Statistical Inference for Large Scale Data

April 20 to 24, 2015
Simon Fraser University

Speakers, titles and abstracts

1. **Speaker**: Jacob Bien, Cornell University

   Title: Convex Banding of the Covariance Matrix

   Abstract: *We introduce a sparse and positive definite estimator of the covariance matrix designed for high-dimensional situations in which the variables have a known ordering. Our estimator is the solution to a convex optimization problem that involves a hierarchical group lasso penalty. We show how it can be efficiently computed, compare it to other methods such as tapering by a fixed matrix, and develop several theoretical results that demonstrate its strong statistical properties. Finally, we show how using convex banding can improve the performance of high-dimensional procedures such as linear and quadratic discriminant analysis.*

2. **Speaker**: Andreas Buja, The Wharton School, University of Pennsylvania

   Title: (1) Large-p Visualization ; (2) Moderate-p Post-Selection Inference

   Abstract: *This talk will be divided into two short talks.*

   (1) Large-p Visualization: Data with large numbers of variables pose a challenge for data exploration and visualization. Facing hundreds of variables, standard tools such as scatterplot matrices and parallel coordinate plots will quickly reach their limit. It may be necessary to abandon detailed visualization of variables and instead show simple summary measures, as in heat maps of correlation tables that are often found in genomics. In this talk we use a modification of heat maps, here called "blockplots", and we demonstrate an associated software tool that is both interactive and dynamic, yet wholly written in R. Called "Association Navigator", its basic operation is zooming and panning of large "blockplots" of correlation tables involving hundreds of variables. The tool includes, among other things: (1) display of p-values and missing value patterns in addition to correlations, (2) mark-up facilities to highlight variables and sub-tables as landmarks when navigating the larger table, (3) histograms / barcharts, scatterplots and scatterplot matrices as "lenses" into the distributions of variables and variable pairs, as well as several other functionalities. The usefulness of the tool is less in beholding gigantic tables in their entirety and more in searching for interesting association patterns by navigating manageable but numerous and interconnected sub-tables.

   (2) Moderate-p Post-Selection Inference: Berk et al. (AoS 2013) proposed a stringent approach, called "PoSI", to producing statistical inference that is valid after arbitrary variable selection in regression. The approach is controversial because of its conservative nature, yet it is the appropriate method of securing valid statistical inference after what is called p-hacking or p-value hunting. This, however, is not the main subject of the talk. Rather, it is the computational limitations of the PoSI approach. Currently our algorithm can handle up to about $p = 20$ predictor variables in an unconstrained model search. For "sparse PoSI" we can handle up to about $p = 80$ predictors if model search is limited to model sizes 4 and lower. The algorithm relies both on a deterministic integration and a Monte Carlo simulation. It would be desirable for speeding up the algorithm if we better understood the underlying geometry of predictor adjustment. This part of the talk will therefore end with open questions.

3. **Speaker**: Venkat Chandrasekaran, California Institute of Technology

   Title: Latent-Variable Graphical Model Selection via Convex Optimization

Abstract:

Suppose we have a Gaussian graphical model with sample observations of only a subset of the variables. Can we separate the extra correlations induced due to marginalization over the unobserved, hidden variables from the structure among the observed variables? In other words is it still possible to consistently perform model selection despite the unobserved, latent variables? As we shall see the key problem that arises is one of decomposing the concentration matrix of the observed variables into a sparse matrix (representing graphical model structure among the observed variables) and a low rank matrix (representing the effects of marginalization over the hidden variables). Such a decomposition can be accomplished by an estimator that is given by a tractable convex program. This estimator performs consistent model selection in the high-dimensional scaling regime in which the number of observed/hidden variables grows with the number of samples of the observed variables. The geometric aspects of our approach are highlighted, with the algebraic varieties of sparse matrices and of low rank matrices playing an important role.

Joint work with Armeen Taeb, Pablo Parrilo, and Alan Willsky.

