we must be careful not to include variables explaining a large portion of the variance of the treatment relative to the variance explained in the outcome. There are in fact cases where we are better off not conditioning on variables, even those which form confounding paths either through the unmeasured confounding or independently and indirectly with the outcome. In other words,some confounding paths should not be blocked in order to minimize bias. Variables which strongly predict variation in the treatment, in particular, should not be included in ignorability approaches since they have the greatest potential to amplify existing bias.

In this paper, we extend the bias amplification literature in three important ways. First, we extend the OLS bias expectation results to a larger class of plausible DAGs and non-standaradized variables. Second, using the insights from the unstandardized variables we reframe the origin of bias amplification in terms of variables which account for a large amount of variance in the treatment and relatively little additional variance in the outcome. In OLS, this can be shown geometrically and following Middleton et al, we show that we can estimate the amplifying factor in the probability limit. Third, we show that although we do not have closed form solutions that bias amplification extends to several notable non-linear and machine learning techniques such as neural networks and lasso regression. Further, we show in simulation studies that even given access to a large number of observable variables which account for the large majority of unmeasured confounding, bias amplification can still be dramatic, as demonstrated theoretically in Pearl (2011). Automated variable selection techniques are not adequate to exclude amplifying variables, particularly when prediction accuracy is the selection criteria used for determining the sparseness penalty. This suggests that model selection techniques must consider amplification in both machine learning and high dimensional (large p) settings.

Estimation of extreme storm-surges using a spatial GEV model

Nicholas Beck

All around the world, floods are considered to be the most catastrophic of natural disasters, both in damages and the number of victims. Moreover, given how the environment is rapidly changing, accurately being able to predict and protect ourselves from such catastrophes is of paramount importance. In Canada, insurers began offering protection against such events in 2015. While products generally cover both rainfall and river flood protection, there is little offered in the way of protection from coastal floods. The goal of this project is to estimate the occurrence of extreme storm-surges and the subsequent risk of coastal flooding in Atlantic Canada. To this end, a spatial Bayesian hierarchical model is used to model annual maximum storm-surges. Additionally, we introduce a copula to our hierarchy as a means to further strengthen our measure of dependence between stations. With this information we could, for example, determine the appropriate premium for coastal flood protection.

Parametric modelling of combined incident and length-biased failure time data

James Hugh McVittie

In certain settings, it may be possible to collect independent samples of incident and prevalent cohort data in which both data sets are relevant to the same research question. Under the assumption of a parametric failure time model, we propose a combined cohort maximum likelihood estimator for the unknown failure time distribution parameters. Using simulated data sets, we empirically examine how the inclusion of data from both types of cohort improves the asymptotic variance of the estimator. We apply the combined maximum likelihood estimator to estimate the career lengths of players in the National Basketball Association between 1990 and 2008.

Tree-Dependent Decomposable Graphs

Mohamad Elmasri

Decomposable graphs are known for their tedious and complicated Markov update steps. Instead of modelling them directly, this work introduces a novel representation of decomposable graphs based on a class of tree-dependent bipartite graphs that span the projective space of the former. The novel representation has a few main benefits. First it allows for a new node-driven Markov chain Monte Carlo sampler of decomposable graphs that can easily parallelize and scale. The proposed sampler also benefits from the computational efficiency of junction-tree-based samplers. Second, it enables a form of sub-clustering within maximal cliques of the graph, adding informational richness to the general use of decomposable graphs that could be harnessed in applications with behavioural type of data.

Poster Presentations

The Significance of the Adjusted R Squared Coefficient

Anthony Coache

This poster explains what, exactly, the adjusted R squared coefficient adjusts for. Motivated by the theoretical soundness of the explanatory framework, we suggest new exact model selection tests for GLMs. Joint work with Olivier Binette.

A Non-parametric Estimator of Time-dependent AUC in the Presence of Competing Risks

Rajib Dey

Competing risks is a very important problem in survival analyses. In competing risk the event time for an individual can be classified as one of the several distinct causes. Evaluating a candidate biomarker or developing a predictive model score for event-time outcomes is one of the significant biomedical goals in order to distinguish accurately between incident cases from the controls surviving beyond t throughout the entire study period.

In this presentation, we will propose a non-parametric method to estimate the time-dependent Area under the curve (AUC) using weighted mean rank(WMR) estimator when we have censored survival times and competing risks. The proposed method will extend the time-dependent predictive accuracy measures of P. Saha and Heagerty (2012, Biostatistics, 1-18). We will investigate the performance of the proposed WMR estimator through both simulation studies and real-life data analysis before the presentation.

