**Tutorial: Statistical Analysis of Financial Data with R**

**Speaker: David Matteson, Cornell University**

**Abstract:** We will introduce statistical concepts and models for finance and illustrate financial data analysis using R. Topics will include financial asset returns, multivariate distributions and copulas, ARIMA models, volatility models, regression with time series data, vector autoregression and functional time series analysis. Basic knowledge of statistics and probability, matrices and linear algebra, and calculus will be assumed, while prior knowledge of finance is not necessary.

**Lecture 1**
Financial Assets and Return
Fitting Univariate and Multivariate t-distributions
Copula models

**Lecture 2**
Stationarity and Autocorrelation
White Noise Models
Autoregressive Moving Average (ARMA) Models

**Lecture 3**
Generalized Autoregressive Conditional Heteroskedasticity (GARCH) Models
Regression with Time Series Errors

**Lecture 4**
Vector autoregression
Functional time series

**Title: R-Estimation for Asymmetric Independent Component Analysis**

**Speaker: Mark Hallin, ECARES, Universite libre de Bruxelles, Belgium**

**Abstract:** Independent Component Analysis (ICA) recently has attracted much attention in the statistical literature as an appealing alternative to elliptical models. Whereas k-dimensional elliptical densities depend on one single unspecified radial density, however, k-dimensional independent component distributions involve k unspecified component densities. In practice, for given sample size n and dimension k, this makes the statistical analysis much harder. We focus here on the estimation, from an independent sample, of the mixing/demixing matrix of the model. Traditional methods (FOBI, Kernel-ICA, FastICA) mainly originate from the engineering literature. Their consistency requires moment conditions, they are poorly robust, and do not achieve any type of asymptotic efficiency. When based on robust scatter matrices, the so-called two-scatter methods enjoy better robustness features, but their optimality properties remain unclear. The "standard semiparametric" approach achieves semiparametric efficiency, but requires the estimation of the densities of the k unobserved independent components. As a reaction, an efficient (signed-)rank-based approach has been proposed by Ilmonen and Paindaveine (2011) for the case of symmetric component densities. The performance of their estimators is quite good, but they unfortunately fail to be root-n consistent as soon as one of the component densities violates the symmetry assumption. In this paper, using ranks rather than signed ranks, we extend their approach to the asymmetric case and propose a one-step R-estimator for ICA mixing matrices. The finite-sample performances of those estimators are investigated and compared to those of existing methods under moderately large sample sizes. Particularly good performances are obtained from a version involving data-driven scores taking into account the skewness and kurtosis of residuals. Finally, we show, by an empirical exercise, that our methods also may provide excellent results in a context such as image analysis, where the basic assumptions of ICA are quite unlikely to hold.

Based on joint work with Chintan Mehta (Yale School of Public Health)

**Title: Testing High Dimensional Mean Under Sparsity**

**Speaker: Xianyang Zhang, Texas A&M University**

**Abstract:** Motivated by the likelihood ratio test under the Gaussian assumption, we develop a maximum sum-of-squares test for conducting hypothesis testing on high dimensional mean vector. The proposed test which incorporates the dependence among the variables is designed to ease the computational burden and to maximize the asymptotic power in the likelihood ratio test. A simulation-based approach is developed to approximate the sampling distribution of the test statistic. The validity of the testing procedure is justified under both the null and alternative hypotheses. We further extend the main results to testing the quality of two mean vectors without imposing the equal covariance assumption. Numerical results suggest that the proposed test can be more powerful than some existing alternatives.

