Program

Thirteenth Annual Winter Workshop
High Dimensional Inference

Department of Statistics
University of Florida
January 14-15, 2011

Contents

Sponsors

The Graduate School, Info Tech and Department of Statistics

Organizing Committee

Michael Daniels
Malay Ghosh
Brett Presneli

Invited Speakers

F. Dubois Bowman, Emory University
Debashis Ghosh, Penn State University
Yang Feng, Columbia University
Gareth James, University of Southern California
Jeff Leek, Johns Hopkins University
Elizaveta Levina, University of Michigan
Robert Strawderman, Cornell University
Marten Wegkamp, Florida State University
Hao Helen Zhang, North Carolina State University
Hui Zou, University of Minnesota

Acknowledgements

The organizers thank the Department of Statistics staff, Robyn Crawford, Tina Greenly, Summer Rozear and Marilyn Saddler for their tremendous efforts in helping to set up this meeting and make it run smoothly.

Conference Schedule

All workshop sessions meet in the 209 Teaching Classroom, Emerson Alumni Hall

Thursday, January 13,2011

7:00-9:00 pm Reception Emerson Hall, Al and Judy Warrington Room

Friday, January 14,2011

8:00-8:45 am **Breakfast:** Continental

8:45-9:00 am Conference Welcome:

9:00-10:40 am **Sessioni:** **Chair:** *Michael Daniels, University of Florida*

Multiple Testing Procedures: Shrinkage and Clustering Aspects
Debashis Ghosh, Penn State University

Multiple Testing for Dependent Data
Jeff Leek, Johns Hopkins University

10:40-11:10 am **Break:** Refreshments

ll:10-Noon **Session 2:** ***Chair·.*** *Malay Ghosh, University of Florida*

Joint Estimation of Multiple Graphical Models
Elizaveta Levina, University of Michigan

Noon-12:30 pm **Conference photo on the Emerson Alumni Hall stairs to the second floor**

12:30-2:00 pm **Lunch:** (Gator Corner Dining Center)

2:00-3:40 pm **Session 3:** **Chair:** *Trevor Park, University of Florida*

Interesting Genes, Informative Genes and Microarray Classification
Hui Zou, University of Minnesota

Minimax Adaptive and Consistent Effective Rank Estimation of Matrices in High Dimensional Multivariate Response Regression
Marten Wegkamp, Florida State University

3:40-5:00 pm **Poster Session:** Emerson Alumni Hall, Presidents Room B (Refreshments)

All workshop sessions meet in the 209 Teaching Classroom, Emerson Alumni Hall

Saturday, January 15,2011

8:30-9:30 am **Breakfast:** Continental

9:30-11:10 am **Session 4:** **Chair:** *Brett Presneli, University of Florida*

MM Algorithms for Penalized Regression Problems
Robert Strawderman, Cornell University

Variable Selection using Adaptive Non-linear Interaction Structures in High Dimensions
Gareth James, University of Southern California

11:10-11:40 am **Break:** Refreshments

11:40-12:30 pm **Session 5:** **Chair:** *Kshitij Khare, University of Florida*

A Spatial Modeling Framework for Functional Neuroimaging Data
DuBois Bowman, Emory University

12:30-2:30 pm **Lunch:** Free time

2:30-4:10 pm **Session 6:** ***Chair*·.** *Hani Doss,University of Florida*

Nonparametric Independence Screening for Ultrahigh-dimensional Sparse Modeling
Yang, Feng, Columbia University

Automatic Structure Selection for Partially Linear Model
Hao Helen Zhang, North Carolina State University

Invited Talks

A Spatial Modeling Framework for Functional Neuroimaging Data

DuBois Bowman
Emory University

Functional magnetic resonance imaging (fMRI) studies have been used to characterize local properties of behavior-related neural activity and to investigate regional associations in brain activity. fMRI is also useful for determining differences in distributed neural processing between subgroups, e.g. schizophrenia patients and healthy controls. fMRI studies produce massive data sets that pose challenges for the development of appropriate statistical methods. The data from fMRI studies consist of 3-D movies for each subject, contain a large number of spatial locations (voxels) within each scan, and exhibit complex patterns of spatial and temporal correlations. In this talk, we develop modeling procedures that capture aspects of the spatial correlations between voxels and temporal correlations between repeated measures on each subject. Our methods provide a unified framework for the distinct objectives of detecting localized alterations in brain activity and determining associations between different brain regions. We demonstrate the applicability of our model approaches using data from addiction and psychiatric disorders.

