Workshop on Analysis of High-Dimensional and Functional Data in Honor of Peter Hall.
May 19-20, 2012
Location: Ballroom B, UC Davis Conference Center

Saturday, May 19, 2012

9:15-9:55: Peter Bickel, University of California, Berkeley.
More on Network Models

This is a report on joint work done with a number of collaborators whom I shall name in the talk. The topic is methods of inference for nonparametric models for unlabeled graphs introduced in the form introduced in B. and Chen (2009) PNAS. In particular we will discuss some ways for the rapid fitting of blockmodels and some types of bootstraps.

9:55-10:35: Jianqing Fan, Princeton University
Endogeneity in Ultra High Dimensional Models

Most papers on high-dimensional statistics are based on the assumption that none of the regressors are correlated with the regression error, namely, they are exogeneous. Yet, endogeneity arises easily in high-dimensional regression due to a large pool of regressors and this causes the inconsistency of the penalized least-squares methods. A necessary condition for model selection of a very general class of penalized regression methods is given, which allows us to prove formally the inconsistency claim. To cope with the possible endogeneity, we construct a novel penalized generalized method of moments (PGMM) criterion function and offer a new optimization algorithm. The PGMM is not a smooth function. To establish its asymptotic properties, we first study the model selection consistency and an oracle property for a general class of penalized regression methods. These results are then used to show that the PGMM possesses an oracle property even in presence of endogenous predictors, the solution is also near

global minimum under the over-identification assumption. Finally, we also show how the semi-parametric efficiency of estimation can be achieved via a two-step approach.

10:05-11:45: Craig Tracy, University of California, Davis.
Turbulent Liquid Crystals, KPZ Universality and the Asymmetric Simple Exclusion Process.

We report on (1) recent experimental work on stochastically growing interfaces and (2) new theoretical developments for the KPZ equation, a stochastic nonlinear PDE, and the closely related asymmetric simple exclusion process.

11:45-12:25: Iain Johnstone, Department of Statistics, Stanford University
Minimax Risk for Sparse Orthogonal Regression Revisited

Consider the classical problem of estimation of the mean of an n-variate normal distribution with identity covariance under squared error loss. We review exact and approximate models of sparsity for the mean vector and some of the associated minimax mean squared error properties. In particular, we describe some apparently unpublished results, developed for a forthcoming book, for the 'highly sparse' regime in which the number of non-zero components remains bounded as n increases.

14:00-14:40: Lawrence D. Brown, University of Pennsylvania
SURE Estimates for a Heteroscedastic Hierarchical Model

Hierarchical models provide an effective tool for combining information from similar resources and achieving partial pooling of inference. Since the seminal work by James and Stein (1961) and Stein (1962), shrinkage estimation has become one major focus for hierarchical models. For the homoscedastic normal model, it is well known that shrinkage estimators, especially the James- Stein estimator, have good risk properties. The heteroscedastic model, though more appropriate for a variety of practical applications, is less well studied, and it is unclear what types of shrinkage estimators are superior in terms of the risk. We propose a class of shrinkage estimators with shrinkage constant based on Stein's unbiased estimate of risk (SURE). We study asymptotic properties of various common estimators as the number of means to be estimated grows ($p \to \infty$). We establish that the SURE estimators are asymptotically optimal among the class of shrinkage estimators being studied, and some other commonly used estimators are not.

We then extend our construction to create a much broader class of semi-parametric shrinkage estimators and establish corresponding asymptotic optimality results. We emphasize that though the forms of our SURE estimators are partially motivated through a hierarchical

normal model, their optimality properties do not heavily depend on the distributional assumptions. We apply the methods to two real data sets and obtain encouraging results. This class of shrinkage estimators is related to estimators derived from regularized MLEs. The optimality property shows that it is possible to adaptively choose an optimal convex regularization function. [This is joint work with S. Kou and X. Xie.]

14:40-15:20: Tony Cai, University of Pennsylvania
A Framework for Shape Constrained Inference

A general non-asymptotic framework, which evaluates the performance of any procedure at individual functions, is introduced in the context of point estimation and confidence interval under shape constraints of monotonicity and convexity. This framework, which is significantly different from the conventional minimax theory for nonparametric function estimation, is also applicable to other problems in shape constrained inference. A benchmark is provided for the mean squared error of any estimate for each function in a similar way that Fisher Information depends on the unknown parameter in a regular parametric model. A local modulus of continuity is introduced and is shown to capture the difficulty of estimating individual functions. Data-driven procedures are introduced and are shown to perform uniformly within a constant factor of the ideal benchmark for every function in the function class. Such adaptivity is much stronger than adaptive minimaxity over a collection of large parameter spaces.

