The Fourth International Workshop on the Perspectives on High-dimensional Data Analysis
August 8–10, 2014

MEALS

Coffee Breaks: As per daily schedule, in the foyer of the TransCanada Pipeline Pavilion (TCPL) (*included in workshop*)

***For meal options at the Banff Centre, there are food outlets on The Banff Centre campus such as Vistas Main Dining Room on the 4th floor of Sally Borden Building (breakfast: 7:00-9:30am; lunch: 11:30am-1:30pm; dinner: 5:30-7:30pm), Le Cafe (ground floor, Sally Borden Building) and the Maclab Bistro (Kinnear Centre). You will also find a good selection of restaurants in the town of Banff which is a 10-15 minute walk from Corbett Hall.**)*

MEETING ROOMS

All lectures will be held in the lecture theater in the TransCanada Pipelines Pavilion (TCPL). An LCD projector, a laptop, a document camera, and blackboards are available for presentations.

SCHEDULE

Friday

16:00 Check-in begins (Front Desk - Professional Development Centre - open 24 hours)
Lecture rooms available after 16:00

Saturday

7:00–8:35 Breakfast
8:40–8:50 Welcoming Remarks from BIRS
8:50–9:00 Opening Remarks (Ejaz)
9:00–10:30 Session 1 (1.5 hours, each talk 30 min):
- Anand N. Vidyashankar
- Abbas Khalili
- Yang Feng

10:30–11:00 Coffee Break, TCPL
11:00–12:30 Session 2 (1.5 hours, each talk 30 min):
- Peter Song
- Karen Kafadar
- S. Ejaz Ahmed

12:30–1:30 Lunch
1:30–3:00 Session 3 (1.5 hours, each talk 30 min):
- Grace Y. Yi
- Joseph Verducci
- Ali Shojaie

3:00–3:30 Coffee Break, TCPL
3:30–5:30 Session 4 (1.5 hours, each talk 30 min):
- Bin Nan
- Andrei Volodin
- Matt Taddy
- Guoqing Diao

5:30–5:45 HDDA-V (Ejaz & Farouk)
5:45– Dinner

Sunday

7:00–9:00 Breakfast
9:00–10:30 Session 5 (1.5 hours, each talk 30 min):
- Michael J. Daniels
- Farouk S. Nathoo
- David A. Stephens

10:30–11:00 Coffee break, TCPL
11:00–12:30 Session 6 (1.0 hour, each talk 30 min):
- George Michailidis
- Pilar Muñoz
- Shuangge Ma

Checkout by 12 noon.
(Please checkout before before 9:00am or during the coffee break.)

The Fourth International Workshop on the Perspectives on High-dimensional Data Analysis
August 8–10, 2014

ABSTRACTS
(in alphabetic order by speaker surname)

Speaker: **Ahmed, S. Ejaz** (Brock University)
Title: *Big Data Big Bias Small Surprise!*
Abstract: In high-dimensional data settings where number of variables is greater than observations, or when number of variables are increasing with the sample size many penalized regularization approaches were studied for simultaneous variable selection and estimation. However, with the existence of covariates with weak effect, many existing variable selection methods may not distinguish covariates with weak signals and no signal. In this case, the prediction based on a selected submodel may not be desirable. In this talk, we propose a high-dimensional shrinkage estimation strategy to improve the prediction performance of a given submodel. Such a high-dimensional shrinkage estimator (HDSE) is constructed by shrinking a ridge estimator in the direction of a predefined candidate submodel. Under an asymptotic distributional quadratic risk criterion, its prediction performance is toughly investigated. We reveal that the proposed HDSE performs better than the full model estimator. More importantly, it significantly improves the prediction performance of any candidate submodel generated from most existing variable selection schemes. The relative performance of the proposed HDSE strategy is appraised by both simulation and the real data analysis.