Nonparametric Independence Screening for Ultrahigh-dimensional Sparse Modeling

Yang Feng
Columbia University

A variable screening procedure via correlation learning was proposed in Fan and Lv (2008) to reduce dimensionality in sparse ultra-high dimensional models. Even when the true model is linear, the marginal regression can be highly nonlinear. To address this issue, we further extend the correlation learning to marginal nonparametric learning. Our nonparametric independence screening, NIS, is a specific member of the sure independence screening. Several closely related variable screening procedures are proposed. It is shown that under some mild technical conditions, the proposed independence screening methods enjoy a sure screening property. The extent to which the dimensionality can be reduced by independence screening is also explicitly quantified. As a methodological extension, an iterative nonparametric independence screening (INIS) is also proposed to enhance the finite sample performance for fitting sparse additive models. The simulation results and a real data analysis demonstrate that the proposed procedure works well with moderate sample size and large dimension and performs better than competing methods.

This is a joint work with Jiangqing Fan, Princeton University, Rui Song, Corolado State University

Multiple Testing Procedures: Shrinkage and Clustering Aspects

Debashis Ghosh
Penn State University

In the analysis of high-dimensional data, one very common approach has to approach problems from a multiple testing perspective. This has led to a renewed interest in the consideration of multiple comparisons problem, of which there has been an emphasis on procedures that control the false discovery rate. In this talk, we will consider two aspects of the problem. First, we will discuss shrinkage approaches to the multiple testing problem. Advantages of this approach will include insensitivity to dependence, a common problem in genomic settings. Second, we will discuss a reinterpretation of the celebrated Benj amini-Hochberg procedure in terms of a type of clustering problem. This reinterpretation will allow for the development of variations that of the B-H method that have a Bayesian justification as a type of empirical null hypothesis in the sense of Efron. Simulated and real-data examples will be used throughout the talk for illustration.

Variable Selection using Adaptive Non-linear Interaction Structures in High Dimensions

Gareth James
University of Southern California

Numerous penalization based methods have been proposed for fitting a traditional linear regression model in which the number of predictors, p, is large relative to the number of observations, n. Most of these approaches assume sparsity in the underlying coefficients and perform some form of variable selection. Recently, some of this work has been extended to non-linear additive regression models. However, in many contexts one wishes to allow for the possibility of interactions among the predictors. This poses serious statistical and computational difficulties when p is large, as the number of candidate interaction terms is of order p squared.

In this talk I will introduce a new approach, "Variable selection using Adaptive Non-linear Interaction Structures in High dimensions" (VANISH), that is based on a penalized least squares criterion and is designed for high dimensional non-linear problems. The criterion is convex and enforces the heredity constraint, in other words if an interaction term is added to the model, then the corresponding main effects are automatically included. Detailed simulation and real world results will also be provided, demonstrating that VANISH is computationally efficient and can be applied to non-linear models involving thousands of terms while producing superior predictive performance over other approaches.

This is joint work with Peter Radchenko.

Multiple Testing for Dependent Data

Jeff Leek
Johns Hopkins University

We develop a framework for performing large-scale significance testing in the presence of arbitrarily strong dependence. We derive a low-dimensional set of random vectors, called a dependence kernel, that fully captures the dependence structure in an observed high-dimensional data set. This represents a surprising reversal of the curse of dimensionality in high-dimensional hypothesis testing. We then show theoretically that conditioning on a dependence kernel is sufficient to render statistical tests independent regardless of the level of dependence in the observed data. This solution to multiple testing dependence has implications in a variety of popular testing problems, such as in gene expression studies, brain imaging, and spatial epidemiology.

