# Rutgers Statistics Workshop

## *Innovations in Statistics and Data Analysis*

**Department of Statistics & Biostatistics**

***Friday, January 30, 2015***

***CoRE Auditorium, Busch Campus, Rutgers University***

# Abstracts

### High-Dimensional Testing Against Sparse Alternatives in Genetic Association Studies

**Xihong Lin, Harvard School of Public Health**

Massive genetic data present many exciting opportunities as well as challenges in data analysis and result interpretation, e.g., how to develop effective strategies for signal detection using massive genetic data when signals are weak and sparse. Many variable selection methods have been developed for analysis of high-dimensional data in the statistical literature. However limited work has been done on statistical inference for massive data. In this talk, I will discuss hypothesis testing for analysis of high-dimensional data motivated by gene, pathway/network based analysis in genome-wide association studies using arrays and sequencing data. I will focus on signal detection when signals are weak and sparse, which is the case in genetic association studies. I will discuss hypothesis testing for signal detection using penalized likelihood-based methods and the generalized higher criticism test. The results are illustrated using data from genome-wide association studies.

## Robust Recovery and Detection of Structured Signals

**Tony Cai, University of Pennsylvania**

A large collection of statistical methods has been developed for estimation and detection of structured signals in the Gaussian and sub-Gaussian settings. In this talk, we present a general approach to robust recovery and detection of structured signals for a wide range of noise distributions. We illustrate the technique with nonparametric regression and detecting and identifying sparse short segments hidden in an ultra long linear sequence of data. A key step is the development of a quantile coupling theorem that is used to connect our problem with a more familiar Gaussian setting. An application to copy number variation (CNV) analysis based on next generation sequencing (NGS) data is also discussed.

## Scaling and Generalizing Variational Inference

**David Blei, Columbia University**

Latent variable models have become a key tool for the modern statistician, letting us express complex assumptions about the hidden structures that underlie our data. Latent variable models have been successfully applied in numerous fields including natural language processing, computer vision, electronic medical records, genetics, neuroscience, astronomy, political science, sociology, the digital humanities, and many others.

The central computational problem in latent variable modeling is posterior inference, the problem of approximating the conditional distribution of the latent variables given the observations. Posterior inference is central to both exploratory tasks, where we investigate hidden structures that underlie our data, and predictive tasks, where we use

the inferred structures to generalize about future data. Approximate posterior inference algorithms have revolutionized Bayesian statistics, revealing its potential as a usable and general-purpose language for data analysis.

Bayesian statistics, however, has not yet reached this potential. First, statisticians and scientists regularly encounter massive data sets, but existing approximate inference algorithms do not scale well. Second, most approximate inference algorithms are not generic; each must be adapted to the specific model at hand. This often requires significant model-specific analysis, which precludes us from easily exploring a variety of models.

In this talk I will discuss our recent research on addressing these two limitations. First I will describe stochastic variational inference, an approximate inference algorithm for handling massive data sets. Stochastic inference is easily applied to a large class of Bayesian models, including time-series models, factor models, and Bayesian nonparametric models. I will demonstrate its application to probabilistic topic models of text conditioned on millions of articles. Stochastic inference opens the door to scalable Bayesian computation for modern data analysis.

Then I will discuss black box variational inference. Black box inference is a generic algorithm for approximating the posterior. We can easily apply it to many models with little model-specific derivation and few restrictions on their properties. Black box inference performs better than similarly generic sampling algorithms, such as Metropolis-Hastings inside Gibbs, and can be composed with stochastic inference to handle massive data. I will demonstrate its use on a suite of nonconjugate models of longitudinal healthcare data.

This is joint work based on two papers:

M. Hoffman, D. Blei, J. Paisley, and C. Wang. Stochastic variational inference. Journal of Machine Learning Research, 14:1303-1347.

R. Ranganath, S. Gerrish, and D. Blei. Black box variational inference. Artificial Intelligence and Statistics, 2014.

# Bayesian Inference on Network Data

**David Dunson, Duke University**

Network data are increasingly available along with other variables of interest. Our motivation is drawn from neurophysiology studies measuring a brain activity network for each subject along with a categorical variable, such as presence or absence of a neuropsychiatric disease, creativity groups or type of ability. We develop a Bayesian approach for inferences on group differences in the network structure, allowing global and local hypothesis testing adjusting for multiplicity. Our approach allows the probability mass function for network-valued data to shift nonparametrically between groups, via a dependent mixture of low-rank factorizations. An efficient Gibbs sampler is defined for posterior computation. We provide theoretical results on the flexibility of the model and assess testing performance in simulations. The approach is applied to provide novel results showing relationships between human brain networks and creativity.

Joint work with Daniele Durante.