Speaker: **Daniels, Michael J.** (University of Texas, Austin)
Title: *Causal inference on quantiles*
Abstract: We explore causal inference on quantiles using Bayesian nonparametric (BNP) methods in the presence of many confounders. In particular, we define relevant causal quantities and specify BNP models to avoid bias from restrictive parametric assumptions. Computational issues will be discussed including how computations scale up in the presence of many confounders. (Joint work with Dandan Xu, University of Florida.)

Speaker: **Diao, Guoqing** (George Mason University)
Title: *High Dimensional Covariate-Adjusted Semiparametric Transformation Gaussian Graphical Models*
Abstract: Estimating high-dimensional un-directed graphical models has received much attention in recent years. The standard method typically assumes that the variables are multivariate normally distributed leading to the so-called Gaussian graphical model. The nonparanormal graphical models (Liu, Lafferty and Wasserman, 2009) relax this assumption but do not account for covariates. We propose a semiparametric transformation Gaussian graphical model, in which each variable of interest is conditionally Gaussian given the covariates after an unspecified transformation. Additionally, the proposed model allows for heteroscedasticity. We develop likelihood-based inferential procedures and establish their asymptotic properties. Extensive simulation studies demonstrate that the proposed model outperforms the existing methods when their model assumptions are violated and is comparable to the existing methods under the true model specification. An application to a real data set is provided. (Joint work with Anand Vidyashankar)

Speaker: **Feng, Yang** (Columbia University)
Title: *Model Selection in High-Dimensional Mis-specified Models*
Abstract: Model selection is vital to high-dimensional modeling in selecting the best set of covariates among a sequence of candidate models. Most existing work assumes implicitly that the model under study is correctly specified or of fixed dimensions. Both model misspecification and high dimensionality are, however, common in real applications. In this paper, we investigate two classical Bayesian and Kullback-Leibler divergence principles of model selection in the setting of high-dimensional misspecified models. Asymptotic expansions of these model selection principles in high dimensions reveal that the effect of model misspecification is crucial and should be taken into account, leading to the generalized BIC and generalized AIC. With a natural choice of prior probabilities, we suggest the generalized BIC with prior probability ($GBIC_p$) which involves a logarithmic factor of the dimensionality in penalizing model complexity. We further establish the consistency of the covariance contrast matrix estimator in the general setting. Our results and new method are also supported by numerical studies.

Speaker: **Kafadar, Karen** (Indiana University; University of Virginia (8/26/14))
Title: *Distinguishing "typical" from "exotic" in streaming data sets*
Abstract: The analysis of massive, high-volume data sets stresses usual statistical software systems and requires new ways of drawing inferences beyond the conventional paradigm (optimal estimation of parameters from a hypothesized distribution), as the entire data set often cannot be read into the software system. Internet traffic data and data from high-energy particle physics experiments raise additional challenges: nearly continuous streams of observations from multiple systems or channels that interact and exchange information in nondeterministic ways. Internet data invite cyber attacks which can spread rapidly and thus require methods that can quickly detect abnormal behavior. New discoveries in particle physics arise when unexpected particle counts arise at specific high energies. This talk discusses analyses of data primarily from high-energy physics experiments, noting general considerations in analyzing high-volume complex data. (Physics collaborator: R.L. Jacobsen, UC-Berkeley.)

Speaker: **Khalili, Abbas** (McGill University)
Title: *Simultaneous variable selection and de-coarsening in multi-path change-point models*
Abstract: Follow-up studies on a group of units are commonly carried out to explore the possibility that a response distribution has changed at some unobservable time point. Often, in practice, this change-point model will include many potential covariates, which may not only be associated with the response distribution but also with the distribution of the unobservable change-points. Here, the covariates are allowed to enter the change-point distribution through a proportional odds model whose baseline odds is assumed to be piecewise constant as a function of time. The combination of a large number of putative regression coefficients in the response distributions as well as the change-point distribution, alone leads to a challenging simultaneous variable selection and estimation problem. Moreover, selection and estimation of the parameters that determine the coarseness of the baseline odds function adds a further level of complexity. Using penalized likelihood methods we are able to simultaneously perform variable selection and determine the coarseness of the baseline odds function. Our approach is computationally efficient and shown to be consistent in variable selection and parameter estimation. We assess its performance through simulations, and demonstrate its usage in fitting a model for cognitive decline in subjects with Alzheimer's disease. (Joint work with A. Shohoudi, D. Wolfson and M. Asgharian.)

