Poster Abstracts

Growth Dynamics of Acceleration: examining whether and how prenatal environmental exposures of BPA and phthalates in women are associated with infant growth dynamics of acceleration in BMI and height.

Baek, Jonggyu (jongguri@umich.edu)
University of Michigan

Type: Poster

Abstract. This is the first study to apply an explicit modeling on growth dynamics of acceleration for infants in 0-3 years. The child period of 0-3 years is of the greatest importance for cognitive, emotional and social development. The near ubiquitous environmental exposure to endocrine disrupting compounds (EDCs) such as BPA, phthalates and heavy metals experienced among children and women during sensitive developmental periods are potentially associated with the potential developmental and reproductive effects. To investigate whether and how prenatal environmental exposures of BPA and phthalates in women in the first three trimesters of pregnancy are associated with infant growth dynamics of BMI and height, the semi-parametric stochastic velocity model (SSVM) with Ornstein-Uhlenbeck (OU) process as a prior for the rate function is used in a Bayesian framework. Furthermore, inter-relationship of growth acceleration between BMI and height was investigated in a bivariate SSVM.

Keywords: Semiparametric stochastic velocity model; gaussian process; Ornstein-Uhlenbeck process.

On the Optimal and Adaptive Robust Estimation of the Heavy-tail Exponent

Bhattacharya, Shrijita (shrijita@umich.edu)
University of Michigan

Type: Poster

Abstract. We consider the problem of robust estimation of the tail index of a regularly varying distribution. We propose a robust version of the Hill estimator, which is immune to extreme outliers in the data. The estimator is shown to be asymptotically efficient among a class of estimators with given strict upper breakdown point. We also develop three automatic procedures for the selection of a tuning parameter controlling the degree of robustness. These methods lead to estimators that adapt to the degree of contamination in the extremes. Simulation studies to evaluate the finite sample properties of these estimators together with their computational costs are given. This work is motivated and applied to the online detection of heavy hitters in fast multi-gigabit network traffic streams.

Keywords: Robust, Hill, Traffic Streams.

Spectral Clustering for Dynamic Stochastic Block Model

Bhattacharyya, Sharmodeep (bhattash@science.oregonstate.edu)
Oregon State University

Type: Poster

Abstract. One of the most common and crucial aspect of many network data sets is the dependence of network link structure on time or other attributes. This has led the researchers to study dynamic, time-evolving networks. In this work, we consider the problem of finding a common clustering structure in time-varying networks. We consider two simple extension of spectral clustering methods to dynamic settings and give theoretical justification that the spectral clustering methods produce consistent community detection for such dynamic networks. We also propose an extension of the static version of nonparametric latent variable models to the dynamic setting and use a special case of the model to justify the spectral clustering methods. We show the validity of the theoretical results via simulations too and apply the clustering methods to real-world dynamic biological networks.
Keywords: Networks; Spectral Clustering; Dynamic Networks; Squared Adjacency Matrix; Stochastic Block Model.

An SDP Relaxation for Structured Blockmodels

Choi, David (davidch@andrew.cmu.edu)
Carnegie Mellon University

Type: Poster

Abstract. Semidefinite programs (SDP) have recently been developed for community detection, which is a special case of the stochastic blockmodel. Here, we develop an SDP for general blockmodels, in which the parameter matrix P of within- and between- class connection probabilities is assumed to be known, but need not correspond to community detection. Instead, P can encode a wide variety of network structures, including hierarchy, latent distances, or time-varying blockmodels. Given knowledge of P, the SDP is provably able to estimate the unknown class labels.
Keywords: network data; stochastic blockmodel; semidefinite programming; time-varying blockmodel.

An Exact, Non-Parametric Test for Treatment Effects on Networks

Fredrickson, Mark (fredric3@illinois.edu)
Univeristy of Illinois, Urbana-Champaign

Type: Poster

Abstract. Recent years have seen renewed interest in "design based" approaches to analyzing randomized controlled trials (RCTs). Design based inference, or randomization inference, relies on the known treatment assignment mechanism of a RCT to derive tests and estimators, rather than specifying parametric forms for outcomes. One area where randomization inference has been less developed is network analysis. Typical graph models, such as exponential random graphs, posit a parametric distribution governing the generation of edges between nodes. Just as Fisher's exact test provides an non-parametric alternative to logistic regression, in this paper we propose a non-parametric test of network formation as an alternative to parametric graph models. Additionally, we extend Rosenbaum's attributable effects framework to provide non-parametric tests of the magnitude of the treatment's effect on the network. These tests can be used to understand many aspects of network topology and require few assumptions beyond the design of the RCT itself.

