WHOA -PSI 2017 Workshop on Higher-Order Asymptotics and Post-Selection Inference Washington University in St.Louis AUG 12-14

WHOA-PSI 2017

Second Workshop on Higher-Order Asymptotics and Post-Selection Inference

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

12 – 14 August, 2017

Schedule of Talks, Abstracts

Organizers:

John Kolassa, Todd Kuffner, Ryan Tibshirani

Rina Foygel Barber, University of Chicago

Title: Distribution-free inference for estimation and classification

Abstract: *Given a "black box" algorithm that estimates Y given a high-dimensional feature vector X, is it possible to perform inference on the quality of our estimates, without any distributional assumptions? Specifically, the black box algorithm fits some regression function $\mu(x)$ that attempts to estimate the mean of Y given $X = x$ (or, in a binary response setting, $P(Y = 1|X = x)$). We would like to then construct a confidence interval for this estimate without any assumptions on the distribution of the data. This talk will cover some ongoing work on this open problem. I will discuss preliminary results for data-driven inference on isotonic regression, where a black box algorithm computing an estimated regression function $\mu(x)$ can be recalibrated to provide a confidence interval for this estimate. Joint work with Fan Yang.*

Yoav Benjamini, Tel Aviv University

Title: Challenges in Large Selective Inference Problems

Abstract: *I shall classify current approaches to multiple inferences according to goals, and discuss the basic approaches being used. I shall then highlight a few challenges that await our attention: some are simple inequalities, others arise in particular applications.*

Yuval Benjamini, Hebrew University of Jerusalem

Title: Measuring bumps: Selection-corrected inference for detected regions in high-throughput data

Abstract: *High-throughput technologies allow scientist to simultaneously measure signal at numerous locations simultaneously. Such data allows scientists to search for spatially-coherent regions which are correlated with a covariate of interest. A popular approach is to (a) estimate a statistic at each point, (b) threshold this map and (c) merge neighboring points passing the threshold into regions. However, treating these regions as if they were known a-priori leads to biases in the estimation, especially considering the large spatial process from which they were selected. Currently, methods for removing these biases accommodate only significance testing, and require strong assumptions or very strict procedures.*

In this talk I will present a conditional-inference based approach that also allows estimating and forming confidence intervals for regional parameters. The proposed method is based on sampling from a conditional distribution, and therefore can accommodate the non-stationary covariance in each region. I will discuss common use cases such as evaluating differentially methylated regions (mDNA) between tissue types, and well as cortical (fMRI) region detection. This is based on work with Jonathan Taylor, Rafael Irizarry and Amit Meir.

Andreas Buja, University of Pennsylvania

Title: Higher-Order von Mises Expansions, Bagging, and Assumption-Lean Inference

Abstract: *The starting point of this work is the fact that models obtained from model selection procedures cannot be assumed to be correct. They therefore require assumption-lean inference and, in particular, assumption-lean standard errors. These can be obtained in two ways: (1) asymptotic plug-in, resulting in sandwich standard errors, and (2) bootstrap and, more specifically in regression, "pairs" or "x-y" bootstrap (as opposed to the residual bootstrap). Two questions arise: Is there a connection between the two types of standard errors? Is one superior to the other? We give some partial answers by showing that (a) asymptotic plug-in standard errors are a limiting case of bootstrap standard errors when the resample size varies, (b) bootstrap standard errors can be thought of as resulting from a bagging operation, (c) bagging results in statistical functionals that have finite von Mises expansions, and finally (d) bootstrap standard errors are "smoother" than sandwich standard errors in the sense of von Mises expansions. We conclude that bootstrap standard errors have some advantages over sandwich standard errors. However, we do not yet fully understand the trade-offs in choosing specific bootstrap resample sizes. Joint with: W.Stuetzle, L.Brown, R.Berk, L.Zhao, E.George, A.Kuchibhotla, K.Zhang*

Florentina Bunea, Cornell University

Title: Overlapping clustering with LOVE

Abstract: *The area of overlapping variable clustering, with statistical guarantees, is largely unexplored. We propose a novel Latent model-based OVErlapping clustering method (LOVE) to recover overlapping sub-groups of a p dimensional vector X from a sample of size n on X, with p allowed to be larger than n. In our model-based formulation, a cluster is given by variables associated with the same latent factor. Clusters are anchored by a few components of X that are respectively associated with only one latent factor, while the majority of the X-components may have multi-factor association. We prove that, under minimal conditions, these clusters are identifiable, up to label switching. LOVE estimates first the set of variables that anchor each cluster, and also estimates the number of clusters. In a second step, clusters are populated with variables with multiple associations. Under minimal signal strength conditions, LOVE recovers the population level overlapping clusters consistently. The practical relevance of LOVE is illustrated through the analysis of a RNA-seq data set, devoted to determining the functional annotation of genes with unknown function.*

