WHOA-PSI 2016

Workshop on Higher-Order Asymptotics and Post-Selection Inference

Washington University in St. Louis, St. Louis, Missouri, USA

30 September - 2 October, 2016

Schedule of Talks, Abstracts

Organizers:

John Kolassa, Todd Kuffner, Nancy Reid, Ryan Tibshirani, Alastair Young

Friday 30th September

7:30 – 9:00 Breakfast and registration

9:00 – 9:15 Introductions

9:15 – 10:05 Tutorial on *Post-Selection Inference* (Todd Kuffner)

10:05 – 10:25 Coffee break

10:25 – 11:15 Tutorial on *Higher-Order Asymptotics* (Todd Kuffner)

11:30 – 1:00 Lunch and registration

1:00 – 1:10 Opening remarks

1:10 – 2:50 **Session 1**; Chair: Nan Lin, Washington University in St. Louis

- **1:10 – 1:35** Ryan Martin, North Carolina State University
  A new double empirical Bayes approach for high-dimensional problems
- **1:35 – 2:00** Anru Zhang, University of Wisconsin
  Cross: efficient low-rank tensor completion
- **2:00 – 2:25** Shujie Ma, UC Riverside
  Wild bootstrap confidence intervals in sparse high dimensional heteroscedastic linear models
- **2:25 – 2:50** Discussion

2:50 – 3:05 Coffee break

3:05 – 4:15 **Session 2**; Chair: Debraj Das, North Carolina State University

- **3:05 – 3:30** Xiaoying Tian, Stanford University
  Selective inference with a randomized response
- **3:30 – 3:55** Yuekai Sun, University of Michigan
  Fast convergence of Newton-type methods on high dimensional problems
- **3:55 – 4:15** Discussion

4:15 – 4:30 Coffee break

4:30 – 5:40 **Session 3**; Chair: Sangwon Hyun, Carnegie Mellon University

- **4:30 – 4:55** Hongyuan Cao, University of Missouri Columbia
  Change point estimation: another look at multiple testing problems
- **4:55 – 5:20** Jelena Bradic, UC San Diego
  Inference in Non-Sparse High-Dimensional Models: going beyond sparsity and de-biasing.
- **5:20 – 5:40** Discussion

5:40 – 6:40 Pub in the conference center; dinner on your own (suggestions in other document)

Saturday 1st October

6:30 – 8:30 Breakfast and registration

8:30 – 9:50 Session 4; Chair: Ryan Tibshirani, Carnegie Mellon University

8:30 – 9:00 Soumendra Lahiri, North Carolina State University
Higher order asymptotic properties of the Bootstrap in post model selection inference

9:00 – 9:30 Larry Brown, University of Pennsylvania
Mallows C_p for Realistic Out-of-sample Prediction

9:30 – 9:50 Discussion

9:50 – 10:10 Coffee break

10:10 – 11:30 Session 5; Chair: Alastair Young, Imperial College London

10:10 – 10:40 Florentina Bunea, Cornell University
Model Based Variable Clustering: PECOK vs CORD

10:40 – 11:10 John Robinson, University of Sydney
Nonparametric Tests for Multi-parameter M-estimators

11:10 – 11:30 Discussion

11:30 – 1:00 Lunch

1:00 – 2:30 Session 6; Chair: Alessandra Salvan, University of Padova

1:00 – 1:40 Richard Samworth, University of Cambridge
Efficient multivariate entropy estimation with hints of an application to testing shape constraints

1:40 – 2:10 Peter McCullagh, University of Chicago
Asymptotic inference for sparse priors

2:10 – 2:30 Discussion

2:30 – 2:50 Coffee break

2:50 – 4:10 Session 7; Chair: Stefan Wager, Columbia University

2:50 – 3:20 Annie Qu, University of Illinois Urbana-Champaign
Weak Signal Identification and Inference in Penalized Model Selection

3:20 – 3:50 Don Fraser, University of Toronto
How Saddlepoint Replaced Sufficiency and Likelihood, and Went Beyond

3:50 – 4:10 Discussion

4:10 – 4:30 Coffee break

4:30 – 5:50 Session 8; Chair: Jacob Bien, Cornell University

4:30 – 5:00 Rob Tibshirani, Stanford University
Recent Advances in Post-Selection Statistical Inference

5:00 – 5:30 Xiao-Li Meng, Harvard University
Building a Statistical Theory for Individualized Treatments: A Multi-resolution Perspective

5:30 – 5:50 Discussion

6:30 – 9:30 Banquet

Sunday 2nd October

6:30 – 8:30 Breakfast and registration

8:30 – 10:20 Session 8; Chair: Aaditya Ramdas, UC Berkeley

8:30 – 9:00 Jianqing Fan, Princeton University
Guarding against Spurious Discoveries in High Dimension

9:00 – 9:30 Nicola Sartori, University of Padova
Median bias reduction of maximum likelihood estimates