Solo authored.

Keywords: network analysis; randomization inference; Fisher's exact test; attributable effects; table sampling.

Multigraphical Lasso for Tensor-valued Data

Greenewald, Kristjan (greenewk@umich.edu)
University of Michigan

Type: Poster

Abstract. Recently, the BiGLasso was proposed to parsimoniously model the covariance of matrix-valued data, enforcing both sparsity and explicit structure on the inverse covariance, significantly reducing the number of parameters and empirically reducing the number of training samples required to learn the covariance. In this work, we extend the BiGLasso to the Multigraphical Lasso, an analogous method modeling the covariance of tensor-valued data, and present an algorithm for the solution of both methods. We derive non-asymptotic performance bounds for our estimator and the BiGLasso estimator. Finally, we apply our methods to an EEG seizure dataset, demonstrating the improvements in required sample size that the imposed structure provides.

Coauthors: Shuheng Zhou, Alfred Hero III

Keywords: covariance; sparsity; concentration bounds; graphical lasso; kronecker sum.

Singularity Structures and Parameter Estimation Behavior in Finite Mixtures of Distributions

Ho, Nhat (minhnhat@umich.edu)
Department of Statistics, University of Michigan, Ann Arbor

Type: Poster

Abstract. Understanding singularity structures of the Fisher information matrix in finite mixture models has been a very challenging problem since this matrix is usually singular and has very low rank around specific values of parameters. In this talk, we develop a general and comprehensive theory to capture these singularity structures under all possible settings of finite mixtures. The general theory is then illustrated under the specific setting of skew normal mixtures. These models have become increasingly popular in recent years due to their flexibility in modeling asymmetric data. However, they appear to contain various singularities under both the exact-fitted setting, i.e the setting when the number of mixing components is known, or the over-fitted setting, i.e the setting where the number of mixing components is bounded but unknown. These singularities happen not only in the vicinity of symmetry but also in the setting of homologous sets, a new phenomenon due to the complex interaction among the parameters of the mixing measures. Apart from these singularities, skew normal density also possesses the system of non-linear partial differential equations structure. It leads to two interesting ways of characterizing the singularity levels of the Fisher information matrix under skew normal mixtures. One way is based on the solvability of inhomogeneous system of polynomial equations while the another way is based on the combining strength phenomenon among multiple homogeneous and inhomogeneous systems of polynomial equations. The rich spectrum of the singularity structures consequently leads to various intricate degrees of parameter estimation.
Coauthor: XuanLong Nguyen
Keywords: Fisher singularities; mixture models; partial differential equations; system of polynomial equations; Wasserstein distances.

Penalized Ensemble Kalman Filters for High Dimensional Non-linear Systems

Hou, Elizabeth (emhou@umich.edu)
University of Michigan; LANL

Type: Poster

Abstract. The ensemble Kalman filter is one of the fastest tracking algorithms for complicated non-linear systems; however, when the state space is high dimensional, the ensemble Kalman filter suffers from propagating sampling errors. This is due to the computational burden of numerically integrating a large ensemble forward through a complicated system, making it typical to have far fewer ensemble members than state dimensions. To deal with tracking in this high dimensional regime, we propose a computationally fast and easy to implement algorithm which we call the penalized ensemble Kalman filter (PEnKF). Under certain conditions, we can prove that the PEnKF does not require more ensemble members than state dimensions in order to have good performance. We support these theoretical results through superior performance in simulations of multiple non-linear and high dimensional systems.
Coauthors: Earl Lawrence, Alfred O. Hero
Keywords: Data Assimilation; Kalman Filter; Regularization; High-Dimensional; Non-linear Systems.