**Title: Multi-scale Factor Analysis of High Dimensional Time Series**

**Speaker: Hernando Ombao, University of California at Irvine**

**Abstract**: We consider the problem of modeling dependence between components of a high-dimensional time series. In this paper, we put this in the context of investigating connectivity between different regions of the brain using functional magnetic resonance imaging (fMRI) data. The primary challenge here is the high dimensionality of fMRI data which consists of time series recorded across over a hundred thousand voxels in the entire brain volume. The common approach is to first divide the brain volume into about one anatomically-determined regions; summarize brain activity in each region by taking the spatially-averaged time series; and finally compute cross-correlation or coherence between region-averaged fMRI time series. The key limitation of this approach is that region-specific brain activity is summarized by a single time series which is predefined to be the average. First, a single time series cannot sufficiently represent complex processes in a brain region. Second, the method for determining the optimal signal summaries should be data-adaptive rather than preselected to be of some form (e.g., the mean).

In this paper, we develop the multi-scale factor analysis (MSFA). The first stage in our proposed framework is to reduce dimensionality by applying principal components analysis (PCA) within each anatomically parcellated region of interest (ROI). This dimension reduction approach is summarizes localized activity by selecting, in a data-adaptive manner, the components series that best explain localized (within-ROI) variance. The second step is to model connectivity between the ROIs or system networks by computing the dependence measure between components series extracted from each of the ROIs. In this paper, we measure connectivity by the RV-coefficient which is essentially the generalized correlation between a pair of multi-dimensional time series (in this case, the components series at each pair of ROIs or networks). The proposed procedure simultaneously accomplishes the following desired goals: (1.) it gives a representation of localized brain activity that is an optimal solution to PCA criterion which objectively selects components that maximize the explained variation within each ROI; (2.) it captures the multi-scale dependence structure at both local (within-ROI) level and global (between ROIs and between networks) level; and (3.) it achieves dimension reduction therefore can efficiently handle the massive fMRI data. The novel MSFA approach is used to study functional connectivity in resting-state fMRI data, which reveals interesting modular and hierarchical structure of human brain networks.

This is collaborative work between Dr. Chee-Ming Ting and Dr. Hussain Salleh of the Universiti Teknologi Malaysia and Yuxiao Wang of the University California at Irvine.

**Title: TBA**

**Speaker: Suhasini Subba Rao, Texas A&M University**

**Abstract:** TBA

**Title: Inference in high-dimensional varying coefficient models**

**Speaker: Mladen Kolar, The University of Chicago, Booth School of Business**

**Abstract:** Varying coefficient models have been successfully applied in a number of scientific areas ranging from economics and finance to biological and medical science. Varying coefficient models allow for flexible, yet interpretable, modeling when traditional parametric models are too rigid to explain heterogeneity of sub-populations collected. Currently, as a result of technological advances, scientists are collecting large amounts of high-dimensional data from complex systems which require new analysis techniques. We focus on the high-dimensional linear varying-coefficient model and develop a novel procedure for estimating the coefficient functions in the model based on penalized local linear smoothing. Our procedure works for regimes which allow the number of explanatory variables to be much larger than the sample size, under arbitrary heteroscedasticity in residuals, and is robust to model misspecification as long as the model can be approximated by a sparse model. We further derive an asymptotic distribution for the normalized maximum deviation of the estimated coefficient function from the true coefficient function. This result can be used to test hypotheses about a particular coefficient function of interest, for example, whether the coefficient function is constant, as well as construct confidence bands for covering the true coefficient function. Construction of the uniform confidence bands relies on a double selection technique that guards against omitted variable bias arising from potential model selection mistakes. We demonstrate how these results can be used to make inference in high-dimensional dynamic graphical models.

Joint work with Damian Kozbur.

**Title: A non-parametric method for joint association analysis of sequencing and Imaging data**

**Speaker: Hao Wang, Michigan State University**

**Abstract:** The rapid development of whole genome sequence (WGS) technology coupled with magnetic resonance image (MRI) data mandates the development of analytical methods that are capable of utilizing both WGS and MRI data to identify predictive biomarkers associated with neurodegenerative diseases, such as Alzheimer's disease. The rich WGS/MRI data, however, brings the issue of "the curse of dimensionality" due to the vast number of sequencing variants and brain surface vertexes. In this work, we tackled the dimensionality issue of MRI data through a stacked denoising autoencoder (SDA) constructed using the deep learning algorithm, which reduces the dimensionality and maintains the majority of the information. For the WGS data, we use a weighted identity-by-state (IBS) kernel to aggregate information over multiple sequencing variants in a genetic region. A weighted U statistic is then used to evaluate the joint association of both imaging and sequencing data with the phenotype of interest. We show that our method maintains the correct type 1 error rate, while achieving high statistical power in comparison to methods using either sequencing or image data alone. To illustrate our approach, we apply the proposed method to the sequencing and image data from the Alzheimer's disease Neuroimaging Initiative.