9:30 – 9:55 Dalia Ghanem, UC Davis
Post-Selection Inference with High-Frequency Time Series Predictors

9:55 – 10:20 Discussion

10:20 – 10:35 Coffee break

10:35 – 11:30 Session 9; Chair: Kai Zhang, University of North Carolina Chapel Hill

10:35 – 11:20 Daniel Yekutieli, Tel Aviv University
From post-hoc analysis to post-selection inference

11:20 – 11:30 Discussion

11:30 – 1:00 Lunch and poster session; Chair: Jessie Jeng, North Carolina State University

1:00 – 2:50 Session 10; Chair: Nancy Reid, University of Toronto, and Qi Wang, Washington University in St. Louis

1:00 – 1:30 Xihong Lin, Harvard University
Hypothesis Testing for Dense and Sparse Signal Detection with Applications in Whole-Genome Array and Sequencing Studies

1:30 – 1:55 Genevera Allen, Rice University
Population Inference post Model Selection in Neuroscience

1:55 – 2:25 Andreas Buja, University of Pennsylvania
From Post-Selection to Misspecification

2:25 – 2:50 Discussion

2:50 – 3:30 Coffee break

3:30 – 5:30 Blue-Sky Session

Genevera Allen, Rice University

Title: Population Inference post Model Selection in Neuroscience

Abstract: *Analyzing large multi-subject neuroimaging or multiple neural recording studies leads to a new type of inferential problem that we term, Population Inference post Model Selection (PIMS). In this type of problem, a model selection procedure is used to summarize images or recordings from each subject in the study; statistical tests are then used to compare the subject-level summaries across the population. Consider an example of inference for functional brain connectivity where neuroscientists seek to understand how the brain communicates at a systems level and how these systems are disrupted in neurological conditions and diseases. Typically this is accomplished by estimating brain networks for each subject in the study and conducting inference across the subjects to find network metrics that are different between groups or are associated with symptom severity. This approach, however, ignores the fact that subject-level networks are estimated via a model selection procedure; hence, this is a PIMS problem. In this talk, we present the PIMS problem, discuss how this arises in many areas of neuroscience, and carefully study the specific example of population inference for functional neural connectivity. For the latter, we show that failure to account for the model selection stage leads to a dramatic increase in false positives and reduced statistical power for population inference. Finally, we introduce an estimation and inferential procedure that uses bootstrapping and multi-level models to improve the problem, demonstrating this through simulations and an fMRI study on autism spectrum disorder.*

Jelena Bradic, UC San Diego

Title: Inference in Non-Sparse High-Dimensional Models: going beyond sparsity and debiasing.

Abstract: *Hypothesis testing in models whose dimension far exceeds the sample size is of fundamental importance in both statistical theory and applications such as biological and social sciences. The state-of-the-art methods heavily rely on the assumption of sparsity of the model parameters. However, sparsity assumption, despite its popularity, is largely violated in many of the contemporary scientific fields. In this paper, we address the essential question of interest whether it is possible to conduct a valid inference without directly requiring sparsity of the model. We develop a novel test, named CorrT, that provides valid inference as a result of its new, decoupling structure utilizing explicitly the null hypothesis. CorrT does not make any assumptions on the sparsity of the model coefficients under the common regularity condition of row-wise sparse precision matrix of the features. Additionally, we develop an extension of the CorrT test for the purpose of simultaneous testing. In a class of non-sparse models we show that the developed test is asymptotically valid. The result is built upon the novel asymptotic theory of the approximate multiplier bootstrap, which may be of independent interest. We allow both the number of tests and the effective dimension of model parameters to be larger than the sample size. Moreover, we show that the proposed test asymptotically achieves minimax optimality in the case of strictly sparse models. Furthermore, CorrT applies to a wide variety of non-linear high-dimensional models that are not covered by the existing literature – including single-index models and Heckman selection models. Empirically,*

the proposed test outperforms the simple and robust debiasing estimator and the score estimator in terms of achieving accurate inference on simulated data. Joint work with a PhD Student Yinchu Zhu.

Lawrence Brown, University of Pennsylvania

Title: Mallows C_p for Realistic Out-of-sample Prediction

Abstract: *(Joint work with A. Buja, R. Berk, A. Kuchibotla and L. Zhao) Mallows' C_p is a frequently used tool for variable selection in linear models. (For the original discussion see Mallows (1973), building on Mallows (1964, 1966).) In practice it may be used in conjunction with forward stepwise selection or all-subsets selection, or some other selection scheme. It can be derived and interpreted as an estimate of (normalized) predictive squared error in a very special situation. Two key features of that situation are: 1) The observed covariate variables and the covariates for the predictive population are, "not to be regarded as being sampled randomly from some population, but rather are taken as fixed design variables". (Mallows (1973).); and 2) The observations in the sample and in the predictive universe follow a homoscedastic linear model. Assumption 1) does not accord with most of the common statistical settings in which C_p is employed, and assumption 2) is very frequently undesirably optimistic in practical settings.*

We derive an easily computed variant of Mallows' expression that does not rely on either of these assumptions. The new variant, denoted as $C_p^{\oplus}$, estimates the predictive squared error when the best linear estimator with the currently selected variables is used for future observations drawn from the same population. The formulation is "assumption lean" in that there are virtually no assumptions on the true sampling distribution.