Gerda Claeskens, KU Leuven

Title: Asymptotic post-selection inference after model selection by the Akaike information criterion

Abstract: *The asymptotic distribution of parameter estimators after model selection by the Akaike information criterion is studied. We exploit the overselection property of this criterion in the construction of a selection region, and obtain the asymptotic distribution of parameter estimators and linear combinations thereof in the selected model. The proposed method does not require the true model to be in the model set. We investigate the method by simulations in linear and generalized linear models. This is joint work with A. Charkhi (KU Leuven).*

Noureddine El Karoui, UC Berkeley

Title: Can we trust the bootstrap? (for moderately difficult statistical problems)

Abstract: *(Joint work with Elizabeth Purdom) The bootstrap is an important and widely used tool for answering inferential questions in Statistics. It is particularly helpful in many analytically difficult situations.*

I will discuss the performance of the bootstrap for simple inferential problems in moderate and high-dimension. Those are typically analytically still challenging and practically relevant, hence a good testing ground for resampling methods.

For instance, one can ask whether the bootstrap provides valid confidence intervals for individual parameters in linear regression when the number of predictors is not infinitely small compared to the sample size. Similar questions related to Principal Component analysis are also natural from a practical standpoint.

We will see that the answer to these questions is generally negative. Our assessment will be done through a mix of numerical and theoretical investigations, and include a few other resampling methods.

Jianqing Fan, Princeton University

Title: Robust inference for high-dimensional data with application to false discovery rate control

Abstract: *Heavy tailed distributions arise easily from high-dimensional data and they are at odd with commonly used sub-Gaussian assumptions. This prompts us to revisit the Huber estimator from a new perspective. The key observation is that the robustification parameter should adapt to sample size, dimension and moments for an optimal bias-robustness tradeoff: a small robustification parameter increases the robustness of estimation but introduces more biases. Our framework is able to handle heavy-tailed data with bounded* $(1+\delta)$*-th moment for any* $\delta > 0$*. We establish a sharp phase transition for robust estimation of regression parameters in both low and high dimensions: when* $\delta > 1$*, the estimator exhibits the optimal sub-Gaussian deviation bound, while only a slower rate is available in the regime* $0 < \delta < 1$*. The transition is smooth and optimal. Moreover, a nonasymptotic Bahadur representation for finite-sample inference is derived with finite variance, which further yields two important normal approximation results, the Berry-Esseen inequality and Cramér-type moderate deviation. As an important application, we apply these robust normal approximation results to analyze a dependence-adjusted multiple testing procedure. It is shown that this robust, dependence-adjusted procedure asymptotically controls the overall false discovery proportion (FDP) at the nominal level under mild moment conditions. Thorough numerical results on both simulated and real datasets are also provided to back up our theory.*

Will Fithian, UC Berkeley

Title: Adaptive Sequential Model Selection

Abstract: *Many model selection algorithms produce a path of fits specifying a sequence of increasingly complex models. Given such a sequence and the data used to produce them, we consider the problem of choosing the least complex model that is not falsified by the data.*

Extending the selected-model tests of Fithian, Sun and Taylor (2014), we construct p-values for each step in the path which account for the adaptive selection of the model path using the data. In the case of linear regression, we propose two specific tests, the max-t test for forward stepwise regression (generalizing a proposal of Buja and Brown (2014)), and the next-entry test for the lasso. These tests improve on the power of the saturated-model test of Tibshirani et al. (2014), sometimes dramatically. In addition, our framework extends beyond linear regression to a much more general class of parametric and non- parametric model selection problems. To select a model, we can feed our single-step p-values as inputs into sequential stopping rules such as those proposed by G'Sell et al. (2016) and Li and Barber (2016), achieving control of the family- wise error rate or false discovery rate (FDR) as desired. The FDR- controlling rules require the null p-values to be independent of each other and of the non-null p-values, a condition not satisfied by the saturated-model p-values of Tibshirani et al. (2014). We derive intuitive and general sufficient conditions for independence, and show that our proposed constructions yield independent p-values.

Max G'Sell, Carnegie Mellon University

Title: Adaptive sequential model selection for the graphical lasso

Abstract: *The graphical lasso is often used in practice to model the dependence structure among variables. However, inferential questions about the resulting solution have traditionally been difficult to answer. We focus on sequential goodness-of-fit tests for the graphcal lasso, using the sequential framework of Fithian et. al (2015+). This serves as an illustration of this approach in a reasonably complex, non-regression setting, and highlights some interesting issues that arise in its application. Time permitting, we will also introduce significance testing for particular edges, and its interplay with the reliability of graphical lasso estimates.*

Dalia Ghanem, UC Davis

Title: Model selection and prediction after selection in linear fixed effects models with high-frequency regressors