Convex Modeling Interactions with Strong Heredity

Asad Haris

We consider the task of fitting a regression model involving interactions among a potentially large set of covariates, in which we wish to enforce strong heredity. We propose FAMILY, a very general framework for this task. Our proposal is a generalization of several existing methods, such as VANISH [Radchenko and James, 2010], hierNet [Bien et al., 2013], the all-pairs lasso, and the lasso using only main effects. It can be formulated as the solution to a convex optimization problem, which we solve using an efficient alternating directions method of multipliers (ADMM) algorithm. This algorithm has guaranteed convergence to the global optimum, can be easily specialized to any convex penalty function of interest, and allows for a straightforward extension to the setting of generalized linear models.

Non-regular Inference for Dynamic Weighted Ordinary Least Squares: Understanding the Impact of Solid Food Intake in Infancy on Childhood Weight

Gabrielle Simoneau

A dynamic treatment regime (DTR) is a set of decision rules to be applied across multiple stages of treatments. The decisions are tailored to individuals, by inputing an individual's observed characteristics and outputting a treatment decision at each stage for that individual. Dynamic weighted ordinary least squares (dWOLS) is a theoretically robust and easily implementable method for estimating an optimal DTR. As many related DTR methods, the dWOLS treatment effects estimators can be non-regular when true treatment effects are zero or very small, which results in invalid Wald-type or standard bootstrap confidence intervals. Inspired by an analysis of the effect of diet in infancy on measures of weight and body size

in later childhood – a setting where the exposure is distant in time and whose effect is likely to be small – we investigate the use of the m-out-of-n bootstrap with dWOLS as method of analysis for valid inferences of optimal DTR. We provide an extensive simulation study to compare the performance of different choices of resample size m in situations where the treatment effects are likely to be non-regular. We illustrate the methodology using data from the PROmotion of Breastfeeding Intervention Trial to study the effect of solid food intake in infancy on long-term health outcomes.

A novel mixed method approach for Bayesian prior elicitation and its application in factor analysis

Hao Zhang

Questionnaire validation is fundamental when developing instruments for assessing latent psychometric properties. To estimate item-domain correlations, Bayesian Confirmatory Factor Analysis (CFA) supports the utilization of expert prior knowledge with smaller sample sizes compared to classic frequentist CFA. This assumes that investigators are able to provide relevant prior information with respect to the latent model, which could be potentially challenging. We propose a qualitative approach to collecting and decoding such information and a quantitative method aggregating the belief of all experts.

We apply the proposed prior acquisition to data of a diabetes empowerment questionnaire (28 questions with priors input from 6 domain experts). Priors were selected by the experts from six graphs of beta distribution. We synthesize in total 28 priors each by integrating six beta densities. We compared the estimated factor loadings and credible intervals with the results of a classic CFA using the same data, yielding substantial gain in precision of estimated parameters. Feasibility and acceptability of the expert survey was high.

We propose that a mixed method approach should be more routinely employed to inform Bayesian CFA for better knowledge translation and to improve the access of primary health care researchers to available and appropriate statistical methods.

On Simulation Design for Evaluating Type I Error: What is the 'correct' null model?

Ting Zhang

When evaluating a newly developed statistical test, an important step is to check its type 1 error (T1E) control using simulations. This is often achieved by the standard simulation design S0 under the so-called 'theoretical' null of no association. In practice, whole-genome association analyses scan through a large number of genetic markers (Gs) for the ones associated with an outcome of interest (Y), where Y comes from an alternative while the majority of Gs are not associated with Y; the Y - G relationships are under the 'empirical' null. This reality can be better represented by two other simulation designs, where design S1.1 simulates Y from an alternative model based on G, then evaluates its association with independently generated G_{new}; while design S1.2 evaluates the association between permutated Y and G. More than a decade ago, Efron (2004) has noted the important distinction

between the 'theoretical' and 'empirical' null in false discovery rate control. Using scale tests for variance heterogeneity, direct univariate, and multivariate interaction tests as examples, here we show that not all null simulation designs are equal. In examining the accuracy of a likelihood ratio test, while simulation design S0 suggested the method being accurate, designs S1.1 and S1.2 revealed its increased empirical T1E rate if applied in real data setting. The inflation becomes more severe at the tail and does not diminish as sample size increases. This is an important observation that calls for new practices for methods evaluation and T1E control interpretation.

Smooth modelling of covariate effects in bisulfite sequencing derived measures of DNA methylation

Kaiqiong Zhao

Identifying disease-associated changes in DNA methylation can help us gain a better understanding of disease etiology. Bisulfite sequencing allows the generation of high-throughput methylation profiles at single-base resolution. However, optimally modeling and analyzing these sparse and discrete sequencing data is still very challenging due to variable read depth, missing data patterns, long-range correlations, data errors, and confounding from cell type mixtures. We propose a regression-based hierarchical model that allows covariate effects to vary smoothly along genomic positions. We build a specialized EM algorithm which explicitly allows for experimental errors and cell type mixtures, to make inference about smooth covariate effects in the model. Simulations show that the proposed method provides accurate estimates of covariate effects and captures the major underlying methylation patterns. We also apply our method to analyze data from rheumatoid arthritis patients and controls.