Joint Estimation of Multiple Graphical Models

Elizaveta Levina
University of Michigan

Graphical models are a popular tool for exploring dependence relationships between variables. Here we develop methods for simultaneously estimating graphical models on several classes that share some common dependence structure, but also have individual differences, as may happen with different subtypes of the same disease. Joint estimation, which allows for sharing information across classes, results in more accurate estimation than fitting each class separately. We develop joint estimati onn methods for both the Gaussian graphical models and the Ising models for discrete data, establish their asymptotic properties, and illustrate the methodology on examples of links between university webpages and the U.S. Senate voting records.

Joint work with Jian Guo, George Michailidis, and Ji Zhu.

MM Algorithms for Penalized Regression Problems

Rob Strawderman
Cornell University

The study of penalized objective functions in connection with high dimensional regression problems has increased dramatically in recent years. In problems where the goal is to simultaneously select variables and estimate regression coefficients in settings of relatively high dimension, the use of nondifferentiable penalty functions, such as the convex LÍ norm and concave SCAD and MCP penalty functions, have been proposed for this purpose. In the case of convex and concave penalty functions, substantial attention has been paid to the understanding of statistical behavior (e.g., oracle properties) of the resulting estimators; however, in the case of concave penalties, considerably less attention has been paid to the development of general algorithms with known convergence properties.

In this talk, I describe a general class of majorization minimization (MM) algorithms for this purpose. The resulting (M)inimization by (I)terated (S)oft (T)hresholding (ie, MIST) algorithm relies on iterated soft-thresholding, implemented componentwise, and is characterized by very fast, stable parameter updating that avoids the need to invert high-dimensional or otherwise numerically unstable matrices. We summarize the local convergence properties for this new class of algorithms, introduce some interesting variations designed specifically for penalized GLM regression problems, and evaluate performance for "raw" and "accelerated" versions of these methods. As an illustration, we propose a new algorithm for fitting Cox regression models subject to the MCP penalization recently introduced in Zhang (2008, 2010) and use it to analyze the relationship between gene expression data and survival in lymphoma patients.

This talk is based on joint work with Liz Schifano and Marty Wells (Electronic Journal of Statistics, 2010).

Minimax Adaptive and Consistent Effective Rank Estimation Of Matrices in High Dimensional Multivariate Response Regression

Marten Wegkamp
Florida State University

We introduce a new criterion, the Rank Selection Criterion (RSC), for selecting the optimal reduced rank estimator of the coefficient matrix in multivariate response regression models. The corresponding RSC estimator minimizes the Frobenius norm of the fit plus a regularization term proportional to the number of parameters in the reduced rank model.

The rank of the RSC estimator provides a consistent estimator of the rank of the coefficient matrix; in general the rank of our estimator is a consistent estimate of the effective rank, which we define as the number of singular values of the target matrix that are appropriately large. The consistency results are valid not only in the classic asymptotic regime, when n, the number of responses, and p, the number of predictors, stay bounded, and m, the number of observations, grows, but also when either, or both, n and p grow, possibly much faster than m.

We establish minimax optimal bounds on the mean squared errors of our estimators. Our finite sample performance bounds for the RSC estimator show that it achieves the optimal balance between the approximation error and the penalty term.

Furthermore, our procedure has very low computational complexity, linear in the number of candidate models, making it particularly appealing for large scale problems. We contrast our estimator with the nuclear norm penalized least squares estimator (NNP), which has an inherently higher computational complexity than RSC. We show that NNP has estimation properties similar to those of RSC, albeit under stronger conditions. However, it is not as parsimonious as RSC. We offer a simple correction of the NNP estimator which leads to consistent rank estimation.

We verify and illustrate our theoretical findings via an extensive simulation study.

This is joint work with Florentina Bunea and Yiyuan She.