Abstract: *We propose a cross-validation model selection procedure for linear fixed effects modeling of mixed-frequency time series data. Conditions ensuring consistency of this procedure for model selection are considered. As such modeling techniques are commonly used for forecasting, e.g. in climate change impact studies, we construct valid post-selection prediction intervals. Simulations illustrate the improvements over existing methods, and the properties of our post-selection prediction intervals. Joint work with Todd Kuffner.*

Lucas Janson, Harvard University

Title: Knockoffs as Post-Selection Inference

Abstract: *The motivation behind post-selection inference is that data analysts often look at their data before choosing inferential questions, and that doing so invalidates inferential results when using procedures that assume the inferential question is fixed ahead of time. A natural (although highly non-trivial) solution that has gained quite a bit of traction has been to develop corrected or altogether new procedures for performing inference* *that adjust for whatever selection or data snooping preceded that inference. Knockoffs, particularly the more recent model-free knockoffs, was developed without post-selection inference in mind, but can be interpreted and used to do just that for the controlled variable selection problem, i.e., selecting a set of important variables that controls the false discovery rate. The knockoffs procedure augments the data with artificial null variables *before* the data analyst views it, and does so in a way that automatically and exactly protects against selection effects in the final inference step. Because the main work of correcting for selection occurs before the analyst sees the data, the knockoffs approach is procedurally somewhat similar to the use of differential privacy in adaptive data analysis, wherein noise is added to the data before the analyst views it to choose inferential questions, although differential privacy results usually bound how far from valid the inference is, while with knockoffs the inference is exact. An advantage of the knockoffs approach is that it places absolutely no restriction on how the analyst views, models, or performs selection on the augmented data, and yet this generality does not make its inference conservative at all. In this talk, I will briefly review the knockoffs procedure and then go into detail about how it can be used to immunize inference to the effects of selection.*

Arun Kumar Kuchibhotla, University of Pennsylvania

Title: An approach to valid post selection inference in assumption-lean linear regression analysis

Abstract: *In this talk, I will present three different but closely related confidence regions that provide valid post-selection inference in an assumption-lean linear regression analysis. Here the phrase "assumption-lean linear regression" is used to mean that we do not assume any of the classical linear regression assumptions on the data generating process. We allow the observations to be structurally dependent and non-identically distributed. In trend with the current high dimensional literature we let the number of covariates that can be used in model selection be almost exponential in the sample size. The guarantee provided by these confidence regions is uniform in the sense of Berk et al. (2013) in that they have valid coverage (asymptotically) for any (random) model-selection procedure. Most importantly, all these three confidence regions can be computed in polynomial-time. Some important extensions and other methods will be discussed time permitting. This is a joint work with Andreas Buja, Richard Berk, Lawrence Brown, Ed George and Linda Zhao.*

Stephen M.S. Lee, University of Hong Kong

Title: A Mixed Residual Bootstrap Procedure for Least-Squares Regression Post-Model Selection

Abstract: *Common practice in linear regression often assumes a model predetermined by a data-driven procedure. Taking into consideration the extra uncertainty introduced by data-driven model selection, recent studies on distributions of post-model-selection least squares estimators have uncovered considerable departures from the standard theory established under the assumption of a fixed model. Consistent estimation of such distributions poses a challenge with important implications for valid regression inferences. We propose a* *mixed residual bootstrap procedure for estimating the true distributions of post-model-selection least squares estimators. The procedure integrates residual bootstrap estimates constructed from candidate models adaptively into a mixture distribution, which is shown to be consistent under the conditions that the approximately correct candidate models are sufficiently separated from the wrong ones and that the model selector is approximately conservative or consistent. Empirical performance of our estimator is illustrated via simulation, with models selected by AIC and the LASSO.*

Hannes Leeb, University of Vienna

Title: Prediction when fitting simple models to high-dimensional data

Abstract: *We study linear subset regression in the context of a high-dimensional linear model. Consider $y = \vartheta + \theta' z + \epsilon$ with univariate response y and a d-vector of random regressors z, and a submodel where y is regressed on a set of p explanatory variables that are given by $x = M'z$, for some $d \times p$ matrix M. Here, 'high-dimensional' means that the number d of available explanatory variables in the overall model is much larger than the number p of variables in the submodel. In this paper, we present Pinsker-type results for prediction of y given x. In particular, we show that the mean squared prediction error of the best linear predictor of y given x is close to the mean squared prediction error of the corresponding Bayes predictor $E[y|x]$, provided only that $p/\log d$ is small. We also show that the mean squared prediction error of the (feasible) least-squares predictor computed from n independent observations of (y, x) is close to that of the Bayes predictor, provided only that both $p/\log d$ and p/n are small. Our results hold uniformly in the regression parameters and over large collections of distributions for the design variables z.*