Use of this variant will be demonstrated via simulations in a simple regression setting that enables easy visualization and also exact computation of some relevant quantities. For a more practical demonstration we also apply the methodology to variable selection in a data set involving criminal sentencing.

Andreas Buja, University of Pennsylvania

Title: From Post-Selection to Misspecification

Abstract: *(Joint work with Richard Berk, Lawrence Brown, Mikhail Traskin, Kai Zhang, Emil Pitkin, Linda Zhao, and Ed George) We recount briefly the PoSI approach to post-selection inference, which consists essentially of a reduction of the problem to simultaneous inference. Resulting inference is conditional on the observed regressors and it is valid under first order misspecification but assumes homoskedasticity. Dissatisfied with making assumptions, we develop a framework of inference for regression that allows degrees of misspecification as the normal situation in data analysis. The framework makes no other assumptions than iid sampling of observations, and it treats the regressors as random. Parameters are reinterpreted as statistical functionals, and a notion of well-specification is proposed that describes ideal circumstances under which a statistical functional describes aspects of the conditional distribution of the response alone, irrespective of the regressor distribution. The framework shows that, under misspecification, the treatment of regressors as fixed and the ancillarity argument that justifies it are wrong. Asymptotically valid inference can be based either on the x-y bootstrap that*

resamples observations, or on sandwich estimators of standard error. It can be shown that in a sense sandwich estimators are a limiting case of bootstrap estimators.

Florentina Bunea, Cornell University

Title: Model Based Variable Clustering: PECOK vs CORD

Abstract: *The problem of variable clustering is that of grouping similar components of a p-dimensional vector $X = (X_1, \ldots, X_p)$, and estimating these groups from n independent copies of X. Traditionally, variable clustering has been treated in an algorithmic manner, making the estimated clusters difficult to interpret and analyze, from a statistical perspective. We take a different approach in this talk, and suggest model based variable clustering.*

For a partition G of the index set $\{1, \ldots, p\}$, we consider the class of G-latent models, in which each group of the X-variables is assumed to have a common latent generator. At first sight, the most natural way to estimate such clusters is via K-means. We explain why this strategy cannot lead to perfect cluster recovery in G-latent models. We offer a correction, based on semi-definite programing, that can be viewed as a penalized convex relaxation of K-means (PECOK). We introduce a cluster separation measure tailored to G-latent models, which can be viewed as a measure of the signal in these models. We derive its minimax lower bound for perfect cluster recovery. The clusters estimated by PECOK are shown to recover G at a near minimax optimal cluster separation rate, a result that holds true even if K, the number of clusters, is estimated adaptively from the data. We also compare PECOK with appropriate corrections of spectral clustering-type procedures, and show that the former outperforms the latter for perfect cluster recovery of minimally separated clusters.

We also introduce a more general class of models for clustering, that of G-block correlation matrix models. We explain when this class can offer more flexibility than the class of G-latent models. We identify the appropriate cluster separation metric in these models, different than the one above, and derive its minimax lower bound for cluster recovery. We derive a new clustering method, CORD, tailored to the class of G-block correlation models.

Extensions to overlapping clustering and inference in G-graphical models will be discussed briefly, time permitting.

Hongyuan Cao, University of Missouri Columbia

Title: Change point estimation: another look at multiple testing problems

Abstract: *We consider the problem of large scale multiple testing for data that have locally clustered signals. With this structure, we apply techniques from change point analysis and propose a boundary detection algorithm so that the local clustering information can be utilized. We show that by exploiting the local structure, the precision of a multiple testing procedure can be improved substantially. We study tests with independent as well as dependent p-values. Monte Carlo simulations suggest that the methods perform well with realistic sample sizes and demonstrate the improved detection ability compared with competing methods. The practical utility of our methods is illustrated from a genome-wide association study of blood lipids.*