Speaker: **Muñoz, M. Pilar** (Universitat Politècnica de Catalunya)
Title: *A Bayesian spatio-temporal framework for massive data*
Abstract: Many of the current problems today are related to stochastic processes that consider both temporal and spatial components, such as the processes related to the environment, econometrics, epidemiology, health sciences, and a long etcetera. The spatial components explain the relationships that occur between observations in an area that take place at fixed moments over time and, therefore, these spatial relationships change over time. This makes both estimation and predictions of these processes non-trivial. In addition,

the data set needed for estimating and predicting these processes correctly is very large, in general, when considering the computational problems that this entails. Among the statistical procedures that have contributed most to estimating these kinds of problems are the hierarchical Bayesian space time models (HBSTM). These procedures consist of starting with a complex structure and hierarchically flexibilizing this structure up to a basic structure within a Bayesian perspective. The number of articles published following this methodology has at this moment an exponential growth since the 1990s, due to the increase in CPU speed and RAM memory in affordable personal computers.

The aim of this work is to give the necessary recommendations to efficiently implement these models for massive space time data and give some tools in order to develop more complex models. Our implementation allows easy modification of the seasonalities, the spatial structure and the temporal autoregressive components. This framework has been programmed in R code and all the code has been implemented in the R package HBSTM `http://cran.r-project.org/web/packages/HBSTM/index.html`, and applied to the estimation of temperature data sets in the Strait of Gibraltar. (Joint work with Alberto López.)

Speaker: **Ma, Shuangge** (Yale University)
Title: *Robust identification of gene-environment interactions*
Abstract: Gene-environment interactions play an important role in disease development beyond the main effects of genetic and environmental factors. We pursue methods that can be robust to model misspecification or contamination in response variables. Penalization is adopted for regularized estimation and marker selection. For computational feasibility, a progressive approach is developed, which can significantly reduce computer time while generating more reliable results. Simulation and data analyses have been extensively conducted.

Speaker: **Michailidis, George** (University of Michigan)
Title: *Change-Point Estimation in High-Dimensional Markov Random Field Models*
Abstract: In this talk we discuss a change-point estimation problem in the context of high-dimensional Markov Random Field models. Change-points represent a key feature in many dynamically evolving network structures. The change-point estimate is obtained by maximizing a profile penalized pseudo-likelihood function under a sparsity assumption. We also derive a tight bound for the estimate, up to a logarithmic factor, even in settings where the number of possible edges in the network far exceeds the sample size. The performance of the proposed estimator is evaluated on synthetic data sets and is also used to explore voting patterns in the US Senate in the 1979–2012 period.

Speaker: **Nan, Bin** (University of Michigan)
Title: *Non-Asymptotic Oracle Inequalities for the High-Dimensional Cox Regression via Lasso*
Abstract: We consider finite sample properties of the regularized high-dimensional Cox regression via lasso. Existing literature focuses on linear models or generalized linear models with Lipschitz loss functions, where the empirical risk functions are the summations of independent and identically distributed (iid) losses. The summands in the negative log partial likelihood function for censored survival data, however, are neither iid nor Lipschitz. We first approximate the negative log partial likelihood function by a sum of iid non-Lipschitz terms, then derive the non-asymptotic oracle inequalities for the lasso penalized Cox regression using point-wise arguments to tackle the difficulties caused by lacking iid Lipschitz losses. (Joint work with Shengchun Kong.)