Jing Lei, Carnegie Mellon University

Title: Cross-validation with Confidence

Abstract: *Cross-validation is one of the most popular model selection methods in statistics and machine learning. Despite its wide applicability, traditional cross-validation methods tend to overfit, unless the ratio between the training and testing sample sizes is very small. We argue that such an overfitting tendency of cross-validation is due to the ignorance of the uncertainty in the testing sample. We develop a new, statistically principled inference tool based on cross-validation, that takes into account the uncertainty in the testing sample. Our method outputs a small set of highly competitive candidate models that contains the best one with probabilistic guarantees. In particular, our method leads to consistent model selection in a classical linear regression setting, for which existing methods require unconventional split ratios. We demonstrate the performance of the proposed method in simulated and real data examples.*

Xihong Lin, Harvard University

Title: Detection of Signal Regions Using Scan Statistics With Applications in Whole Genome Association Studies

Abstract: *We consider in this paper detection of signal regions associated with disease outcomes in whole genome association studies. The existing gene- or region-based methods test for the association of an outcome and the genetic variants in a pre-specified region, e.g.,* *a gene. In view of massive intergenic regions in whole genome association studies, we propose a quadratic scan statistic based method to detect the existence and the locations of signal regions by scanning the genome continuously. The proposed method accounts for the correlation (linkage disequilibrium) among genetic variants, and allows for signal regions to have both causal and neutral variants, and causal variants whose effects can be in different directions. We study the asymptotic properties of the proposed scan statistics. We derived an asymptotic threshold that controls for the family-wise error rate, and show that under regularity conditions the proposed method consistently selects the true signal regions. We performed simulation studies to evaluate the finite sample performance of the proposed method. Our simulation results showed that the proposed procedure outperforms the existing methods, especially when signal regions have causal variants whose effects are in different directions, or are contaminated with neutral variants, or the variants in signal regions are correlated. We applied the proposed method to analyze a lung cancer genome-wide association study to identify the genetic regions that are associated with lung cancer risk.*

Han Liu, Princeton University

Title: Combinatorial inference

Abstract:

Po-Ling Loh, University of Wisconsin

Title: Robust estimation, efficiency, and Lasso debiasing

Abstract: *We present results concerning high-dimensional robust estimation for linear regression with non-Gaussian errors. We provide error bounds for certain local/global optima of penalized M-estimators, valid even when the loss function employed is nonconvex – giving rise to more robust estimation procedures. We also present a new approach for robust location/scale estimation with rigorous theoretical guarantees. We conclude by discussing high-dimensional variants of one-step estimation procedures from classical robust statistics and connections to recent work on confidence intervals based on Lasso debiasing.*

Jelena Markovic, Stanford University

Title: Formalizing data science through selective inference: Selective inference after multiple views / queries on the data

Abstract: *We present the recent developments in the conditional approach to selective inference starting with Lee et al. (2016) through three examples. In all the examples, we perform a model selection procedure on our data and choose the coefficients to report inference for after looking at the outcome of this procedure. In order to have valid post-selection inference for the selected coefficients, we base our inference using the conditional distribution of the data, where the conditioning is on the observed outcome of the model selection procedure.*

In the first example, we use cross-validated LASSO to select the model. Then we use the inferential tools applicable in a even more general framework that require the *constraints coming from a model selection procedure to be written in terms of asymptotically Gaussian random vectors. By adding a small randomization to cross-validation error vector, we show cross-validation error vector and the data vector corresponding to LASSO are jointly asymptotically Gaussian.*

As most data analysts will want to try various model selection algorithms when choosing a model, we present a way to do inference after multiple views / queries on the data as our second example. Here, we do marginal screening first and then fit LASSO to the resulting model consisting of the variables selected by marginal screening and their interactions. A technique that enables us to achieve the valid post-selection here relies on adding randomization to each objective of the selection procedures.

Our third example shows the advantage in power of data carving compared to data splitting. Data splitting uses a part of the data for model selection and the leftover data for inference. Data carving reuses the leftover information in the first part of the data to construct valid p-values and confidence intervals for the selected coefficients. As a further application, we also present an example of doing valid inference after multiple splits of the data.

Ryan Martin, North Carolina State University

Title: On valid post-selection prediction in regression

Abstract: *TBA*

Peter McCullagh, University of Chicago

Title: Laplace approximation of high-dimensional integrals

Abstract: *TBA*

Ian McKeague, Columbia University

Title: Marginal screening for high-dimensional predictors of survival outcomes

Abstract: *A new marginal screening test is developed for a right-censored time-to-event outcome under a high-dimensional accelerated failure time (AFT) model. Establishing a rigorous screening test in this setting is challenging, not only because of the right censoring, but also due to the post-selection inference in the sense that the implicit variable selection step needs to be taken into account to avoid an inflated Type I error. McKeague and Qian (2015) constructed an adaptive resampling test to circumvent this problem under ordinary linear regression. To accommodate right censoring, we introduce an approach based on a maximally selected Koul-Susarla-Van Ryzin estimator from a marginal AFT working model. A regularized bootstrap method is developed to calibrate the test. The talk is based on joint work with Tzu-Jung Huang and Min Qian.*

Xiao-Li Meng, Harvard University

Title: Struggling with large p and small n? How about (potentially) infinite p and (essentially) zero n?