Automatic Structure Selection for Partially Linear Models

Hao Helen Zhang
North Carolina State University

Partially linear models provide good compromises between linear and nonparametric models. However, given a large number of covariates, it is often difficult to objectively decide which covariates are linear and which are nonlinear. Common approaches include hypothesis testing methods and screening procedures based on univariate scatter plots. These methods are useful in practice; however, testing the linearity of multiple functions for large dimensional data is both theoretically and practically challenging, and visual screening methods are kind of ad hoc. In this work, we tackle this structure selection problem in partially linear models from the perspective of model selection. A unified estimation and selection framework is proposed and studied. The new estimator can automatically determine the linearity or nonlinearity for all covariates and at the same time consistently estimate the underlying regression functions. Both theoretical and numerical properties of the resulting estimators are presented.

Interesting Genes, Informative Genes and Microarray Classification

Hui Zou
University of Minnesota

This talk focuses on classification with microarray gene expression data which is one of the leading examples behind the recent surge of interests in high dimensional data analysis. It is now known that feature selection is crucial and often necessary in high dimensional classification problems. In the current literature there are two main approaches to gene (feature) selection in microarray classification. The first approach is basically a two-stage procedure. First, one performs large-scale multiple hypothesis testing to find the “interesting genes". After reporting a subset of “interesting genes", one then builds a discriminative function only using these “interesting genes". The second approach integrates feature selection and discriminant analysis so that one can directly target on finding the “informative genes" that are responsible to classifying the binary response.

Consider feature selection under the linear discriminant analysis model. We carefully examine the fundamental difference between the ”interesting genes" and “informative genes" and show that the first approach could lead to a fraud classification rule. On the other hand, the computational challenges in deriving sparse discriminant analysis rules force people to use the independent rules that basically ignore the correlations among features in the classification procedure. In this talk we argue that such compromise is not necessary and present a new sparse discriminant analysis method that has the same computation complexity as sparse regularized least squares. We further provide theoretical justifications for our method. Our analysis concerns the scenario where both n (sample size) and p (dimension) can go to infinity and p can grow faster than any polynomial order of n. We show that, under reasonable regularity conditions, the proposed sparse LDA classifier can simultaneously achieve two goals: (1) consistently identify the subset of informative predictors; (2) consistently estimate the Bayes classification direction. Numerical examples will be presented as well.

Poster Abstracts

Simultaneous Inference For The Mean Function Of Dense Longitudinal Data

Guanqun Cao
Michigan State University

We propose a polynomial spline estimator for the mean function of dense longitudinal data together with a simultaneous confidence band which is asymptotically correct. In addition, the spline estimator and its accompanying confidence band enjoy semiparametric efficiency in the sense that they are asymptotically the same as if all random trajectories are observed entirely and without errors, a view taken in Ferraty and Vieu (2006). We also build a confidence band to test the difference of two groups of functional data. Simulation experiments provide strong evidence that corroborates the asymptotic theory while computing is efficient. The confidence band procedure is illustrated by analyzing the near infrared spectroscopy data.

This is a joint work with Lijian Yang and David Todem.

Modeling and Forecasting the Time Series of Treasury Bond Yield Curves

Cong Feng
University of Georgia

A novel method is proposed for forecasting a time series of smooth curves, using functional principal component analysis in combination with time series modeling and forecasting the scores. In this way, we can achieve the smoothing, dimension reduction and prediction at the same time with the expedient computation. The research problem is motivated by the demand to forecast the time series of economic functions, such as Treasury bond yield curves. Extensive simulation studies have been carried out to compare the prediction accuracy of our method with other competitor's methods. The proposed methodology is applied to forecasting the yield curves of UK government bond.

HORSES: Hexagonal Operator for Regression with Shrinkage and Equality Selection

Woncheol Jang
University of Georgia

Identifying homogeneous subgroups of variables can be challenging in high dimensional data analysis with highly correlated predictors. We propose a new method called HORSES (Hexagonal Operator for Regression with Shrinkage and Equality Selection) which simultaneously selects positively or spatially correlated variables and identifies them as predictive clusters. HORSES can be implemented via a constrained least-squares problem with regularization involving a linear combination of an LÍ penalty for the coefficients and another LÍ penalty for pairwise differences of the coefficients. The penalty function encourages grouping of positively correlated predictors with a sparsity solution. We show via simulation that the proposed method outperforms other variable selection methods in terms of prediction error and model complexity. The technique is demonstrated on two data sets: soil analysis from Appalachia (relatively small scale problem) and a functional magnetic resonance imaging study (very large scale problem), showing the exibility of the methodology.