Jianqing Fan, Princeton University

Title: Guarding against Spurious Discoveries in High Dimension

Abstract: *(Joint work with Wenxinz Zhou) Many data-mining and statistical machine learning algorithms have been developed to select a subset of covariates to associate with a response variable. Spurious discoveries can easily arise in high-dimensional data analysis due to enormous possibilities of such selections. How can we know statistically our discoveries better than those by chance? In this paper, we define a measure of goodness of spurious fit, which shows how good a response variable can be fitted by an optimally selected subset of covariates under the null model, and propose a simple and effective LAMM algorithm to compute it. It coincides with the maximum spurious correlation for linear models and can be regarded as a generalized maximum spurious correlation. We derive the asymptotic distribution of such goodness of spurious fit for generalized linear models and L_1-regression. Such an asymptotic distribution depends on the sample size, ambient dimension, the number of variables used in the fit, and the covariance information. It can be consistently estimated by multiplier bootstrapping and used as a benchmark to guard against spurious discoveries. It can also be applied to model selection, which considers only candidate models with goodness of fits better than those by spurious fits. The theory and method are convincingly illustrated by simulated examples and an application to the binary outcomes from German Neuroblastoma Trials.*

Don Fraser, University of Toronto

Title: How Saddlepoint Replaced Sufficiency and Likelihood, and Went Beyond

Abstract: *p-values are in the public eye, and one sees mostly turmoil; and the American Statistical Association has recently scolded their misuse but offered little to go beyond Fisher's "probability of as far or farther from expectation". Saddlepoint methods entered statistics rather slowly: Henry Daniels in 1954 then Barndorff-Nielsen and Cox 1979, just 25 years! But more recently the saddlepoint methods have radically changed the landscape for core methods of inference; and p-values no longer need to be in the wild west stage. We offer a core view of p-value and outline an integrity for evolved statistical inference.*

Dalia Ghanem, UC Davis

Title: Post-Selection Inference with High-Frequency Time Series Predictors

Abstract: *Consider a data set composed of an outcome variable which is observed annually, and some predictor variables which are observed hourly or daily. Variations of this scenario are fairly common in the environmental and atmospheric sciences, in economics, and in finance. In such empirical settings, there are many different models that a practitioner might select between, and hence our goal is to propose a route to valid post-selection inference in this setup. Specifically, we seek valid post-selection inference in the conditional mean regression model for the outcome variable. We consider two approaches to this problem: (i) utilizing a sequential procedure to select a suitable dimension reduction for the predictor time series, and (ii) a hierarchical approach beginning with selecting among competing time series models for the predictor time series. We discuss an application to modeling US county-level mortality rates as a function of daily temperature and precipitation, using real data. Joint work with Todd Kuffner.*

Soumendra Lahiri, North Carolina State University

Title: Higher order asymptotic properties of the Bootstrap in post model selection inference

Abstract: *(Joint work with Arindam Chatterjee and Debraj Das) Chatterjee and Lahiri (2013) showed that under suitable conditions, the residual Bootstrap is second order correct for studentized pivots based on the ALASSO. One of the major limitations of their result is the existence of a preliminary estimator satisfying certain probabilistic bounds that are hard to verify in the $p > n$ case. In this talk, we show that the second order correctness property holds quite generally for a number of penalized regression methods satisfying a version of the Oracle property of Fan and Li (2001). In particular, we show that under some suitable conditions, the LASSO and some popular nonconvex penalization functions including the SCAD and the MCP also enjoy second order correctness.*

Xihong Lin, Harvard University

Title: Hypothesis Testing for Dense and Sparse Signal Detection with Applications in Whole-Genome Array and Sequencing Studies

Abstract: *Massive genetic and genomic data present many exciting opportunities as well as challenges in data analysis and result interpretation, e.g., how to develop effective strategies for signal detection using massive genetic and genomic data when signals are weak and sparse. Many variable selection methods have been developed for analysis of high-dimensional data in the statistical literature. However limited work has been done on statistical inference for massive data. In this talk, I will discuss hypothesis testing for analysis of high-dimensional data motivated by gene, pathway/network based analysis in whole-genome array and sequencing studies. I will focus on signal detection when signals are weak and sparse, which is the case in genetic and genomic association studies. I will discuss hypothesis testing for signal detection using penalized likelihood based methods, and aggregated marginal test statistics based method using the Generalized Higher Criticism (GHC) and Berk-Jones tests. The results are illustrated using data from genome-wide association studies.*

Shujie Ma, UC Riverside

Title: Wild bootstrap confidence intervals in sparse high dimensional heteroscedastic linear models

Abstract: *In recent years, statistical inference for high dimensional regression models has received considerable attention. In this talk, I will introduce a wild bootstrap (WB) method we have proposed for constructing confidence intervals in sparse high dimensional heteroscedastic linear models. Specifically, we estimate the parameters by a partial concave penalized method. This method avoids shrinking the estimate of the parameter of interest to exactly zero. We show that the proposed estimator follows an asymptotic normal distribution. We further develop a heteroscedasticity estimator for the covariance matrix, based on which we define a heteroscedasticity-robust pivotal statistic for the parameter of interest. Then we propose a wild bootstrap procedure, and show that it provides valid approximation to the distribution of the pivotal statistic. Simulation studies are performed to investigate the finite sample performance of the proposed bootstrap method. It is also illustrated by a real data example.*