**Title: Regression Models for Multivariate Spatially or Longitudinally Correlated Functional Data**

**Speaker: Jeffrey Morris, The University of Texas MD Anderson Cancer Center**

**Abstract:** In this talk, I will describe a series of regression modeling strategies that can be used for multivariate spatially- or longitudinally correlated functional data. Intrafunctional correlation is handled through basis function modeling, while interfunctional correlation is captured by one of three approaches: (1) parametric or nonparametric random effect functions, (2) separable or non-separable spatial (or temporal) inter-functional processes, or (3) tensor-basis function modeling. I will describe these general approaches and illustrate them on a series of complex, high-dimensional, spatially and longitudinally correlated functional data sets coming from strain tensor data from a glaucoma study and event-related potential data from a smoking cessation study. Methods for performing model selection and choosing which basis functions best fit the data will also be described.

**Title: Collective estimation of multiple bivariate density functions with application to angular-sampling-based protein loop modeling**

**Speaker: Lan Zhou, Texas A&M University**

**Abstract:** We develop a method for simultaneous estimation of density functions for a collection of populations of protein backbone angle pairs using a data-driven, shared basis that is constructed by bivariate spline functions defined on a triangulation of the bivariate domain. The circular nature of angular data is taken into account by imposing appropriate smoothness constraints across boundaries of the triangles. Maximum penalized likelihood is used to fit the model and an alternating blockwise Newton-type algorithm is developed for computation. A simulation study shows that the collective estimation approach is statistically more efficient than estimating the densities individually. The proposed method was used to estimate neighbor-dependent distributions of protein backbone dihedral angles (i.e., Ramachandran distributions). The estimated distributions were applied to protein loop modeling, one of the most challenging open problems in protein structure prediction, by feeding them into an angular-sampling-based loop structure prediction framework. Our estimated distributions compared favorably to the Ramachandran distributions estimated by fitting a hierarchical Dirichlet process model; and in particular, our distributions showed significant improvements on the hard cases where existing methods do not work well.

This is a joint work with Mehdi Maadooliat, Seyed Morteza Najibi, Xin Gao and Jianhua Huang.

**Title: On high-dimensional robust regression and the bootstrap**

**Speaker: Noureddine El Karoui, University of California, Berkeley**

**Abstract:** In this talk, I'll briefly review some of my recent work on high-dimensional robust regression.

Very interestingly, ideas connected to the analysis of robust regression estimators in high-dimension gives insight into the performance of the bootstrap. I will discuss a number of surprising results, including the fact that two equally intuitive (in low-dimension) bootstraps perform very differently in high-dimension: one leads to extremely conservative confidence intervals, the other to anti-conservative confidence intervals.

Time permitting, I will discuss many more problems with the bootstrap in moderate to high-dimension.