This is joint work with Johan Lim, Ji Meng Loh, Nicole Lazar and Donghyun Yu,

Survival Prediction with Simultaneous Shrinkage and Grouping Prior

Kyu Ha Lee
University of Missouri Columbia

Variable selection for high dimensional data has recently received a great deal of attention. However, due to the complex structure of the likelihood, only limited developments have been made for time-to-event data where censoring is present. In this paper, we propose several Bayesian variable selection schemes for Bayesian semiparametric survival models for right-censored survival data. Special shrinkage priors on the coe_cients corresponding to the predictor variables are used to handle cases when the explanatory variables are of very high-dimension. As grouping is natural in microarray studies where often the genes belonging to the same biological pathways are grouped together and perform as a single unit, we extend the idea of the shrinkage prior such that it can incorporate any group structure among the covariates. The shrinkage priors are obtained through a scale mixture representation of Normal and Gamma distributions. Our proposed variable selection priors correspond to the well known frequentist the lasso, elastic-net, group lasso, and fused lasso penalty. The likelihood function is constructed based on the Cox proportional hazards model framework, where the cumulative baseline hazard function is estimated nonparametrically using a discrete gamma process. We assign priors on the tuning parameters of the shrinkage priors and adaptively control the sparsity of our models. The primary use of the proposed models is to identify the important covariates relating to the survival curves. To implement our methodologies, we have developed fast Markov chain Monte Carlo algorithms with adaptive jumping rule. We have successfully applied our methods on simulated data sets and real microarray data sets which contain right censored survival time. The performance of our Bayesian variable selection models compared with other standard competing methods is also provided to demonstrate the superiority of our methods.

This is a joint work with Sounak Chakraborty, and Jianguo Sun.

Estimation and Algorithm for Joint Linkage and Linkage Disequilibrium Analysis in Lamily Data

Jiangtao Luo
University of Nebraska Medical Center

The joint linkage and linkage disequilibrium for family data are studied in this paper. The algorithm for the estimates of the parameters is given here. We also show that convergence of the algorithm has some nice properties.

Sufficient Dimension Reduction Based on the Hellinger Integral

Qin Wang
Virginia Commonwealth University

Sufficient dimension reduction provides a useful tool to study the dependence between a response Y and a multidimensional regressor X. A new formulation is proposed here based on the Hellinger integral of order two — and so jointly local in (X,Y) — together with an efficient estimation algorithm. The link between χ^2-divergence and dimension reduction subspaces is the key to our approach, which has a number of strengths. It requires minimal (essentially, just existence) assumptions. Relative to other forward regression based methods, it is faster, allowing larger problems to be tackled, more general, multidimensional (discrete, continuous or mixed) Y as well as X being allowed, and includes a sparse version enabling variable selection. Finally, it unifies three existing methods, each being shown to be equivalent to adopting suitably weighted forms of the Hellinger integral.

This is a joint work with Dr. Xiangrong Yin at University of Georgia and Dr. Frank Critchley at Open University (UK).

A Fast Inference On Default Probability For Credit Rating Via Generalized Additive Model

Shuzhuan Zheng
Michigan State University

Generalized additive models (GAM) is excessively applied in bioscience and finance with non-Gaussian responses including binary and count data. Due to its high dimensional feature, the computational expedient inference for estimation and variable selection remains an open problem. We propose spline-backfitted kernel (SBK) estimator for the GAM time series data with oracle efficiency. It is not only theoretically reliable but also computational expedient, which meets the needs of high-dimensional data analysis. In addition, we constructed the simultaneous confidence band for each component functions, in order to test their global shapes. Particularly, a BIC criterion has been developed in the context of GAM, which has appealing performances for variable selection in GAM. The simulation study strongly corroborates with the asymptotics we developed, and our method is applied for credit rating of the listed companies in Japan.

This is a joint worth with Rong Liu, Lijian Yang and Lifeng Wang