15:50-16:10: Jiashun Jin, Carnegie Mellon University.
Optimal Variable Selection by Graphlet Screening

Consider a linear model $Y = X\beta + z$, where $X = X_{n,p}$, $z \sim N(0, I_n)$, and $p \gg n \gg 1$. The signal vector β is sparse in the sense that most of its coordinates is 0. The goal is to separate the nonzero coordinates of β from the zero ones (i.e., variable selection). We assume the Gram matrix $G = X'X$ is sparse in the sense that each of its row has relatively few large coordinates. The Gram matrix naturally induces a sparse graph, which, by interacting with the signal sparsity, enables us to decompose the signals into many small-size isolated signal islands (if only we know where they are!). As a result, the original large-scale regression problem can be viewed as the aggregation of many small-size subproblems. Naturally, this Such insight motivates a two-stage Screen and Clean method which we call the *graphlet screening*, where we first identify candidates for such signal islands by multivariate screening, and then re-examine each candidate. The screening is guided by the sparse graph, so the computation cost is $O(p)$ (up to some multi-$\log(p)$ term) instead of $O(p^m)$ for some $m > 1$.
We develop a theoretic framework where we measure the errors of variable selection by the Hamming distance, and show that the graphlet screening attains the optimal rate of convergence.

Somewhat surprisingly, the well-known L^1 and L^0 penalization methods are non-optimal even in very simple settings and even when the tuning parameters are ideally set. These methods are non-optimal for they neglect the graph structure in the Gram matrix.

16:10-16:50: Raymond J. Carroll, Texas A&M University
Deconvolution and Classification

In a series of papers on Lidar data, magically good classification rates are claimed once data are deconvolved and a dimension reduction technique applied. The latter can certainly be useful, but it is not clear a priori that deconvolution is a good idea in this context. After all, deconvolution adds noise, and added noise leads to lower classification accuracy. I will give a more or less formal argument that in a closely related class of deconvolution problems, what statisticians call "Measurement Error Models", deconvolution typically leads to increased classification error rates. An empirical example in a more classical deconvolution context illustrates the results, and new methods and results relevant to the Lidar data will be discussed.

16:50-17:05: Debashis Paul, University of California, Davis
Functional and High-dimensional PCA when Samples are Correlated

Principal components analysis in the context of functional data and high-dimensional data involve typically different computational techniques. But the behavior of the standard estimates under both settings bear some similarity under a broad setting which includes observations that are weakly correlated across realizations.

17:05-17:20: Hans-Georg Müller, University of California, Davis
Stringing: From High-Dimensional to Functional Data

Stringing takes advantage of high dimensionality of data vectors by representing such data as discretized and noisy observations that are generated from a smooth stochastic process. Assuming that data vectors result from scrambling the original ordering of the observations, Stringing proceeds by reordering the components of the high-dimensional vectors, thereby transforming the vectors into functional data. This talk is based on joint work with Kehui Chen, Kun Chen, Jane-Ling Wang and Ping-Shi Wu.

Sunday, May 20

9:00-9:20: Byeong U. Park, Seoul National University
Structured Nonparametric Regression

It is widely admitted that structural modeling is a useful tool for circumventing the curse of dimensionality in nonparametric function estimation. In this talk I will introduce the idea of smooth backfitting as a powerful technique of fitting structured nonparametric regression models. I will discuss recent developments of the technique in several structured models.

9:20-9:40: Qiwei Yao, London School of Economics.
Factor Modelling for High-Dimensional Time Series: a Dimension-Reduction Approach

Following a brief survey on the factor models for multiple time series in econometrics, we introduce a statistical approach from the viewpoint of dimension reduction. Our method can handle nonstationary factors. However under stationary settings, the inference is simple in the sense that both the number of factors and the factor loadings are estimated in terms of an eigenanalysis for a non-negative definite matrix, and is therefore applicable when the dimension of time series is in the order of a few thousands. Asymptotic properties of the proposed method are investigated under two settings: (i) the sample size goes to infinity while the dimension of time series is fixed; and (ii) both the sample size and the dimension of time series go to infinity together. In particular, our estimators for zero-eigenvalues enjoy the faster convergence (or divergence) rates, which makes the estimation for the number of factors easier. Furthermore the estimation for both the number of factors and the factor loadings shows the so-called ``blessing of dimensionality" property. A two-step procedure is investigated when the factors are of different degrees of strength. Numerical illustration with both simulated and real data is also reported.