Ryan Martin, North Carolina State University

Title: A new double empirical Bayes approach for high-dimensional problems

Abstract: *In high-dimensional problems, selecting a good prior—one that leads to a posterior with optimal concentration properties and efficient computation—can be a challenge for Bayesians. In this talk I will present a new kind of empirical Bayes that uses data in the prior in two ways: first, the prior is suitably centered on the data, and second, a regularization step is taken to prevent the greedy centering from driving the behavior of the posterior. In the context of a sparse high-dimensional linear model, a variety of posterior concentration results will be presented, along with simulation results that demonstrate the method's quality performance. Extensions to other high-dimensional models, as well as nonparametric problems, will also be discussed.*

Peter McCullagh, University of Chicago

Title: Asymptotic inference for sparse priors

Abstract: *(Joint work with N. Polson) The goal of this work is to develop an asymptotic approximation for the posterior distribution of the signal X for a sample of size one in an additive Gaussian model with known variance. Sparseness is defined by the limiting prior exceedance rate $\nu h(x)$ for fixed thresholds $x > 0$. The posterior approximation holds for a wide range of sparse prior distributions in an asymptotic setting where the sparseness parameter ν is small.*

Xiao-Li Meng, Harvard University

Title: Building a Statistical Theory for Individualized Treatments: A Multi-resolution Perspective

Abstract: *Personalized treatment sounds heavenly, but where on Earth did they find the right guinea pig for me? What data are relevant when making a treatment decision for me? What replications are relevant for quantifying the uncertainty of this personalized decision? What does ?relevant? even mean here? The multi-resolution (MR) perspective from the wavelets literature provides a convenient theoretical framework for contemplating such questions. Within the MR framework, signal and noise are two sides of the same coin: variation. They differ only in the resolution of that variation?a threshold, the primary resolution, divides them. We use observed variations at or below the primary resolution (signal) to estimate a model and those above the primary resolution (noise) to estimate our uncertainty. The search for the appropriate primary resolution is a quest for the age old bias-variance trade-off: estimating more precisely a less relevant treatment decision versus estimating less precisely a more relevant one. In this paper, we investigate this trade-off by considering the problem of predicting an individualized outcome using potentially infinitely many covariates, where the prediction performance, regardless of the specific criterion used, depends critically on the number of covariates used, a choice of the primary resolution. We decompose the total prediction loss into three components: (1) the misspecification loss, depending purely on the functional forms used for prediction; (2) the approximation loss, decaying to zero with the increase of the primary resolution; and (3) the estimation loss, depending on both*

the available data and the primary resolution level. We demonstrate how the choice and optimal rate of the primary resolution depends on the decaying rates of the approximation and estimation losses. We also suggest a TIC-like criterion for selecting the primary resolution in practice, taking into account the criterion for evaluating the prediction performance. (This is a joint work with Xinran Li)

Annie Qu, University of Illinois Urbana-Champaign

Title: Weak Signal Identification and Inference in Penalized Model Selection

Abstract: *Weak signal identification and inference are very important in the area of penalized model selection, yet they are under-developed and not well-studied. Existing inference procedures for penalized estimators are mainly focused on strong signals. In this paper, we propose an identification procedure for weak signals in finite samples, and provide a transition phase in-between noise and strong signal strengths. We also introduce a new two-step inferential method to construct better confidence intervals for the identified weak signals. Our theory development assumes that variables are orthogonally designed. Both theory and numerical studies indicate that the proposed method leads to better confidence coverage for weak signals, compared with those using asymptotic inference. In addition, the proposed method outperforms the perturbation and bootstrap resampling approaches. We illustrate our method for HIV antiretroviral drug susceptibility data to identify genetic mutations associated with HIV drug resistance. This is joint work with Peibei Shi.*

John Robinson, University of Sydney

Title: Nonparametric Tests for Multi-parameter M-estimators

Abstract: *Tests of hypotheses concerning subsets of multivariate means or coefficients in linear or generalized linear models depend on parametric assumptions which may not hold. One nonparametric approach to these problems uses the standard nonparametric bootstrap using the test statistics derived from some parametric model but basing inferences on bootstrap approximations. We derive different test statistics based on empirical exponential families and use a tilted bootstrap to give inferences.The bootstrap approximations can be accurately approximated to relative second order accuracy by a saddlepoint approximation. This generalises earlier work in two ways. First, we generalise from bootstraps based on resampling vectors of both response and explanatory variables to include bootstrapping residuals for fixed explanatory variables, and second, we obtain a theorem for tail probabilities under weak conditions justifying approximation to bootstrap results for both cases.*

Richard Samworth, University of Cambridge

Title: Efficient multivariate entropy estimation with hints of an application to testing shape constraints