A New Graphical Approach for Quick and Exact Simulation of Correlated Discrete Random Variables

Jiang, Bei (bei1@ualberta.ca)
University of Alberta

Type: Poster

Abstract. Simulation of correlated discrete variables with specified marginals and covariances have important applications. For example, there is a need in neuroscience research to simulate discrete neural spike train data across brain regions. In this paper, we present a novel graphical approach for simulating such random variables using an efficient one-pass algorithm, where the random sample is drawn for each variable in one iteration. We also give the conditions for compatibility of the marginal probabilities and covariances. This one-pass algorithm also leads to the construction of a family of Markov random fields on a directed acyclic graph with conditional and joint field distributions. A necessary and sufficient condition that guarantees the permutation property of the derived random field is studied.
Keywords: Correlated Random Field; Directed Acyclic Graph; Permutation Property.

Modeling Multiple Brain Networks through Linear Mixed Effects Models

Kim, Yura (kimyr@umich.edu)
University of Michigan

Type: Poster

Abstract. Data on the brain's structural or functional connections are frequently represented in the form of networks, with a different network for each subject in the study. These networks all share the same set of nodes and can thus be analyzed jointly. Current work tends to either reduce them to global summaries such as modularity, or vectorize the edge values and ignore network structure. Here we propose a method for modeling brain networks via linear mixed effects models which takes advantage of the community structure, or regions, known to be present in the brain. The model allows us to compare different populations (for example, healthy and mentally ill patients) both globally and at the edge level, and find significant areas of difference. Further, we can incorporate the correlation between edges inherent in brain data by allowing for a general variance structure in the mixed effects model. We illustrate the method by analyzing data from a study comparing schizophrenics to healthy controls.

Joint work with prof. Elizaveta Levina
Keywords: brain network; community structure; linear mixed effects model.

High-dimensional Matrix Linear Regression Model

Kong, Dehan (kongdehanstat@gmail.com)
University of Toronto

Type: Poster

Abstract. We develop a high-dimensional matrix linear regression model to correlate matrix responses with a high dimensional vector of covariates when coefficient matrices have low-rank structure. We propose a fast and efficient screening procedure based on the spectral norm of each coefficient matrix in order to deal with the case when the number of covariates is large. We develop an efficient estimation procedure based on the trace norm regularization, which explicitly imposes the low rank structure of coefficient matrices. We systematically investigate some theoretical properties of our estimators, including estimation consistency, rank consistency, and the sure independence screening property under some mild conditions. We examine the finite-sample performance of our screening and estimation methods by using simulations and a large-scale imaging genetic dataset collected by the Alzheimer's Disease Neuroimaging Initiative study.
Keywords: Imaging Genetics, Low Rank, Matrix Linear Regression, Spectral norm, Trace norm..

Network Cross-validation by Edge Sampling

Li, Tianxi (tianxili@umich.edu)
University of Michigan

Type: Poster

Abstract. Many models and methods are now available for network analysis, but model selection and tuning remain challenging. Cross-validation is a useful general tool for these tasks in many settings, but is not directly applicable to networks since splitting network nodes into groups requires deleting edges and destroys some of the network structure. We propose a new edge sampling cross-validation strategy applicable to a wide range of network problems. We provide theoretical justifications on the effectiveness of our method in a general setting, and in particular show that the method has good asymptotic properties under the stochastic block model. Numerical results on both simulated and real networks show that our approach performs well for a number of model selection and parameter tuning tasks.
Keywords: network modeling; cross-validation; model selection; matrix completion.

Change-Point Estimation using Shape-Restricted Regression Splines

Liao, Xiyue L (liaoxiyue2011@gmail.com)
School of Medicine, New York University

Type: Poster

Abstract. Change-Point estimation is in need in fields like climate change, signal processing, economics, dose-response analysis etc, but it has not yet been fully discussed. We consider estimating a regression function and its change-point m, where m is a mode, an inflection point, or a jump point. Linear inequality constraints are used with spline regression functions to estimate m and the regression function simultaneously using profile methods. For a given m, the maximum-likelihood estimate of the regression function is found using constrained regression methods, then the set of possible change-points is searched to find the m that maximizes the likelihood. Convergence rates are obtained for each type of change-point estimator, and we show an oracle property, that the convergence rate of the regression function estimator is as if m were known. Parametrically modeled covariates are easily incorporated in the model. The scenario when the random error is from a stationary autoregressive process is also presented. Penalized spline-based regression is also discussed as an extension.

Co-author: Mary C. Meyer

Keywords: change-point estimation; monotone; convex; convergence rate.