9:40-10:20: Ker-Chau Li, UCLA and Institute of Statistical Science, Academia Sinica
Statistical Issues of Nonlinearity and Interaction in Voluminous Genomic Data Analysis

Research in translational medicine requires the intensive use of statistical methods to help the integration of voluminous genomic data with patient outcomes. Such data include microarray gene expression data, array CGH for DNA copy number variation, SNP, transcription factor binding data, microRNA as well as next generation sequencing data of various kinds. While the separate analysis of each dataset has already generated many difficult problems, the co-investigation of multiple datasets poses even greater challenges. In this talk we will address the

recurrent issue of nonlinearity and complex patterns of interaction, which is often encountered in exploring complex biological data of huge dimensions. We will review some statistical methods and discuss their practical limitations.

10:50-11:05: Jiming Jiang, University of California, Davis
The EMAF and E-MS Algorithms: Model Selection with Incomplete Data

In this talk, I will present two computer-intensive strategies of model selection with incomplete data that we recently developed. The E-M algorithm is well-known for parameter estimation in the presence of missing data. On the other hand, model selection, as another key component of model identification, may also be viewed as parameter estimation, with the parameter being [the identification (ID) number of] the model and the parameter space being the (ID numbers of the) model space. From this point of view, it is intuitive to consider extension of the E-M to model selection in the presence of missing data. Our first strategy, called the EMAF algorithm, is motivated by a recently developed procedure for model selection, known as the adaptive fence (AF), that is incorporated with the E-M algorithm in the missing data situation. Our second strategy, called the E-MS algorithm, is a more direct extension of the E-M algorithm to model selection problems with missing data. This work is joint with Thuan Nguyen of the Oregon Health and Science University and J. Sunil Rao of the University of Miami.

11:05-11:20: Paul Baines, University of California, Davis.
Semi-parametric tests for equality of means of single-location and multiple-location processes.

We develop a framework for testing for equality of the means of multiple random processes, with an application to ozone data analysis. We consider two main settings: (I) testing for equality of multiple processes, and, (II) testing for equality across sets of random processes with spatial dependence. In the single location case, L^2 loss is used to measure the divergence between multiple mean functions in a semi-parametric regression framework. We develop a bootstrap test procedure that can be applied to temporally dependent data with ARMA-type serial correlation. Consistent estimators of the parameters can be achieved under both the null and a range of alternative hypothesis under mild regularity conditions. In order to test `local' features in the time series, we also propose a wavelet-domain version of the test that allows for testing equality at specific time-scales of interest. In the multiple location setting, a space-time model is proposed to capture both the temporal and spatial dependence. We propose a similar bootstrap test procedure, and wavelet domain counterpart, that can be efficiently applied in the space-time context. We conclude with some simulation studies and an application to testing for model-data agreement with multiple sources of ozone data.

11:20-11:35: Thomas Lee, University of California, Davis.
Generalized Fiducial Inference for Ultrahigh Dimensional Regression

In this talk we describe a modified version of Fisher's fiducial idea, termed generalized fiducial inference, and apply it to conduct statistical inference for ultrahigh dimensional regression. It is shown that our fiducial based inference procedure produces confidence intervals that possess asymptotically correct coverages. Simulation results demonstrate that our fiducial based procedure also enjoys promising empirical properties. This is joint work with Jan Hannig (UNC Chapel Hill) and Randy C. S. Lai (UC Davis).

11:35-12:15: Peter Hall, University of California, Davis, and University of Melbourne.
Distribution Approximation, Roth's Theorem, and Looking for Insects in Shipping Containers

Methods for distribution approximation, including the bootstrap, do not perform well when applied to lattice-valued data. For example, the inherent discreteness of lattice distributions confounds both the conventional normal approximation and the standard bootstrap when used to construct confidence intervals. However, in certain problems involving lattice-valued random variables, where more than one sample is involved, this difficulty can be overcome by ensuring that the ratios of sample sizes are quite irregular. For example, at least one of the ratios of sample sizes could be a reasonably good rational approximation to an irrational number. Results from number theory, in particular Roth's theorem (which applies to irrational numbers that are the roots of polynomials with rational coefficients), can be used to demonstrate theoretically the advantages of this approach. This project was motivated by a problem in risk analysis involving quarantine searches of shipping containers for insects and other environmental hazards, where confidence intervals for the sum of two binomial proportions are required.