Abstract: *Many statistical procedures, including goodness-of-fit tests and methods for independent component analysis, rely critically on the estimation of the entropy of a distribution. In this talk, we seek entropy estimators that are efficient in the sense of achieving the local asymptotic minimax lower bound. To this end, we initially study a generalisation*

of the estimator originally proposed by Kozachenko and Leonenko (1987), based on the k-nearest neighbour distances of a sample of n independent and identically distributed random vectors in $\mathbb{R}^d$. When $d \leq 3$ and provided $k/log^5 n \to \infty$ (as well as other regularity conditions), we show that the estimator is efficient; on the other hand, when $d \geq 4$, a non-trivial bias precludes its efficiency regardless of the choice of k. This motivates us to consider a new entropy estimator, formed as a weighted average of Kozachenko–Leonenko estimators for different values of k. A careful choice of weights enables us to obtain an efficient estimator in arbitrary dimensions, given sufficient smoothness. We conclude with hints of how these results can be used to propose a new test of log-concavity in low dimensions.

Nicola Sartori, University of Padova

Title: Median bias reduction of maximum likelihood estimates

Abstract: *(joint with E. C. Kenne Pagui and A. Salvan) For regular parametric problems, we show how median centering of the maximum likelihood estimate can be achieved by a simple modification of the score equation. For a scalar parameter of interest, the estimator is second-order median unbiased and equivariant under interest respecting reparameterizations. With a vector parameter of interest, componentwise equivariance and second-order median centering are obtained. The new method does not depend on the existence of the maximum likelihood estimate and is effective in preventing infinite estimates, like Firths (1993, Biometrika) implicit method for bias reduction. Simulation results for continuous and discrete models, including binary regression, confirm that the method succeeds in solving the infinite estimate problem and in achieving componentwise median centering, while keeping comparable dispersion and the same approximate distribution as its main competitors.*

Yuekai Sun, University of Michigan

Title: Fast convergence of Newton-type methods on high dimensional problems

Abstract: *We study the convergence rate of Newton-type methods on high-dimensional problems. The high-dimensional nature of the problem precludes the usual global strong convexity and smoothness that underlie the classical analysis of such methods. We find that restricted version of these conditions which typically arise in the study of the statistical properties of the solutions are also enough to ensure good computational properties of Newton-type methods. We explore the algorithmic consequences in distributed and streaming settings.*

Xiaoying Tian, Stanford University

Title: Selective inference with a randomized response

Abstract: *Inspired by sample splitting and the reusable holdout introduced in the field of differential privacy, we consider selective inference with a randomized response. Using a randomized response can ensure that the leftover information of Fithian et al. (2014) is bounded below ensuring that selective intervals are better behaved than without randomization. Under independent sampling, we prove a selective (or privatized) central*

limit theorem that transfers procedures valid under asymptotic normality without selection to their corresponding selective counterparts. This allows selective inference in the nonparametric settings. Finally, we describe a method for selective inference following cross-validation using slightly more randomization than the split into groups of standard cross-validation. We focus on the classical asymptotic setting, leaving the interesting high-dimensional asymptotic questions for future work.

Rob Tibshirani, Stanford University

Title: Recent Advances in Post-Selection Statistical Inference

Abstract:

Daniel Yekutieli, Tel Aviv University

Title: From post-hoc analysis to post-selection inference

Abstract: *I will give an introductory talk explaining the connection between the work of Tukey and Scheffe on post-hoc analysis, Benjamini and Hochberg's work on the FDR, the work of Efron and colleagues on the Bayesian FDR, my work with Benjamini on selective inference, the work of Berk et al. on post-selection inference, and recent work on frequentist and Bayesian post-selection inferences based on the conditional likelihood.*

Anru Zhang, University of Wisconsin

Title: Cross: efficient low-rank tensor completion

Abstract: *Tensors, or high-order arrays, attract significant attention in recent research. Current literature on tensor completion primarily focuses on recovery via a number of uniformly sampled entries, which is unclear whether the required sample size can be further reduced. In this article, we propose a framework for low-rank tensor completion via a novel cross tensor measurement scheme. The proposed procedure is efficient and easy to implement. In particular, we show that a third order tensor of rank-(r_1, r_2, r_3) in $p_1 \times p_2 \times p_3$ dimensional space can be recovered from as few as $O(r_1 r_2 r_3 + r_1(p_1 - r_1) + r_2(p_2 - r_2) + r_3(p_3 - r_3))$ measurements, which matches the sample complexity lower-bound. In the noisy case, we also develop theoretical risk upper bound and the matching minimax lower bound over certain class of low-rank tensors for the proposed procedure. The results can be further extended for fourth or higher-order tensors. Simulation studies show that the method perform well under a variety of settings. Finally, the procedure is illustrated through a real data example in neuroimaging.*