Model-based Community Detection for Networks with Node Covariates

Liu, Boang (boangliu@umich.edu)
Department of Statistics, University of Michigan

Type: Poster

Abstract. Network data with node covariates are common in many fields. In principle, the two sources of information can be combined for community detection. However, most existing methods either lack statistical interpretation or make strong conditional independence assumptions. In this project, we develop a general statistical framework to describe the relationship between the link structure, node covariates, and communities. Further, we propose two families of statistical models which are the most general under this framework with the least conditional independence assumptions between the three parts. We develop variational EM algorithms to estimate community memberships as well as model parameters. The proposed methods have been tested on both simulated and real networks. Co-author: Ji Zhu, Department of Statistics, University of Michigan

Keywords: network; node covariates; community detection; variational EM algorithm.

Compressed Sensing without Sparsity Assumptions

Lopes, Miles (melopes@ucdavis.edu)
UC Davis, Statistics Department

Type: Poster

Abstract. The theory of Compressed Sensing (CS) asserts that an unknown p-dimensional signal can be accurately recovered from an underdetermined set of n linear measurements with $n << p$, provided that x is sufficiently sparse. However, in applications, the degree of sparsity $||x||_0$ is typically unknown, and the problem of directly estimating $||x||_0$ has been a longstanding gap between theory and practice. A closely related issue is that $||x||_0$ is a highly idealized measure of sparsity, and for real signals with entries not equal to 0, the value $||x||_0 = p$ is not a useful description of compressibility. In our previous work that examined these problems, we considered an alternative measure of "soft" sparsity, $(||x||_1/||x||_2)^2$, and designed a procedure to estimate $(||x||_1/||x||_2)^2$ that does not rely on sparsity assumptions.

The present work offers a new deconvolution-based method for estimating unknown sparsity, which has wider applicability and sharper theoretical guarantees. In particular, we introduce a family of entropy-based sparsity measures $s_q(x)$, parameterized by $q \in [0, \infty]$, which includes $(||x||_1/||x||_2)^2$ and $||x||_0$ as special cases. Also, we propose an estimator for $s_q(x)$ whose relative error converges at a parametric rate, even when p/n diverges. Our main results describe the limiting distribution of the estimator, as well as some connections to Basis Pursuit Denosing, the Lasso, deterministic measurement matrices, and inference problems in CS. (http://arxiv.org/abs/1507.07094; to appear in IEEE Trans. Info. Theory)
Keywords: Compressed Sensing, Unknown Sparsity, Lasso, High-Dimensional Inference.

Provable Symmetric Nonnegative Matrix Factorization for Overlapping Clustering

MAO, XUEYU (maoxueyu@gmail.com)
University of Texas at Austin

Type: Poster

Abstract. Abstract: The problem of finding overlapping communities in networks has gained much attention recently. Algorithmic approaches often employ non-negative matrix factorization (NMF) or variants, while model-based approaches (such as the widely used mixed-membership stochastic blockmodel, or MMSB) assume a distribution over communities for each node and run standard inference techniques to recover these parameters. However, few of these approaches have provable consistency guarantees. We investigate the use of the symmetric NMF (or SNMF) for the MMSB model, and provide conditions under which an optimal SNMF algorithm can recover the MMSB parameters consistently. Since we are unaware of general-purpose optimal SNMF algorithms, we develop an SNMF variant, called GeoNMF, designed specifically for the MMSB model. GeoNMF is provably consistent, and experiments on both simulated and real-world datasets show its accuracy. Authors: Xueyu Mao, Purnamrita Sarkar, Deepayan Chakrabarti
Keywords: Overlapping clustering, spectral methods, symmetric non-negative matrix factorization.

Ensemble Estimation of Mutual Information

Moon, Kevin (kevimoon@gmail.com)
Yale University

Type: Poster

Abstract. We derive the mean squared error convergence rates of kernel density-based plug-in estimators of mutual information measures without boundary correction. We propose ensemble estimators of these information measures that achieve the parametric convergence rate when the densities are sufficiently smooth by taking a weighted sum of these plug-in estimators with varied bandwidths. A central limit theorem is given and the estimators are extended to the case where the data contain a mixture of discrete and continuous components. To the best of our knowledge, this is the first nonparametric mutual information estimator known to achieve the parametric convergence rate for this mixed data. The estimator has the added benefit of being simple to implement and performs well in higher dimensions.
Coauthors: Kumar Sricharan, Alfred Hero.
Keywords: mutual information; ensemble estimators; kernel density estimators.

Fast, Scalable, Consistent Community Detection by Local Patching Algorithms

Mukherjee, Soumendu Sundar (soumendu@berkeley.edu)
Department of Statistics, University of California, Berkeley

Type: Poster

Abstract. We propose a local clustering algorithm for community detection that performs clustering on a number of subgraphs and finally patches the results into a single clustering. In the local step, one can plug in any existing clustering algorithm like spectral clustering, SDP based methods, likelihood based methods (pseudo-likelihood, variational likelihood), modularity based methods (Newman-Girvan modularity, likelihood modularity) etc. The algorithm is inherently parallelizable and thus much faster and more scalable than a corresponding global algorithm on large graphs. This is particularly useful in case of SDP based, likelihood based and/or modularity based methods which do not scale well to large graphs. We prove (weak) consistency of the local algorithm for the stochastic block model (SBM) under various subgraph selection procedures, including random subgraphs, and random ego neighborhoods. Empirically, in simulations, the local algorithms are not only fast, but also improve upon the misclassification errors of the corresponding global procedures in various regimes.

(Joint work with Peter J. Bickel and Purnamrita Sarkar.)
Keywords: Community detection; stochastic block model; local patching algorithms; scalability; consistency..

Understanding Brain Connectivity by Estimating a Non-stationary Covariance using Multi-resolution Knots

Nandy, Siddhartha (nandysid@stt.msu.edu)
Department of Statistics and Probability, Michigan State University

Type: Poster

Abstract. Human brain mapping from functional Magnetic Resonance Imaging has outgrown in several interdisciplinary subjects and has worthwhile contributions over last few decades.The field of brain connectivity can be broadly classified into three types, anatomical,functional and,effective connectivity.The latest one works at neuroanatomical or neuronal level and, leads to more interesting statistical parametric maps.Structural equation modeling(SEM) and dynamic causal modeling(DCM) are mostly used models to study effective connectivity.Further SEM lies on explaining the underlying covariance structure of the data.We will consider our SEM analysis based on the observed blood oxygen level dependent(BOLD) contrast which has a spatio-temporal covariance structure.We propose a technique of estimating the underlying covariance structure using bivariate knots overlay-ed in multi-resolution fashion.This leads us to the interpretation of connectivity using non-zero off-diagonal entries from our estimated covariance matrix.The observed data are at 85 slices of image with size 6085 on human brain.We select centrally spaced 93079 voxels to exclude low signal voxels outside the brain region.Also at each voxel, we further observe 256 temporal resolution.

Co Authors : Chae-Young Lim, Tapabrata Maiti

Keywords: Structural equation modeling, Multi-resolution knots, Statistical parametric maps, functional Magnetic Resonance Imaging, Blood oxygen level dependent.

Dealing Outliers/Subclusters for Non-linear High-to-low Gaussian Mixture Regression

Tu, Chun-Chen (timtu@umich.edu)
University of Michigan

Type: Poster

Abstract. Deleforge et al (2013) proposed a Gaussian Locally-Linear Mapping(GLLiM) approach to carry out predicting low dimensional responses using high dimensional predictors. The key concept behind their approach is to use mixtures to approximate non-linear patterns locally and to use inverse regression to mitigate the complications due to high-dimensional predictors. In this work, we focus on identifying the potential outliers and subclusters embedded within the mixture components found in the high-dimensional setting in Deleforge et al and then further use them to improve prediction outcomes. Like their approach, the problem can be treated as mixture of affine regressions in a well-designed joint probability model, and the high-dimensionality can be handled with inverse regression. The advantage of assuming Gaussianity is to makes the estimation of parameters computationally efficient. Specifically, our method addresses the issue that estimation of a local linearity group at a high-dimensional setting could erroneously include some data points

that are outliers in the corresponding low-dimensional space. To solve this problem, we propose a Divide-ReMerge method for predicting low dimensional responses using high dimensional predictors. The parameters can be estimated using an Expectation-Maximization algorithm. Numerical results of several dataset that illustrate the improvement of the prediction performance will be provided.
Keywords: Regression; Mixture Models; Expectation-Maximization; Outlier.

Robust Categorical Principal Components Analysis

Turkmen, Asuman (turkmen.2@osu.edu)
The Ohio State University at Newark

Type: Poster

Abstract. Principal component analysis (PCA) is primarily an exploratory technique that is widely used for dimensionality reduction and visualization of multivariate data. The computation of principal components is only tailored to continuous variables. Practically important violation occurs when it is applied to discrete data, which are encountered frequently in empirical analysis. For instance, genotype data used in genome-wide association studies (GWAS) are discrete in nature, but PCA is routinely used to analyze these datasets. In addition traditional PCA is sensitive to outliers, which represent errors or bias in datasets. It is still unknown how to deal with outliers in discrete PCA. For example, genotype data would contain genotyping errors (outliers) caused by a sample mix-up, incorrectly specified relationships, technician error, or technology failure. Detecting genotyping errors, and developing analytical methods that are robust to such errors would help to decrease the risk of potential false findings (positive or negative) and to increase our ability to identify real associations. Therefore, the purpose of this study is to develop a robust categorical PCA that can be applied to categorical datasets containing outliers such as genotype data. We show through a simulation study and an application to genomic data example, that the robust methodology can be more powerful and thus more adequate for association studies than the classical approach. Co-authors: Yuan Yuan (Auburn University) & Nedret Billor (Auburn University)
Keywords: Discrete data; dimension reduction; GWAS; outlier detection.

Unified Empirical Likelihood Ratio Tests for Functional Concurrent Linear Models and the Phase Transition from Sparse to Dense Functional Data

Wang, Honglang (hlwang@iupui.edu)
Indiana University-Purdue University Indianapolis

Type: Poster

Abstract. We consider the problem of testing functional constraints in a class of functional concurrent linear models where both the predictors and the response are functional data measured

at discrete time points. We propose test procedures based on the empirical likelihood with bias-corrected estimating equations to conduct both pointwise and simultaneous inferences. The asymptotic distributions of the test statistics are derived under the null and local alternative hypotheses, where sparse and dense functional data are considered in a unified framework. We find a phase transition in the asymptotic null distributions and the orders of detectable alternatives from sparse to dense functional data. Specifically, the proposed tests can detect alternatives of root-n order when the number of repeated measurements per curve is of an order larger than n^{η_0} with n being the number of curves. The transition points η_0 for pointwise and simultaneous tests are different and both are smaller than the transition point in the estimation problem. Simulation studies and real data analyses are conducted to demonstrate the proposed methods.

co-authers: Ping-Shou Zhong (Michigan State University), Yuehua Cui (Michigan State University), Yehua Li (Iowa State University)

Keywords: Empirical likelihood; Functional ANOVA; Nonparametric hypothesis testing; Unified inference.

Estimation of the Average Treatment Effect in Randomized Experiments using Random Forests

Wu, James (jameswu@umich.edu)
Graduate Student

Type: Poster

Abstract. When assessing the results of a randomized experiment, it is natural to estimate the average effect of the treatment. One way to do this is through the simple difference estimator: the difference between the average observed outcome of the treatment group and the average of the control group. One can estimate the treatment effect with lower variance by adjusting for covariates. This is commonly done through regression; however, the regression estimate is biased in the Neyman model and can, in certain circumstances, be outperformed by the simple difference estimator. Furthermore, randomization does not justify the assumptions of least squares regression. We propose a method that estimates the average treatment effect by leaving each observation out and then estimating that observations treatment and control outcomes using the random forest algorithm. This estimator solves the issues with linear regression: randomization justifies our assumptions, our estimator is unbiased, and it performs no worse than the simple difference estimator (and can substantially improve performance).

Co-Author: Johann Gagnon-Bartsch

Keywords: randomization; random forest; treatment effects; leave one out.

Convex Relaxation for Community Detection with Covariates

Yan, Bowei (boweiy@utexas.edu)
University of Texas at Austin

Type: Poster

Abstract. Community detection in networks is an important problem in many applied areas. In this paper, we investigate this in the presence of node covariates. Recently, an emerging body of theoretical work has been focused on leveraging information from both the edges in the network and the node covariates to infer community memberships. However, so far the role of the network and that of the covariates have not been examined closely. In essence, in most parameter regimes, one of the sources of information provides enough information to infer the hidden clusters, thereby making the other source redundant. To our knowledge, this is the first work which shows that when the network and the covariates carry "orthogonal" pieces of information about the cluster memberships, one can get asymptotically consistent clustering by using them both, while each of them fails individually. This is joint work with Purnamrita Sarkar.
Keywords: semidefinite programming; consistency; covariates; block models; sparse graphs.

Principal Nested Spheres for Time Warped Functional Data Analysis

Yu, Qunqun (qunyu@live.unc.edu)
University of North Carolina at Chapel Hill

Type: Poster

Abstract. There are often two important types of variation in functional data: the horizontal (or phase) variation and the vertical (or amplitude) variation. These two types of variation have been appropriately separated and modeled through a domain warping method (or curve registration) based on the Fisher-Rao metric. This paper focuses on the analysis of the horizontal variation, captured by the domain warping functions. The square-root velocity function representation transforms the manifold of the warping functions to a Hilbert sphere. Motivated by recent results on manifold analogs of principal component analysis, we propose to analyze the horizontal variation via a Principal Nested Spheres approach. Compared with earlier approaches, such as approximating tangent plane principal component analysis, this is seen to be an efficient and interpretable approach to decompose the horizontal variation in both simulated and real data examples.
Keywords: Functional data variability; Principal Nested Spheres; Time warping.

Indian Factorization Machines

Yurochkin, Mikhail (moonfolk@umich.edu)
University of Michigan

Type: Poster

Abstract. In this work we investigate how to incorporate interactions between variables into regression type problems. In the high dimensional setting, the number of possible interactions grows

exponentially. Factorization Machines use lower-dimensional representations of the interaction coefficients, but is limited to two-way, or at most three-way, interactions due to exponential growth of the model complexity. We propose a novel model utilizing lower-dimensional representations of the interactions together with an Indian Buffet Process prior for the hypergraph of interactions. Our MCMC algorithm can jointly learn interactions of arbitrary depth and their weights. Co-authors: XuanLong Nguyen, Nikolaos Vasiloglou.
Keywords: Factorization Machines; Indian Buffet Process; Bayesian Nonparametrics.

Nonparametric Seeded Graph Matching

Zhang, Yuan (zhang.7824@osu.edu)
Statistics, Ohio State University

Type: Poster

Abstract. Seeded graph matching is a semi-supervised learning problem that attracts raising interests in recent years. The task is to pair up nodes in two graphs, provided just a few correctly matched node pairs. The problem is typical in the cheap data, expensive manpower era, and the topic finds a wide range of applications in engineering, social sciences, biomedical and biological studies and so on. Existing methods usually require strong modeling conditions such as that a latent perfect matching exists and that the generative model is low rank and positive semi-definite. In this paper, we propose a novel method that can address much more general cases: 1. partially matched graphs, i.e. graphs that contain unique parts that cannot be paired-up; 2. graphs in different sizes; 3. graphs that are essentially high-rank; 4. graphs that are not positive/negative semi-definite. We also studied the theoretical properties of our method and showed that it achieves consistency under quite weak requirements for the seed nodes. Numerical studies show the advantageous performance of our method.
Keywords: 1. statistical network analysis; 2. semi-supervised learning 3. seeded graph matching 4. nonparametric statistics.

Bias Reduction via Resampling Techniques in Approximate Conditional Likelihoods with a General Missing Data Mechanism

Zhao, Jiwei (zhaoj@buffalo.edu)
State University of New York at Buffalo

Type: Poster

Abstract. Statistical methods for handling missing data depend on how the missing data mechanism is assumed. In this paper, we consider a generally applicable missing data mechanism, which includes both ignorable and nonignorable missing data scenarios. We introduce the conditional likelihood function and its approximation as the base for estimating the unknown parameter in the data

generating process. We find that the bias is a severe issue when using this approximate conditional likelihood function. One reason for this is that we only use the completed observed samples in our proposal. Large bias may deteriorate the power and may jeopardize other quantifications, like the results for the variable selection. In this paper, we propose to use some resampling techniques to reduce the bias in the estimation in our problem. We consider both Jackknife and Bootstrap methods. We develop the theoretical results in each situation, and conduct comprehensive simulation studies to evaluate the finite sample performance of our proposed methods.
Keywords: Missing data mechanism; approximate conditional likelihood; Jackknife; Bootstrap; bias reduction..

Network Inference From Time Varying Grouped Observations

Zhao, Yunpeng (yzhao15@gmu.edu)
George Mason University

Type: Poster

Abstract. In social network analysis, the observed data is usually some social behavior, such as the formation of groups, rather than explicit network structure. Zhao and Weko (2015) proposed a model-based approach called the star model to infer implicit networks from grouped observations. Star models assumed independence among groups, which sometimes is not valid for practical consideration. We generalize the idea of star models into the case of time varying grouped observations. Similarly to star models, we assume the group at each time point is gathered by one leader, but allow dependency among groups in the same time segment. We apply a variant of Expectation-Maximization algorithm – hard EM for identifying group leaders and apply a label switching technique to optimize Bayesian information criterion for identifying segments. The performance of the new model is evaluated under different simulation settings. We apply this model to a data set of Kibale chimpanzee project.
Keywords: grouping behavior; social networks; hard EM.

Using the Cox Model for Current Status Data to Assess Altered Timing of Sexual Maturation by Exposure to Environmental Toxic Mixtures

Zhou, Ling (zholing@umich.edu)
Biostatistics

Type: Poster

Abstract. Motivating by a collaborative project that aims to evaluate the adverse effects of pre- and post-natal exposures to multiple toxic agents (e.g., lead) on onset of sexual maturation, we develop a new Cox model for self-reported event of menarche among adolescent girls living in Mexico City. The proposed Cox model for current status data enables to assess altered effects

of growth and developmental predictors of interest by a bundle of multiple toxicants, in which nonlinear interactions are imposed to capture flexibly the pattern of alterations in covariate effects. Utilizing the splines smoothing technique, we develop an estimation method for parameters and nonparametric functions, and then a statistical inference for parameters based on a second step of updating with local linear fitting. A generalized likelihood ratio method is also established to conduct hypothesis testing for linear or nonlinear interactions. The proposed methodology is examined by simulation studies and is illustrated by the motivating data example of girl's sexual maturation. Co-authors: Peter X.-K. Song, University of Michigan

Keywords: Cox additive model; index model; interaction; principal component analysis; varying coefficient.

Estimation and Prediction of Longitudinal Biomarker Distributions Using Bayesian Nonparametric Beta Regression

Zhou, Shouhao (szhou@mdanderson.org)
UT MD Anderson Cancer Center

Type: Poster

Abstract. This work was motivated by a study of chronic myeloid leukemia (CML), in which the limited-range continuous response variable, BCR-ABL transcript level, is a surrogate of residual disease following treatment with tyrosine kinase inhibitors. We propose Bayesian hierarchical models to jointly estimate subject-specific smooth distribution curves of BCR-ABL transcript levels over time and covariate effects on these transcript levels. Its goal is to help physicians understand the relative extremeness of a recently measured BCR-ABL level at a certain time point after CML treatment, and to predict the future trajectory of BCR-ABL transcript levels for a specific CML patient given his/her historical data. This information will enhance the personalized management of patients treatment, which is an important component of precision medicine.

Keywords: Bayesian beta regression; Fractional polynomials; Longitudinal analysis; Subject-specific time trend.

Simulation-Based Hypothesis Testing of High-Dimensional Means Under Covariance Heterogeneity

Zhou, Wen (riczw@stat.colostate.edu)
Colorado State University

Type: Poster

Abstract. In this paper, we study the problem of testing the mean vectors of high dimensional data in both one-sample and two-sample cases. The proposed testing procedures employ maximum-type statistics and the parametric bootstrap techniques to compute the critical values. Different from

the existing tests that heavily rely on the structural conditions on the unknown covariance matrices, the proposed tests allow general covariance structures of the data and therefore enjoy wide scope of applicability in practice. To enhance powers of the tests against sparse alternatives, we further propose two-step procedures with a preliminary feature screening step. Theoretical properties of the proposed tests are investigated. Through extensive numerical experiments on synthetic datasets and an human acute lymphoblastic leukemia gene expression dataset, we illustrate the performance of the new tests and how they may provide assistance on detecting disease-associated gene-sets.
Keywords: Feature screening; High dimension; Hypothesis testing; Normal approximation; Parametric bootstrap; Sparsity..