Speaker: **Nathoo, Farouk S.** (University of Victoria)
Title: *Statistical Modeling of Electromagnetic Neuroimaging Data*
Abstract: In this talk I will present my recent work involving the development of statistical methods for the analysis of MEG and EEG data: (I) A skew-t space-varying regression model for the spectral analysis of resting state brain activity; (II) A sparse functional linear model for solving the ill-posed neuroelectromagnetic inverse problem based on spatial spike-and-slab priors; (III) A high-dimensional state-space model for the combined analysis of MEG and EEG data. I will discuss model formulations and computational problems, along with solutions based on mean field approximations and hybrid variational Bayes/MCMC algorithms.

Speaker: **Shojaie, Ali** (University of Washington)
Title: *Inference in high dimensions with the penalized score test*
Abstract: In recent years, there has been considerable theoretical development regarding variable selection consistency using penalized regression techniques, such as the lasso. However, there has been relatively little work on quantifying the uncertainty in these selection procedures. In this paper, we propose a new method for inference in high dimensions using a score test based on penalized regression. In this test, we perform penalized regression of an outcome on all but a single feature, and test for correlation of the residuals with the held-out feature. This procedure is applied to each feature in turn. Interestingly, when an l_1 penalty is used, the sparsity pattern of the lasso corresponds exactly to a decision based on the proposed test. Further, when an l_2 penalty is used, the test corresponds precisely to a score test in a mixed effects model, in which the effects of all but one feature are assumed to be random. We formulate the hypothesis being tested, as a compromise between the null hypotheses tested in simple linear regression on each feature, and in multiple linear regression on all features, and develop reference distributions for some well-known penalties. We also examine the behavior of the test on real and simulated data. (Joint work with Daniela Witten.)

Speaker: **Song, Peter** (University of Michigan)
Title: *Composite Likelihood Inference GeoCopula models for high-dimensional spatial-clustered data*
Abstract: Spatial-clustered data refer to high-dimensional correlated measurements collected from units or subjects that are spatially clustered. Such data arise frequently from studies in social and health sciences. We propose a unified modeling framework, termed as GeoCopula, to characterize both large-scale variation and small-scale variation for various data types, including continuous data, binary data and count data as special cases. To overcome challenges in the estimation and inference for the model parameters, we propose an efficient composite likelihood approach in that the estimation efficiency is resulted from a construction of over-identified joint composite estimating equations. Consequently the statistical theory for the proposed estimation is developed by extending the classical theory of Nobel Laureate Hansen's generalized methods of moments (GMM). A clear advantage of the proposed estimation method is the computation feasibility. We present several simulation studies to assess the performance of the proposed models and estimation methods for both Gaussian and binary spatial-clustered data. Results show a clear improvement on estimation efficiency over the conventional composite likelihood method. An illustrative data example is included to motivate and demonstrate the proposed method. (Joint work with Drs. Yun Bai and Jian Kang.)

Speaker: **Stephens, David A.** (McGill University)
Title: *Bayesian reconstruction of metabolomic profiles from NMR spectra*
Abstract: Nuclear Magnetic Resonance (NMR) spectra are widely used in metabolomics to obtain profiles of metabolites dissolved in biofluids. Methods for estimating metabolite concentrations from these spectra are presently confined to manual peak fitting and to binning procedures for integrating resonance peaks. Extensive information on the patterns of spectral resonance generated by human metabolites is now available in online databases. By incorporating this information into a Bayesian model we can deconvolve resonance peaks from a spectrum and obtain explicit concentration estimates for the corresponding

metabolites. Spectral resonances that cannot be deconvolved in this way may also be of scientific interest, so we model them jointly using wavelets.

I will describe via a Markov chain Monte Carlo algorithm which allows us to sample from the joint posterior distribution of the model parameters, using specifically designed block updates to improve mixing. The strong prior on resonance patterns allows the algorithm to identify peaks corresponding to particular metabolites automatically, eliminating the need for manual peak assignment.

Speaker: **Taddy, Matt** (University of Chicago)
Title: *The Gamma Lasso*
Abstract: The gamma lasso is a very fast regression algorithm for obtaining continuous regularization paths corresponding to cost functions spanning the range of concavity between L_0 and L_1 norms. This will serve as a foil for survey of the large literature on sparse penalized deviance estimation. We'll describe the essential role of path stability in predictive performance, computation cost, and in ability to use information-criteria for model selection; the trade-off between support recovery (or low false discovery) and stability; Bayesian interpretation of the methods; and practical implications for Big Data applications in prediction and inference. I'll use estimation of individual contributions from NHL hockey players as a running illustrative example.