Poster Presentations

The poster session is Sunday, 2nd October 11:30am-1:00pm

Debraj Das, North Carolina State University

Title: Perturbation Bootstrap in Adaptive LASSO

Abstract: *(Joint work with Soumendra Lahiri) The Adaptive LASSO (ALASSO) was proposed by Zou [J. Amer. Statist. Assoc. 101 (2006) 1418-1429] as a modification of the LASSO for the purpose of simultaneous variable selection and estimation of the parameters in a linear regression model. Zou (2006) established that the ALASSO estimator is variable-selection consistent as well as asymptotically Normal in the indices corresponding to the nonzero regression coefficients in certain fixed-dimensional settings. Minnier, Tian and Cai [J. Amer. Statist. Assoc. 106 (2011) 1371-1382] proposed a perturbation bootstrap method to approximate the distribution of the ALASSO estimator in the fixed-dimensional setting. In this paper, we show that this (naive) perturbation bootstrap fails to provide a consistent approximation to the distribution of the ALASSO estimator when the dimension of the regression parameter is reasonably large. We propose a modified perturbation bootstrap and establish its distributional consistency even when the dimension of the model is allowed to grow with the sample size in a substantially large rate. Moreover, we show that a suitably studentized version of our modified perturbation bootstrap ALASSO estimator achieves second-order correctness even in high dimension. As a consequence, inferences based on the modified perturbation bootstrap will be more accurate than the inferences based on the oracle Normal approximation.*

Guanshengrui Hao, Washington University in St. Louis

Sangwon Hyun, Carnegie Mellon University

Title: Exact Post-Selection Inference for Changepoint Detection and Other Generalized Lasso Problems

Abstract: *We present tools for inference conditioned on model selection events that are defined by the generalized lasso regularization path. We develop exact hypothesis tests and confidence intervals for linear contrasts of the underlying mean vector, conditioned on any model selection event along the generalized lasso path (assuming Gaussian errors in the observations). By inspecting specific choices of D, we obtain post-selection tests and confidence intervals for specific cases of generalized lasso estimates, such as the fused lasso, trend filtering, and the graph fused lasso. In addition, some practical aspects of our methods such as valid post-processing of generalized estimates before performing inference in order to improve power, and problem-specific visualization aids that may be given to the data analyst for he/she to choose linear contrasts to be tested.*

Jelena Markovic, Stanford University

Title: Bootstrap after model selection

Abstract: *(Joint work with Jonathan Taylor) Recently, [Tian and Taylor, 2015] developed a selective inference approach with the randomized response. They constructed an asymptotically pivotal test statistic that allows for the selective inference in non-parametric settings. In this work, under a regularity condtion on the randomization distribution, we relax their assumptions, notably the local alternatives; furthermore, we propose a bootstrap version of this test statistic. We prove that the bootstrap test statistic is also asymptotically pivotal in the uniform sense across a family of non-parametric distributions. We also describe a way of using the wild bootstrap and projected Langevin Monte Carlo method to compute the bootstrapped test statistic. To illustrate the computations, we consider several examples in which the model selection has been done using algorithms such as randomized Lasso, multiple steps forward-stepwise and marginal screening all with random design matrix.*

Snigdha Panigrahi, Stanford University

Title: Bayesian post-selection inference in the linear model

Abstract: *In this work, we provide Bayesian inference for a linear model selected after observing the data. Adopting Yekutieli's [J. Roy. Statist. Soc. B, 74(3), 515–541, 2012] ideas, the Bayesian model consists of a prior and a truncated likelihood. The resulting posterior distribution, unlike in the setup usually considered when performing Bayesian variable selection, is a affected by the very fact that selection was applied. A major computational challenge in such an approach is the intractability of the truncated likelihood. At the core of our methods is a convex approximation to the truncated likelihood, which facilitates sampling from the (approximate) adjusted posterior distribution. We demonstrate both theoretically and in simulations that employing the proposed approximation results in Bayesian procedures are qualitatively similar to those using the exact truncated likelihood. Replacing the truncated likelihood by its approximation, we can approximate the maximum-likelihood estimate as the MAP estimate corresponding to a constant prior. Our approximation of the full truncated likelihood also has a frequentist appeal, equipped to address frequentist questions that have not been resolved in existing work on exact post-selection inference.*

Aaditya Ramdas, University of California Berkeley

Title: The p-filter: multi-layer FDR control for grouped hypotheses

Abstract: *(Joint work with Rina Foygel Barber, accepted at JRSS-B) In many practical applications of multiple testing, there are natural ways to partition the hypotheses into groups using the structural, spatial or temporal relatedness of the hypotheses, and this prior knowledge is not used in the classical Benjamini-Hochberg (BH) procedure for controlling the false discovery rate (FDR). When one can define (possibly several) such partitions, it may be desirable to control the group-FDR simultaneously for all partitions (as special cases, the "finest" partition divides the n hypotheses into n groups of one hypothesis each, and this corresponds to controlling the usual notion of FDR, while the "coarsest" partition puts all n hypotheses into a single group, and this corresponds to testing the global null hypothesis). In this paper, we introduce the p-filter, which takes*