Abstract: *The arrival of "big data" has expanded the statistical asymptotic land from fixing-p-growing-n to growing-p-&-n in a variety of ways. But this expansion is still not rich enough to establish a rigorous theoretical foundation for individualized prediction and learning (e.g., for deciding personalized treatment) when the individuality corresponds to (potentially) infinite many attributes/variables, such as for human beings. Furthermore, for such individuals, there will be no direct "training sample" coming from their genuine guinea pigs because it is impossible to match on infinitely many attributes. Consequently, the traditional notions of estimation consistency, unbiasedness, etc., all become unreachable even in theory. It is possible, however, to set up an approximation theory via a multi-resolution (MR) perspective, inspired by wavelets (and other) literature, where we use the resolution level to index the degree of approximation to the ultimate individuality. The key strategy here is to seek an "indirect learning population (ILP)" by matching enough attributes to increase the relevance of the results to the individuals and yet still accumulate sufficient sample sizes for robust estimation. Theoretically, MR relies on a standard ANOVA type of decomposition, albeit with indefinitely many terms to permit the resolution level approaching infinity. This talk reports recent progress on MR asymptotic results, in terms of the ILP size, that can render statistical insights, and on adaptive error estimations aiming for practical implementation of optimal individualized learning from a given ILP. (This is a joint work with Xinran Li.)*

Yang Ning, Cornell University

Title: A General Framework for High-Dimensional Inference and Multiple Testing

Abstract: *We consider the problem of how to control the false scientific discovery rate in high-dimensional models. Towards this goal, we focus on the uncertainty assessment for low dimensional components in high-dimensional models. Specifically, we propose a novel decorrelated likelihood based framework to obtain valid p-values for generic penalized M-estimators. Unlike most existing inferential methods which are tailored for individual models, our method provides a general framework for high-dimensional inference and is applicable to a wide variety of applications, including generalized linear models, graphical models, classifications and survival analysis. The proposed method provides optimal tests and confidence intervals. The extensions to general estimating equations are discussed. Finally, we show that the p-values can be combined to control the false discovery rate in multiple hypothesis testing.*

Snigdha Panigrahi, Stanford University

Title: A Cure-All Truncated Approach to Provide Selection Adjusted Effect Size Estimates

Abstract: *My talk will describe new methods to provide adjusted estimates for effects in GWAS studies post a genome-wide selection. The starting point in such studies is an exhaustive collection of genetic variants; the goal being to identify and quantify effect sizes of a very small subset of variants that are believed to determine an outcome. In achieving such a goal, the scientist per- forms a genome wide search of the strongest possible associations and is subsequently, confronted in defending the significance of her findings. Motivated to measure these effect sizes as consistent point estimates and intervals with target coverage, my methods are modeled along the truncated approach to selective inference in Lee and Taylor (2014) and Fithian et al. (2014).*

At the core of my methods is a convex approximation to the typically intractable truncated likelihood as a function of the parameters in the model. This allows frequentist inference free from any MCMC samplers through a tractable and approximate pivotal quantity. Appending the approximate truncated likelihood to a prior on the parameters post selection allows Bayesian inference based on a truncated posterior. Estimates based on the truncated posterior show superior frequentist properties like Bayesian FCR and Bayes risk in comparison to the unadjusted estimates. I will illustrate these inferential gains for effect size estimates using a truncated model, both for a frequentist and a Bayesian in the context of a GWAS experiment. The talk will be based on a combination of Panigrahi et al. (2016, arXiv:1605.08824); Tian et al. (2016, arXiv:1609.05609); Panigrahi et al. (2017, arXiv:1703.06154); Panigrahi and Taylor (2017, arXiv:1703.06176).

Annie Qu, University of Illinois Urbana-Champaign

Title: Variable Selection for Highly Correlated Predictors

Abstract: *(Joint work with Fei Xue) Penalty-based variable selection methods are powerful in selecting relevant covariates and estimating coefficients simultaneously. However, variable selection could fail to be consistent when covariates are highly correlated. The partial correlation approach has been adopted to solve the problem with correlated covariates. Nevertheless, the restrictive range of partial correlation is not effective for capturing signal strength for relevant covariates. In this paper, we propose a new Semi-standard PArtial Covariance (SPAC) which is able to reduce correlation effects from other predictors while incorporating the magnitude of coefficients. The proposed SPAC variable selection facilitates choosing covariates which have direct association with the response variable, via utilizing dependency among covariates. We show that the proposed method with the Lasso penalty (SPAC-Lasso) enjoys strong sign consistency in both finite-dimensional and high-dimensional settings under regularity conditions. Simulation studies and the 'HapMap' gene data application show that the proposed method outperforms the traditional Lasso, adaptive Lasso, SCAD, and Peter–Clark-simple (PC-simple) methods for highly correlated predictors.*