Speaker: **Verducci, Joseph** (The Ohio State University)
Title: *Nonparametric Methods for Detecting Association that is Overly Concentrated in a Subpopulation*
Abstract: We begin with the Frank Copula on a pair (X, Y) of variables and show that, for a large fixed sample size n, given the ranking of the X values, the ranking of the Y values is well approximated by a Mallows ranking model. Moreover, the empirical distribution of the absolute ranked differences $|\text{rank}(X) - \text{rank}(Y)|$ is well approximated as a beta distribution. The approximations hold especially well for copulas with relatively low overall association. This is the context where detection of over-concentration in a subsample is feasible because the empirical distribution of the absolute rank differences may initially lie distinctly above the theoretical distribution under a null hypothesis of homogeneous association. Some robustness issues are checked by simulations under a Gaussian copula. An alternative approach begins with the same null hypothesis, that $\text{rank}(Y)$ follows a Mallows ranking model centered at $\text{rank}(X)$, and nests this into a non-homogeneous alternative that is simple to estimate and to test against. The behaviors of the two methods are compared.

Speaker: **Vidyashankar, Anand N.** (George Mason University)
Title: *Supervised Factor Analysis with High-dimensional Predictors*
Abstract: In a variety of scientific applications, it is desired to identify associations between measurements and high-dimensional correlated features. While confirmatory factor analysis is a useful statistical method in these contexts, presence of high-dimensional features complicates data analysis. Additionally, it is critical to provide standard errors and construct confidence intervals for the parameters of the proposed model after accounting for model selection uncertainty. In this presentation we propose a new methodology, a variant of the screen and clean method of Wasserman and Roeder, for factor modeling and post-model selection inference. We establish consistency, asymptotic normality, and the oracle property of the estimators of the proposed factor model. We illustrate our results using simulations and a variety of examples arising in biological and financial applications.

Speaker: **Volodin, Andrei** (George Mason University)
Title: *Confidence sets based on the positive part James-Stein estimator with the asymptotically constant coverage probability*
Abstract: The asymptotic expansions for the coverage probability of a confidence set centered at the James-Stein estimator presented in our previous publications, show that this probability depends on the noncentrality parameter τ^2 (the sum of the squares of the means of normal distributions). In this talk we

discuss how these expansions can be used for a construction of confidence region with constant confidence level, which is asymptotically (the same formula for both case $\tau \to 0$ and $\tau \to \infty$) equal to some fixed value $1 - \alpha$. We establish the shrinkage rate for the confidence region according to the growth of the dimension p and also the value of τ for which we observe quick decreasing of the coverage probability to the nominal level $1 - \alpha$. When $p \to \infty$ this value of τ increases as $O(p^{1/4})$. The accuracy of the results obtained is shown by the Monte-Carlo statistical simulations. (Joint work with S. Ejaz Ahmed.)

Speaker: **Yi, Grace Y.** (University of Waterloo)
Title: *Variable Selection and Inference Procedures for Marginal Analysis of Longitudinal?Data with Missing Observations or Measurement Error*
Abstract: In contrast to extensive attention on model selection for univariate data, research on correlated data remains relatively limited. Furthermore, in the presence of missing data and/or measurement error, standard methods would typically break down. To address these issues, we propose marginal methods that simultaneously carry out model selection and estimation for longitudinal data analysis. Our methods have a number of appealing features: the applicability is broad because the methods are developed for a unified framework with marginal generalized linear models; model assumptions are minimal in that no full distribution is required for the response process and the distribution of the mis-measured covariates is left unspecified; and the implementation is straightforward. To justify the proposed methods, we provide both theoretical properties and numerical assessments. (Joint work of Xianming Tan and Runze Li.)