4. **Speaker**: Johannes Lederer, Cornell University

   Title: How to calibrate tuning parameters

   Abstract: *High-dimensional statistics is the basis for analyzing large and complex data sets that are generated by cutting-edge technologies in genetics, neuroscience, astronomy, and many other fields. However, Lasso, Ridge Regression, Graphical Lasso, and other standard methods in high-dimensional statistics depend on tuning parameters that are difficult to calibrate in practice. In this talk, I present two novel approaches to overcome this difficulty. My first approach is based on a novel testing scheme that is inspired by Lepskis idea for bandwidth selection in non-parametric statistics. This approach provides tuning parameter calibration for estimation and prediction with the Lasso and other standard methods and is to date the only way to ensure high performance, fast computations, and optimal finite sample guarantees. My second approach is based on the minimization of an objective function that avoids tuning parameters altogether. This approach provides accurate variable selection in regression settings and, additionally, opens up new possibilities for the estimation of gene regulation networks, microbial ecosystems, and many other network structures.*

5. **Speaker**: Jason Lee, Stanford University

   Title: Selective Inference via the Condition on Selection Framework and Communication-efficient Sparse Regression

   Abstract: *Selective Inference is the problem of testing hypotheses that are chosen or suggested by the data. Inference after variable selection in high-dimensional linear regression is a common example of selective inference; we only estimate and perform inference for the selected variables. We propose the Condition on Selection framework, which is a framework for selective inference that allows selecting and testing hypotheses on the same dataset. In the case of inference after variable selection (variable selection by lasso, marginal screening, or forward stepwise), the Condition on Selection framework allows us to construct confidence intervals for regression coefficients, and perform goodness-of-fit testing for the selected model.*

   In the second part of the talk, we consider the problem of sparse regression in the distributed setting. The main computational challenge in a distributed setting is harnessing the computational capabilities of all the machines while keeping communication costs low. We devise an approach that requires only a single round of communication among the machines. We show the approach recovers the convergence rate of the (centralized) lasso as long as each machine has access to an adequate number of samples.

6. **Speaker**: Hannes Leeb, University of Vienna

   Title: On conditional moments of high-dimensional random vectors given lower-dimensional projections

   Abstract:

   One of the most widely used properties of the multivariate Gaussian distribution, besides its tail behavior, is the fact that conditional means are linear and that conditional variances are constant. We here show that this property is also shared, in an approximate sense, by a large class of non-Gaussian distributions. We allow for several conditioning variables and we provide explicit non-asymptotic results, whereby we extend earlier findings of Hall and Li (1993) and Leeb (2013).

   These results have immediate consequences for modern statistical technology, in particular for inference with sparse working models when the true model need not be sparse.

   (This is joint work with Lukas Steinberger.)

7. **Speaker**: Po-Ling Loh, Wharton School, U Penn

   Title: High-dimensional robust M-estimation

   Abstract: *Many popular algorithms for high-dimensional regression (e.g., Lasso-based linear regression) involve optimizing penalized M-estimators. We present some recent results concerning the behavior of stationary points of such objective functions, when both the loss and penalty function are allowed to be nonconvex. We discuss new consequences of our results in the context of high-dimensional robust regression, in which alternative (often nonconvex) losses are used to limit the influence of outliers and heavy-tailed errors. Our main results involve statistical error bounds on the l_2-distance between any stationary point within a local region of convexity and the true regression vector. Furthermore, we show that an efficient two-step optimization algorithm involving a convex robust loss function in the first step may be used to obtain an initial point within the local convexity region, from which we may initialize a first-order algorithm to obtain a statistically consistent stationary point of the target estimator.*

8. **Speaker**: Aurelie Lozano

   Title: Elementary Estimators for High-dimensional Statistical Models

   Abstract: *High-dimensional inference problems, where the number of parameters to be estimated greatly exceeds the number of observations, are increasingly prevalent in the era of Big Data. Despite their high-dimensionality, these problems often contain hidden structures exhibiting some form of sparsity that can be exploited to achieve consistent inference. However, computational efficiency becomes a bottleneck in cases of very large-scale analysis, as current state-of-the-art high-dimensional estimators involving non-differentiable regularization are not readily scalable.*

   In this talk I introduce a new paradigm of elementary estimators for structurally constrained high-dimensional inference that addresses the scaling issue at the source. These estimators are in many cases available in closed-form, cover wide classes of supervised and unsupervised learning problems, and possess strong statistical guarantees despite their extreme simplicity.