Aaditya Ramdas, UC Berkeley

Title: Interactive multiple testing with STAR: selectively traversed accumulation rules for structured FDR control

Abstract: *We propose a general framework (STAR) for interactive multiple testing with a human-in-the-loop that combines ideas from the literatures on post-selection inference, ordered multiple testing and the knockoff filter. In our setup, we have N independent p-values, each corresponding to a different hypothesis. Initially, only "masked" p-values $g(P)$ are revealed to the scientist, while, akin to data-carving, the "leftover" information $h(P)$ is used to estimate the FDP of the masked set. At each step of the interaction, the scientist may freely choose one or more p-values to "unmask", by utilizing information contained in all p-values that were previously unmasked, the remaining masked p-values, as well as any other side information in the form of covariates, prior knowledge or intuition. When the chosen p-value is unmasked and shown to the scientist, they may freely update* *any model or prior being used, and revise the estimated FDP of the remaining masked set. This process continues until the estimated FDP falls below a pre-determined level, at which point all remaining masked p-values are rejected, while all unmasked ones are accepted. We prove that this interactive procedure guarantees FDR control no matter what unmasking choices are made by the scientist. We also demonstrate how to apply this procedure when we have only knockoff statistics instead of p-values, with the absolute value and sign of the statistic respectively playing the role of $g(P)$ and $h(P)$. We provide heuristics to help the scientist with the unmasking process for a variety of structured applications, and show that STAR performs excellently in a wide range of problems such as convex region detection, bump-hunting, and FDR control on trees and DAGs. (joint work with Lihua Lei and Will Fithian, soon to be posted on Arxiv)*

Richard Samworth, University of Cambridge

Title: High-dimensional changepoint estimation via sparse projection

Abstract: *Changepoints are a very common feature of Big Data that arrive in the form of a data stream. In this work, we study high-dimensional time series in which, at certain time points, the mean structure changes in a sparse subset of the coordinates. The challenge is to borrow strength across the coordinates in order to detect smaller changes than could be observed in any individual component series. We propose a two-stage procedure called 'inspect' for estimation of the changepoints: first, we argue that a good projection direction can be obtained as the leading left singular vector of the matrix that solves a convex optimisation problem derived from the CUSUM transformation of the time series. We then apply an existing univariate changepoint estimation algorithm to the projected series. Our theory provides strong guarantees on both the number of estimated changepoints and the rates of convergence of their locations, and our numerical studies validate its highly competitive empirical performance for a wide range of data generating mechanisms. Software implementing the methodology is available in the R package 'InspectChangepoint'.*

Weijie Su, University of Pennsylvania

Title: When Does the First Spurious Variable Get Selected by Sequential Regression Procedures?

Abstract: *Applied statisticians use sequential regression procedures to produce a ranking of explanatory variables and, in settings of low correlations between variables and strong true effect sizes, expect that variables at the very top of this ranking are true. In a regime of certain sparsity levels, however, three examples of sequential procedures – forward stepwise, the lasso, and least angle regression – are shown to include the first spurious variable unexpectedly early. We derive a sharp prediction of the rank of the first spurious variable for the three procedures, demonstrating that the first spurious variable occurs earlier and earlier as the regression coefficients get denser. This counterintuitive phenomenon persists for independent Gaussian designs and an arbitrarily large magnitude of the effects. We further gain a better understanding of the phenomenon by identifying the underlying cause and then leverage the insights to introduce a simple visualization tool termed the "double-ranking diagram" to improve on sequential methods.*

As a byproduct of these findings, we obtain the first provable result certifying the exact equivalence between the lasso and least angle regression in the early stages of solution paths beyond orthogonal designs. This equivalence can seamlessly carry over many important model selection results concerning the lasso to least angle regression.

Stefan Wager, Stanford University

Title: Efficient Policy Learning

Abstract: *There has been considerable interest across several fields in methods that reduce the problem of learning good treatment assignment policies to the problem of accurate policy evaluation. Given a class of candidate policies, these methods first effectively evaluate each policy individually, and then learn a policy by optimizing the estimated value function; such approaches are guaranteed to be risk-consistent whenever the policy value estimates are uniformly consistent. However, despite the wealth of proposed methods, the literature remains largely silent on questions of statistical efficiency: there are only limited results characterizing which policy evaluation strategies lead to better learned policies than others, or what the optimal policy evaluation strategies are. In this paper, we build on classical results in semiparametric efficiency theory to develop quasi-optimal methods for policy learning; in particular, we propose a class of policy value estimators that, when optimized, yield regret bounds for the learned policy that scale with the semiparametric efficient variance for policy evaluation. On a practical level, our result suggests new methods for policy learning motivated by semiparametric efficiency theory. Joint work with Susan Athey.*