as input a list of n p-values and $M \geq 1$ partitions of hypotheses, and produces as output a list of $\leq n$ discoveries such that group-FDR is provably simultaneously controlled for all partitions. Importantly, since the partitions are arbitrary, our procedure can also handle multiple partitions which are nonhierarchical. The p-filter generalizes two classical procedures—when $M = 1$, choosing the finest partition into n singletons, we exactly recover the BH procedure, while choosing instead the coarsest partition with a single group of size n, we exactly recover the Simes test for the global null. We verify our findings with simulations that show how this technique can not only lead to the aforementioned multi-layer FDR control, but also lead to improved precision of rejected hypotheses. We present some illustrative results from an application to a neuroscience problem with fMRI data, where hypotheses are explicitly grouped together according to predefined regions of interest (ROIs) in the brain, thus allowing the scientist to explicitly and flexibly employ field-specific prior knowledge.

Xiwei Tang, University of Illinois Urbana-Champaign

Title: Individualized Subgroup Variable Selection

Abstract: *We propose a novel individualized variable selection method which performs coefficient estimation, subgroup identification and variable selection simultaneously. In contrast to traditional model selection approaches, an individualized regression model allows different individuals to have different relevant variables. The key component of the new approach is to construct a separation penalty which utilizes cross-subject information and assumes that within-group subjects share the same homogeneous effect. This allows us to borrow information from subjects within the same subgroup, and therefore improve the estimation efficiency and variable selection accuracy for each individual. Another advantage of the proposed approach is that it combines strength of homogeneity and heterogeneity in modeling, and therefore enhances the prediction power. We provide theoretical justification for the proposed approach, and propose an effective algorithm to achieve an individualized variable selection. Simulation studies and an application to the HIV longitudinal data are illustrated to compare the new approach to existing penalization methods.*

Suzanne Thornton, Rutgers University

Title: Approximate confidence distribution computing (ACC)

Abstract: *(Joint with Min-ge Xie) Approximate Bayesian computing (ABC) is a likelihood-free method that has grown increasingly popular since early applications in population genetics. However, the theoretical justification for inference based on this method has yet to be fully developed; a key problem is that ABC produces a posterior distribution based on the likelihood of a non-sufficient summary statistic. We introduce a more general computational technique, approximate confidence distribution computing (ACC), developed entirely within a frequentist framework. Inference based on ACC is justified (even if reliant upon a non-sufficient summary statistic) by establishing correct frequentist coverage properties using the theory of confidence distributions. Subsequently, the coverage performance of ACC does not solely rely on Bernstein von Mises asymptotic theories. Furthermore, no prior assumptions are necessary for ACC but can be included when*

available without damaging the integrity of ACC based inference. We supplement the theory with examples that illustrate the applications of ACC and provide some simulation results.

Wei Wang, Washington University in St. Louis

Title: High-Dimensional Covariance and Precision Matrix Estimation

Abstract: *Due to availability of high-dimensional data in various fields, such as genomics and finance, there has been a growing interest in high-dimensional covariance and precision matrix estimation. Three techniques on large covariance and precision matrix estimations are reviewed, the Lediot-Wolf estimator, CLIME and CONDREG. The Lediot-Wolf estimator is a linear shrinkage covariance matrix estimator towards a specified target covariance matrix, and chooses the optimal shrinkage to optimize the variance-bias tradeoff. The CLIME produces a sparse precision matrix estimator based on optimizing a L1-penalized log-likelihood function. The CONDREG can produce a well-conditioned estimator by imposing a constraint on the condition number, and no sparsity assumption on either the covariance matrix or its inverse is needed. Based on the idea of a weighted average from the Lediot-Wolf estimator, the precision matrix estimator as a linear combination of the CLIME and CONDREG was studied, i.e. alpha*CLIME +(1-alpha)*CONDREG, which can be viewed as a balance between sparsity and eigenstructure regularization.*

Liqun Yu, Washington University in St. Louis

Title: Feature selection via approximated information criteria

Abstract: *We propose a new group-type feature selection method called the "group minimum information criteria" (gMIC) for the generalized linear model. The idea is to minimize an approximated Bayesian Information Criterion (BIC). The gMIC is formulated in two steps. First, the BIC is approximated by replacing the ℓ_0- norm in BIC with a smooth group-type penalty. In the second step, a reparameterization that encourages group level sparsity while maintaining the smoothness of the objective function, is applied. The gMIC enjoys two major benefits. First, it does not involve a tuning parameter and is hence computationally appealing. Second, unlike other penalization methods, the gMIC achieves feature selection and estimation in one single optimization step. As we shall see, this offers a unique opportunity for circumventing the post-selection inference. Theoretical properties of the gMIC are derived and numerical simulations are presented to back up the theory.*