   Joint work with Eunho Yang and Pradeep Ravikumar

9. **Speaker**: Richard Samworth, Cambridge University

   Title: Statistical and computational trade-offs in estimation of sparse principal components

   Abstract:

   In recent years, Sparse Principal Component Analysis has emerged as an extremely popular dimension reduction technique for high-dimensional data. The theoretical challenge, in the simplest case, is to estimate the leading eigenvector of a population covariance matrix under the assumption that this eigenvector is sparse. An impressive range of estimators have been proposed; some of these are fast to compute, while others are known to achieve the minimax optimal rate over certain Gaussian or subgaussian classes. We show that, under a widely-believed assumption from computational complexity theory, there is a fundamental trade-off between statistical and computational performance in this problem. More precisely, working with new, larger classes satisfying a Restricted Covariance Concentration condition, we show that no randomised polynomial time algorithm can achieve the minimax optimal rate. On the other hand, we also study a (polynomial time) variant of the well-known semidefinite relaxation estimator, and show that it attains essentially the optimal rate among all randomised polynomial time algorithms.

10. **Speaker**: Noah Simon, University of Washington, Department of Biostatistics, nrsimon@uw.edu

    Title: Adjusting Point Estimates and Confidence Intervals for Selection Bias in High Dimensions as a Frequentist

    Abstract:

    With recent advances in high throughput technology, researchers often find themselves running a large number of hypothesis tests (thousands+) and estimating a large number of effect-sizes. Generally there is particular interest in those effects estimated to be most extreme. Unfortunately naive estimates of these effect-sizes (even after potentially accounting for multiplicity in a testing procedure) can be severely biased. In this talk we explore this bias from a frequentist perspective. We show that were the bias known apriori one could build estimates that (potentially significantly) dominate our usual estimators, and bias corrected confidence intervals. In practice the bias will be unknown — we discuss a bootstrap procedure to estimate it. Unlike other proposals for debiasing estimates, our procedure implicitly adjusts for unknown dependence between the features. Finally, we empirically demonstrate the efficacy of our approach and relate it to ideas in empirical Bayes and compound decision theory. Keywords: High Dimensional, Selection Bias, Empirical Bayes

11. **Speaker**: Jonathan Taylor, Stanford University

    Title: Selective inference in regression

    Abstract: *We describe a framework for exact post-selection inference in regression problems. At the core of our framework is what we call selective inference which allows us to define selective Type I and II errors for hypothesis tests and selective coverage for intervals. Time allowing, several examples will be discussed including the LASSO, forward stepwise, change point detection. This is joint work with several students and collaborators.*

12. **Speaker**: Rob Tibshirani, Stanford University

    Title: Two novel applications of selective inference

Abstract: *I will discuss two applications of the recently developed theory of selective inference. The first application is to high dimensional regression and testing problems. Our idea is pre-cluster the features and extract a prototype from each cluster. Then we run the lasso or a multiple testing procedure on the prototypes. Using the theory of selective inference we derive exact p-values that account for the prototype selection, and any further selection in the model building. The second application is for the problem of comparing a data-derived predictor to an external one. Tibshirani and Efron (2007) proposed the "pre-validation" method for this problem. But this method does not guarantee proper type I error control. Using selective inference theory we derive an exact procedure for this problem.*

The first part is joint with Stephen Reid; the second part is joint with Sam Gross and Jonathan Taylor

13. **Speaker**: Bin Yu, University of California at Berkeley

Title: Stability

Abstract:

Reproducibility is imperative for any scientific discovery. More often than not, modern scientific findings rely on statistical analysis of high dimensional data. At a minimum, reproducibility manifests itself in stability of statistical results relative to reasonable perturbations to data and to the model used. Jacknife, bootstrap, and cross-validation are based on perturbations to data, while robust statistics methods deal with perturbations to models. In this talk, a case is made for the importance of stability in statistics. Firstly, we motivate the necessity of stability of interpretable encoding models for movie reconstruction from brain fMRI signals. Secondly, we find strong evidence in the literature to demonstrate the central role of stability in statistical inference. Thirdly, a smoothing parameter selector based on estimation stability (ES), ES-CV, is proposed for Lasso, in order to bring stability to bear on cross-validation (CV). ES-CV is then utilized in the encoding models to reduce the number of predictors by 60% with almost no loss (1.3%) of prediction performance across over 2,000 voxels. Last, a novel stability argument is seen to drive new results that shed light on the intriguing interactions between sample to sample variability and heavier tail error distribution (e.g. double-exponential) in high dimensional regression models with p predictors and n independent samples. In particular, when p/n belongs to (0.3, 1) and error is double-exponential, the Least Squares (LS) is a better estimator than the Least Absolute Deviation (LAD) estimator. (This talk draws materials from papers with S. Nishimoto, A. T. Vu, T. Naselaris, Y. Benjamini, J. L. Gallant, with C. Lim, and with N. El Karoui, D. Bean, P. Bickel, and C. Lim.)

14. **Speaker**: Ming Yuan, University of Wisconsin

Title: Distance Shrinkage and Euclidean Embedding via Regularized Kernel Estimation

Abstract:

Although recovering an Euclidean distance matrix from noisy observations is a common problem in practice, how well this could be done remains largely unknown. To fill in this void, we study a simple distance matrix estimate based upon the so-called regularized kernel estimate. We show that such an estimate can be characterized as simply applying a constant amount of shrinkage to all observed pairwise distances. This fact allows us to establish risk bounds for the estimate implying that the true distances can be estimated consistently in an average sense as the number of objects increases. In addition, such a characterization suggests an efficient algorithm to compute the distance matrix estimator, as an alternative to the usual second order cone programming known not to scale well for large problems.

15. **Speaker**: Tong Zhang, Rutgers University

Title: Some Recent Progress on Non-Convex Regularization Methods for Sparse Estimation

Abstract:

Non-convex regularization methods provide natural procedures for sparse recovery but are difficult to analyze. In this talk I will review some progress we have made in recent years. I will first show improved sparse recovery performance for local solutions of nonconvex formulations obtained via specialized numerical procedures. I will then present a unified framework describing the relationship of these local minima to the global minimizer of the underlying nonconvex formulation. In particular, we show that under suitable conditions, the global solution of nonconvex regularization leads to desirable recovery performance and it corresponds to the unique sparse local solution, which can be obtained via different numerical procedures. This unified view leads to a more satisfactory treatment of non-convex high dimensional sparse estimation procedures, and has led to additional numerical procedures for handling non-convex sparse regularization.

Collaborators: Cunhui Zhang, Han Liu, Zhaoran Wang, Tuo Zhao, Qiang Sun

16. **Speaker**: Ji Zhu, University of Michigan

Title: Detecting Overlapping Communities in Networks with Spectral Methods

Abstract:

Community detection is a fundamental problem in network analysis. In practice, it often occurs that the communities overlap, which makes the problem more challenging. Here we propose a general, flexible, and interpretable generative model for overlapping communities, which can be thought of as a generalization of the degree-corrected stochastic block model. We develop an efficient spectral algorithm for estimating the community memberships, which deals with the overlaps by employing the K-medians algorithm rather than the usual K-means for clustering in the spectral domain. We show that the algorithm is asymptotically consistent when networks are not too sparse and the overlaps between communities not too large. Numerical experiments on both simulated networks and many real social networks demonstrate that our method performs well compared to a number of benchmark methods for overlapping community detection. This is joint work with Yuan Zhang and Elizaveta Levina.