Huixia Judy Wang, George Washington University

Title: Inference for High Dimensional Quantile Regression

Abstract: *In this talk I will present a new marginal testing procedure to detect the presence of significant predictors associated with the quantiles of a scalar response. The idea is to fit the marginal quantile regression on each predictor separately, and then base the test on the t-statistic associated with the chosen most informative predictor at the quantiles of interest. A resampling method is devised to calibrate this test statistic, which has non-regular limiting behavior due to the variable selection. Asymptotic validity of the procedure is established in a general quantile regression setting in which the marginal quantile regression models can be misspecified. Even though a fixed dimension is assumed to derive the asymptotic results, the proposed test is applicable and computationally feasible for large-dimensional predictors. The method is more flexible than existing marginal screening test methods based on mean regression, and has the added advantage of being robust against outliers in the response. The approach is illustrated using an application to an HIV drug resistance dataset. This is a joint work with Ian McKeague and Min Qian at Columbia University.*

Wei Biao Wu, University of Chicago

Title: Testing for Trends in High-dimensional Time Series

Abstract: *We consider statistical inference for trends of high-dimensional time series. Based on a modified L2-distance between parametric and nonparametric trend estimators, we propose a de-diagonalized quadratic form test statistic for testing patterns on trends, such as linear, quadratic or parallel forms. We develop an asymptotic theory for the test statistic. A Gaussian multiplier testing procedure is proposed and it has an improved finite sample performance. Our testing procedure is applied to a spatial temporal temperature data gathered from various locations across America. A simulation study is also presented to illustrate the performance of our testing method. The work is joint with Likai Chen.*

Min-ge Xie, Rutgers University

Title: A union of BFF (Bayesian, frequentist and fiducial) inferences by confidence distribution and artificial data sampling

Abstract: *In this talk, we will briefly introduce some new developments of confidence distribution and also our new understanding about fiducial inference. A comparative study of the BFF inferences from a unique simulation (artificial sample data) angle will be provided. It will be illustrated that the connections between Bayesian and frequentist inferences (including modern Bayesian computing procedures, bootstrap and many other resampling and simulation methods) are much more congruous and coherent than what have been perceived in the scientific community. Specifically, we will discuss the ideas of uncertainty matching by CD-random variable, by bootstrapping, by fiducial samples and also by Bayesian posterior samples, in which the uncertainty by a model population is matched with the uncertainty from the same or a similar artificial data generation scheme. A general theory of uncertainty matching for exact and asymptotic inference will be provided. Some preliminary ideas of higher order matching and model selection under the same artificial data generation framework will also be explored.*

Daniel Yekutieli, Tel Aviv University

Title: Post selection inference in replicated and repeated experiments

Abstract: *I will discuss selection and post selection inference for cases where an experiment may be replicated under different conditions and may also be repeated several times for the same condition.*

Alastair Young, Imperial College London

Title: Measuring nuisance parameter effects in Bayesian inference

Abstract: *A key concern of statistical theory is interpretation of the effects of nuisance parameters on an inference about an interest parameter. Especially important for statistical practice is quantification of the consequences of including potentially high-dimensional nuisance parameters to provide realistic modelling of a system under study. Through decomposition of the Bayesian version of an adjusted likelihood ratio statistic, we consider easily computed measures of nuisance parameter effects in Bayesian inference and illustrate their utility in practical examples, including hierarchical modelling.*

Russell Zaretzki, University of Tennessee Knoxville

Title: Feasible and powerful simultaneous inference after lasso

Abstract: *We propose methodology for simultaneous inference after lasso in the spirit of the PoSI procedure (Berk et al., 2013), but which has the advantage of yielding considerably narrower intervals, and which is computationally less-demanding. Theoretical control of familywise error rate is considered, and simulations are presented to assess the validity of our coverage error claims in realistic scenarios, as well as compare the interval lengths with more conservative methods. Joint work with Todd Kuffner.*

Mayya Zhilova, Georgia Institute of Technology

Title: Higher-order Berry-Esseen inequalities and accuracy of the bootstrap

Abstract: *In this talk, we study higher-order accuracy of a bootstrap procedure for approximation in distribution of a smooth function of a sample average in high-dimensional non-asymptotic framework. Our approach is based on Berry-Esseen type inequalities which extend the classical normal approximation bounds. These results justify in non-asymptotic setting that bootstrapping can outperform Gaussian (or chi-squared) approximation in accuracy with respect to both dimension and sample size. In addition, the presented results lead to improvements of accuracy of a weighted bootstrap procedure for general log-likelihood ratio statistics (under certain regularity conditions). The theoretical results are illustrated with numerical experiments on simulated data.*