**Title: Penalized matrix decompositions in sparse high-dimensional multivariate analysis**

**Speaker: Irina Gaynanova, Texas A&M University**

**Abstract:** The overwhelming growth of the datasets with the number of measurements p being larger than the number of samples n generated a lot of interest in developing high-dimensional analogs of traditional multivariate analysis techniques. A large body of work focused on the problem of sparse estimation of population eigenvectors, leading to sparse principal component analysis, sparse linear discriminant analysis and sparse canonical correlation analysis. A common approach is to consider a penalized loss estimation procedure, where the loss function exploits the definition of the eigenvector and the penalty function promotes sparsity in the solution. Due to the existence of equivalent definitions for the eigenvectors, this approach gives rise to multiple estimation procedures even if the same penalty function is used. The benefits of using one procedure versus the other are not well-understood, and as a result, current literature lacks guidance on which estimation procedure is best in terms of theoretical guarantees, computational complexity and empirical performance. In this talk, I will provide a partial answer to this question by drawing the connection between the penalized matrix decompositions and the modified power method for finding eigenvectors and for finding singular vectors. This connection allows to demonstrate the superiority of modified power method for singular value decomposition over eigendecomposition when p>>n in terms of both variable selection performance and computational complexity. Both simulation studies and real data analysis support this finding. I will conclude by providing a unified framework for sparse estimation of population eigenvectors, and identifying directions for future research.

**Title: TBA**

**Speaker: Tyler McCormick, University of Washington**

**Abstract:** TBA

**Title: Community detection in stochastic block models with unknown number of communities**

**Speaker: Anirban Bhattacharya, Texas A&M University**

**Abstract:** A fundamental problem in network analysis is clustering the nodes into groups which share a similar connectivity pattern, called community detection. Existing community detection algorithms assume the knowledge of the number of clusters or estimate it apriori using classical selection criteria, followed by finding the clustering configuration. Such two stage procedures do not account for uncertainty in the first stage leading to inaccurate estimation of the clustering configurations. We address this problem by developing a coherent probabilistic framework for simultaneous estimation of the number of clusters and the clustering configurations. In addition, we propose efficient Markov chain Monte Carlo algorithms with empirical guarantees of rapid mixing and convergence. The methodology is shown to outperform well-known competitors in a variety of simulated examples and real network data.

**Title: Spectral analysis of linear time series in high dimensions**

**Speaker: Debashis Paul (Department of Statistics, University of California, Davis)**

**Abstract:** We study the spectral behavior of a class of p-dimensional stationary linear processes in high-dimensional regimes, namely when p/n converges to some finite positive constant. The class of linear processes under consideration is determined by a sequence of real or complex random innovations with i.i.d. entries possessing zero mean, unit variance and finite fourth moments, and that the coefficient matrices in the linear process representation are Hermitian and simultaneously diagonalizable. The object of our study is the symmetrized sample autocovariance matrix for any given lag order. We show that the empirical distribution of the eigenvalues of a these symmetrized autocovariance matrix converges to a nonrandom limiting distribution characterized by a system of nonlinear equations uniquely describing its Stieltjes transform. We also consider the problem estimating the spectra of the coefficients and autocovariance matrices. The estimation procedure involves a nonlinear optimization procedure and utilizes the Stieltjes transforms of the symmetrized autocovariance matrices. We discuss some potential applications.

This is based on joint work with Haoyang Liu and Alexander Aue.

**Title: MOCCA: a primal/dual algorithm for nonconvex composite functions with applications to CT imaging**

**Speaker: Rina Foygel Barber, University of Chicago**

**Abstract:** Many optimization problems arising in high-dimensional statistics decompose naturally into a sum of several terms, where the individual terms are relatively simple but the composite objective function can only be optimized with iterative algorithms. Specifically, we are interested in optimization problems of the form F(Kx) + G(x), where K is a fixed linear transformation, while F and G are functions that may be nonconvex and/or nondifferentiable. In particular, if either of the terms are nonconvex, existing alternating minimization techniques may fail to converge; other types of existing approaches may instead be unable to handle nondifferentiability. We propose the MOCCA (mirrored convex/concave) algorithm, a primal/dual optimization approach that takes local convex approximation to each term at every iteration. Inspired by optimization problems arising in computed tomography (CT) imaging, this algorithm can handle a range of nonconvex composite optimization problems, and offers theoretical guarantees for convergence when the overall problem is approximately convex (that is, any concavity in one term is balanced out by convexity in the other term). Empirical results show fast convergence for several structured signal recovery problems.