Statistics Workshop on High-Dimensional and Spatial Data Analysis

Date: 18 June 2017 (Sunday), 9:40-17:00
Venue: FSC 1217, Hong Kong Baptist University

Speakers:

Wenlin Dai, King Abdullah University of Science and Technology, Saudi Arabia
Marc G. Genton, King Abdullah University of Science and Technology, Saudi Arabia
Bo Li, University of Illinois at Urbana-Champaign, USA
Nan Lin, Washington University in St. Louis, USA
Qi Long, University of Pennsylvania, USA
Ying Sun, King Abdullah University of Science and Technology, Saudi Arabia
Yichuan Zhao, Georgia State University, USA

Program:

9:40-10:20 **Marc G. Genton**
Computational challenges with big environmental data

10:20-11:00 **Qi Long**
Scalable Bayesian variable selection for structured high-dimensional data

11:00-11:20 Break

11:20-12:00 **Bo Li**
Spatially varying autoregressive models for prediction of new HIV diagnoses

12:00-14:00 Lunch

14:00-14:40 **Yichuan Zhao**
Smoothed jackknife empirical likelihood method for ROC curves with missing data

14:40-15:20 **Ying Sun**
Visualization and assessment of spatio-temporal covariance properties

15:20-15:40 Break

15:40-16:20 **Nan Lin**
Group sparsity via approximated information criteria

16:20-17:00 **Wenlin Dai**
Directional outlyingness for multivariate functional data

18:00-21:00 Dinner

Abstracts (in order of appearance):

Marc G. Genton: Computational challenges with big environmental data

Abstract: Two types of computational challenges arising from big environmental data are discussed. The first type occurs with multivariate or spatial extremes. Indeed, inference for max-stable processes observed at a large collection of locations is among the most challenging problems in computational statistics, and current approaches typically rely on less expensive composite likelihoods constructed from small subsets of data. We explore the limits of modern state-of-the-art computational facilities to perform full likelihood inference and to efficiently evaluate high-order composite likelihoods. With extensive simulations, we assess the loss of information of composite likelihood estimators with respect to a full likelihood approach for some widely-used multivariate or spatial extreme models. The second type of challenges occurs with the emulation of climate model outputs. We consider fitting a statistical model to over 1 billion global 3D spatio-temporal temperature data using a distributed computing approach. The statistical model exploits the gridded geometry of the data and parallelization across processors. It is therefore computationally convenient and allows to fit a non-trivial model to a data set with a covariance matrix comprising of 10^{18} entries. We provide 3D visualization of the results with Google glasses. The talk is based on joint work with Stefano Castruccio and Raphael Huser.

Qi Long: Scalable Bayesian variable selection for structured high-dimensional data

Abstract: Variable selection for structured covariates lying on an underlying known graph is a problem motivated by practical applications, and has been a topic of increasing interest. However, most of the existing methods may not be scalable to high dimensional settings involving tens of thousands of variables lying on known pathways such as the case in genomics studies. We propose an adaptive Bayesian shrinkage approach which incorporates prior network information by smoothing the shrinkage parameters for connected variables in the graph, so that the corresponding coefficients have a similar degree of shrinkage. We fit our model via a computationally efficient expectation maximization algorithm which scalable to high dimensional settings. Theoretical properties for fixed as well as increasing dimensions are established, even when the number of variables increases faster than the sample size. We demonstrate the advantages of our approach in terms of variable selection, prediction, and computational scalability via a simulation study, and apply the method to a cancer genomics study. Joint work with Changgee Chang and Suprateek Kundu.

Bo Li: Spatially varying autoregressive models for prediction of new HIV diagnoses

Abstract: In demand of predicting new HIV diagnosis rates based on publicly available HIV data that is abundant in space but has few points in time, we propose a class of spatially varying autoregressive (SVAR) models compounded with conditional autoregressive (CAR) spatial correlation structures. We then propose to use the copula approach and a flexible CAR formulation to model the dependence of adjacent counties. These models allow for spatial and temporal correlation as well as space-time interactions and are naturally suitable for predicting HIV cases and other spatio-temporal disease data that feature a similar data structure. We apply the proposed models to HIV data over Florida, California and New England states and comparethem to a range of linear mixed models that have been recently popular for modeling spatio-temporal disease data. The results show that for such data our proposed models outperform the others in terms of prediction.

Yichuan Zhao: Smoothed jackknife empirical likelihood method for ROC curves with missing data

Abstract: In this talk, we apply smoothed jackknife empirical likelihood (JEL) method to construct confidence intervals for the receiver operating characteristic (ROC) curve with missing data. After using hot deck imputation, we generate pseudo-jackknife sample to develop jackknife empirical likelihood. Comparing to traditional empirical likelihood method, the smoothed JEL has a great advantage in saving computational cost. Under mild conditions, the smoothed jackknife empirical likelihood ratio converges to a scaled chi-square distribution. Furthermore, extensive simulation studies in terms of coverage probability and average length of confidence intervals demonstrate this proposed method has the good performance in small sample sizes. A real data set is used to illustrate our proposed JEL method. This talk is based on joint work with Hanfang Yang.

Ying Sun: Visualization and assessment of spatio-temporal covariance properties

Abstract: Spatio-temporal covariances are important for describing the spatio-temporal variability of underlying random processes in geostatistical data. For second-order stationary processes, there exist subclasses of covariance functions that assume a simpler spatio-temporal dependence structure with separability and full symmetry. However, it is challenging to visualize and assess separability and full symmetry from spatio-temporal observations. In this work, we propose a functional data analysis approach that constructs test functions using the cross-covariances from time series observed at each pair of spatial locations. These test functions of temporal lags summarize the properties of separability or symmetry for the given spatial pairs. We use functional boxplots to visualize the functional median and the variability of the test functions, where the extent of departure from zero at all temporal lags indicates the degree of non-separability or asymmetry. We also develop a rank-based nonparametric testing procedure for assessing the significance of the non-separability or asymmetry. The performances of the proposed methods are examined by simulations with various commonly used spatio-temporal covariance models. To illustrate our methods in practical applications, we apply it to real datasets, including weather station data and climate model outputs.

Nan Lin: Group sparsity via approximated information criteria

Abstract: We propose a new group variable selection and estimation method, and illustrate its application for the generalized linear model (GLM). This new method, termed “gMIC”, was derived from approximating the information criterion by a smooth unit dent function. The gMIC is derived as a smooth approximation of a group-version modification of the information criterion. The approximated information criterion is further reparameterized in a way that not only renders sparse estimation from a smooth programming problem but also facilitates a convenient way of circumventing post-selection inference. Compared to existing group variable selection and estimation methods, the gMIC is free of parameter tuning and hence computationally advantageous. The oracle property of the proposed method was established. Both simulation studies and real examples are provided to support the theory. This is a joint work with Liqun Yu and Xiaogang Su.

Wenlin Dai: Directional outlyingness for multivariate functional data

Abstract: The direction of outlyingness is crucial to describing the centrality of multivariate functional data. Motivated by this idea, we generalize classical depth to directional outlyingness for functional data. We investigate theoretical properties of functional directional outlyingness and find that it naturally decomposes functional outlyingness into two parts: magnitude outlyingness and shape outlyingness which represent the centrality of a curve for magnitude and shape, respectively. Using this decomposition, we provide a visualization tool for the centrality of curves. Furthermore, we design an outlier detection procedure based on functional directional outlyingness. This criterion applies to both univariate and multivariate curves and simulation studies show that it outperforms competing methods. Weather and electrocardiogram data demonstrate the practical application of our proposed framework.