Poster Presentations

The poster sessions are on Sunday, 13 August from 11:30am-1:00pm

Kenneth Hung, UC Berkeley

Title: Rank Verification for Exponential Families

Abstract: *Many statistical experiments involve comparing multiple population groups. For example, a clinical trial may compare binary patient outcomes under several treatment conditions to determine the most effective treatment. Having observed the "winner" (largest observed response) in a noisy experiment, it is natural to ask whether that treatment is actually the "best" (stochastically largest response). This article concerns the problem of rank verification — post hoc significance tests of whether the orderings discovered in the data reflect the population ranks. For exponential family models, we show under mild conditions that an unadjusted two-tailed pairwise test comparing the top two observations (i.e., comparing the "winner" to the "runner-up") is a valid test of whether the winner is the best. We extend our analysis to provide equally simple procedures to obtain lower confidence bounds on the gap between the winning population and the others, and to verify ranks beyond the first.*

John Kolassa, Rutgers University

Title:

Abstract:

Amit Meir, University of Washington

Title: Selective Maximum Likelihood Inference – Theory and Application

Abstract: *Applying standard statistical methods after model selection may yield inefficient estimators and hypothesis tests that fail to achieve nominal type-I error rates. In recent years, much progress has been made in developing inference methods that are valid for selected model. However, relatively little attention has been given to the problem of estimation after model selection.*

Here, we propose to compute selective maximum likelihood estimators and construct confidence intervals based on their asymptotic properties: a route that is a matter of course in many statistical applications, yet novel in the context of selective inference. In order to do so, we develop a stochastic optimization framework for computing the selective MLE and derive its asymptotic properties.

We demonstrate the usefulness of our methodologies in two applied settings. In the context of fMRI studies, we use the proposed optimization framework to estimate the mean signal in selected regions of interest and construct confidence-intervals using a profile-likelihood type approach. In the context of rare-variant GWAS, we perform inference for genetic variants in genes that were selected via an aggregate test. Specifically, *we show that the problem of computing the (multivariate) MLE for regression models that were selected via an aggregate test can often be cast as a line-search problem. This is based on work with Mathias Drton, Yuval Benjamini and Ruth Heller.*

Honglang Wang, IUPUI

Title: Empirical Likelihood Ratio Tests for Coefficients in High Dimensional Heteroscedastic Linear Models

Abstract: *This paper considers hypothesis testing problems for a low-dimensional coefficient vector in a high-dimensional linear model with heteroscedastic variance. Heteroscedasticity is a commonly observed phenomenon in many applications including finance and genomic studies. Several statistical inference procedures have been proposed for low-dimensional coefficients in a high-dimensional linear model with homoscedastic variance. However, existing procedures designed for homoscedastic variance are not applicable for models with heteroscedastic variance and the heterscedasticity issue has been rarely investigated and studied. We propose a simple inference procedure based on empirical likelihood to overcome the heteroscedasticity issue. The proposed method is able to make valid inference even when the conditional variance of random error is an unknown function of high-dimensional predictors. We apply our inference procedure to three recently proposed estimating equations and establish the asymptotic distributions of the proposed methods. Simulation studies and real data analyses are conducted to demonstrate the proposed methods.*

Qi Wang, Washington University in St. Louis

Title: Pseudo-posterior Inference for Volatility in Lévy Models with Infinite Jumps

Abstract: *Jump diffusion models driven by Lévy processes are suitable models to capture the dynamics of financial asset returns. Efficient estimation of the volatility parameter is important for accurate measurement and effective control of risk. A complete Bayesian specification of jump diffusion models with infinite jump activity would require a prior distribution on the jump process, expressing relative beliefs about the size and frequency of jumps, as well as how these characteristics are time-dependent. The model for the jump process is difficult to evaluate, because it is discretely-observed and nonparametric. Furthermore, approximating the likelihood function and sampling from the full posterior can require advanced computational methods. Therefore, it is complicated to use fully Bayesian methods for inference about the volatility parameter, which would be based on its marginal posterior distribution. Martin, Ouyang & Domagni (2017) proposed to misspecify the model on purpose, and to assume only finite jump activity. Then inference can be based on the conditional posterior of the volatility parameter, for a given realization of the jump process. The misspecification is that the jump process is not modeled at all. It can be shown that the misspecified model posterior for the volatility will be centered at the wrong value, but that this can be corrected by a simple estimate of the location shift, and re-scaling the log-likelihood. Moreover, the Bernstein–von Mises theorem can be used to quantify posterior uncertainty, and to construct approximate credible intervals. Kleijn & van der Vaart (2012) gave conditions under which the Bernstein–von Mises theorem holds under model misspecification. Building on this theory, we extend the results of Martin, Ouyang & Domagni (2017) to models with infinite*

jump activity. We provide sufficient conditions for the Bernstein–von Mises theorem to hold in this setting, and formally prove the result. Joint work with Jose Figueroa-Lopez and Todd Kuffner.

Qiyiwen Zhang, Washington University in St. Louis

Title:

Abstract:

Xinwei Will Zhang, Rutgers University

Title:

Abstract: