Abstracts

Statistical research in the era of big data,

August 3, 2018

Jiahua Chen, Department of Statistics, University of British Columbia

Monitoring distributional degradation

An environmentally appropriate approach to engineering materials is to keep the product as natural as possible and quantify their variability. Reducing the variability will be done only as necessary, not through some energy-intensive refining processes. The quality distribution of a grade can be maintained or improved by being selective of the natural product to be included. To achieve this goal, one should have the product constantly monitored. The industry often wishes to detect signs of quality deterioration in form of distribution degradation through a sequential sampling plan.

In this paper, we develop statistical methods for monitoring the quality distribution of natural products suitable when data are available only sequentially in the presence if historical data. The primary application of the proposed methods is for forestry products. These methods serve also generic purposes for detecting changes in low or in high quantiles of the distribution of any evolving population.

The idea is hence applicable to control problems in various applications.

Lixing Zhu, Hong Kong Baptist University and Beijing Normal University

Order determination for large dimensional matrices

This talk describes how to attack two longstanding problems in determining the model dimensionality (order) when criteria that are based on eigen-decomposition of target matrices are used in practice.

First, due to the existence of some dominating eigenvalues compared to other nonzero eigenvalues, the true dimensionality is often underestimated. Second, the estimation accuracy of any existing method often relies on the uniqueness of minimum/maximum of the criterion. Yet, it is often not the case particularly for the models that converge to a limit with smaller dimensionality. To alleviate these difficulties, we propose a thresholding double ridge ratio criterion. Unlike all the existing eigendecomposition-based criteria, this criterion can define a consistent estimate even when there are several local minima. This generic strategy is readily applied to many fields. As the examples, we give the details about dimension reduction in regressions with fixed and divergent dimensions; about when the number of projected covariates can be consistently estimated, when cannot if a sequence of regression models converges to a limiting model with fewer projected covariates; about ultra-high dimensional factor models and about spiked population models.

Numerical studies are conducted to examine the finite sample performance of the method.

Xin Gao, York University

Fusion learning with high dimensionality

We consider situations where the data consist of a number of responses for each individual, which may include a mix of discrete and continuous variables. The data also include a class of predictors, where the same predictor may have different physical measurements across different experiments depending on how the predictor is measured. The goal is to select which predictors affect any of the responses, where the number of such informative predictors tends to infinity as sample size increases. There are marginal likelihoods for each experiment. We specify a pseudolikelihood combining the marginal likelihoods, and propose a pseudolikelihood information criterion. Under regularity conditions, we establish selection consistency for this criterion with unbounded true model size. The proposed method includes a Bayesian information criterion with appropriate penalty term as a special case. Simulations indicate that data integration can dramatically improve upon using only one data source.

Xuekui Zhang, University of Victoria

Bayesian hierarchical models for SNP discovery from genome-wide association studies, a semi-supervised machine learning approach

Genome-wide association studies (GWASs) aim to detect genetic risk factors for complex human diseases by identifying disease-associated single-nucleotide polymorphisms (SNPs). SNP-wise approach, the standard method for analyzing GWAS, tests each SNP individually. Then the P-values are adjusted for multiple testing. Multiple testing adjustment (purely based on p-values) is over-conservative and causes lack of power in many GWASs, due to insufficiently modelling the relationship among SNPs. To address this problem, we propose a novel method, which borrows information across SNPs by grouping SNPs into three clusters. We pre-specify the patterns of clusters by minor allele frequencies of SNPs between cases and controls, and enforce the patterns with prior distributions. Therefore, compared with the traditional approach, it better controls false discovery rate (FDR) and shows higher sensitivity, which is confirmed by our simulation studies. We re-analyzed real data studies on identifying SNPs associated with severe bortezomib-induced peripheral neuropathy (BiPN) in patients with multiple myeloma. The original analysis in the literature failed to identify SNPs after FDR adjustment. Our proposed method not only detected the reported SNPs after FDR adjustment but also discovered a novel SNP rs4351714 that has been reported to be related to multiple myeloma in another study.

Minge Xie, Rutgers University

Individualized Fusion Learning (iFusion) for Making Personalized Inference in Heterogeneous Big Data

Statistical inferences from multiple data sources can often be fused together to yield more effective inference than from individual source alone. Such fusion learning is of vital importance for big data where data are often assembled in various domains. This paper develops a fusion methodology called individualized fusion learning (iFusion), to enhancing inference for an individual via adaptive combination of confidence distributions obtained from its clique (i.e., peers of similar individuals). iFusion begins with obtaining inference for each individual, then adaptively forming a clique, and finally obtaining a combined inference from the clique. iFusion explores heterogeneity in the database to form a clique for each individual and, by drawing inference from the clique, it allows borrowing strength from similar peers to enhance the inference efficiency for each individual. Furthermore, iFusion can be performed without using the entire data simultaneously and thus allow split-&-conquer to be implemented on individuals to substantially reduce the computational expense. We provide supporting theories for iFusion and also illustrate it using numerical examples.

Xuewen Lu, University of Calgary

Hierarchically Penalized Partially Linear Proportional Hazards Model with a Diverging Number of Parameters

In this paper, we study group variable selection in the partially linear proportional hazards (PH) model with right censored data. We assume a grouping structure exists among the linear explanatory variables in the presence of nonparametric risk functions of low-dimensional covariates. Motivated by the hierarchical grouped variable selection in the linear Cox PH model, we propose a hierarchical bi-level variable selection approach for high-dimensional covariates in the partially linear PH model. The proposed methods are capable of conducting estimation and simultaneous group selection and individual variable selection within groups. The rate of convergence of the parameter estimators is derived and the selection consistency is obtained under a hierarchical penalty and an adaptive hierarchical penalty, respectively. Finally, computational algorithms and programs are developed for utilizing the proposed methods. Simulation studies indicate good finite sample performance of the methods. Real data examples are provided to illustrate the application of the methods.

Lan Liu, University of Minnesota at Twin Cites

Efficiency Boosting via Envelope Chain for Task-evoked fMRI study

The state-of-art method in regression analysis for the brain imaging data is to fit a univariate voxel-wise linear regression or mixed effect models. Such a method utilizes the information of only one voxel at a time, however, ignoring the association between voxels leads to severe efficiency loss. In this paper, we propose a novel statistical method that utilizes information of multiple correlated voxels at the same time and therefore are more efficient. The key idea of our methods is to utilize highly correlated voxels to first identify the direction that contains the information on the group difference, then to project the data onto that direction to reduce noise. The proposed method is further illustrated in Human Connectome Project.

Mark Wolters, Fudan University

The Worst Referee Report Ever

What constitutes a worthy contribution to data science? Based on an unusual recent publication experience, it seems that even experts cannot agree. I will use my recent experience to motivate a discussion about merit and impact, from the perspective of a computationally-oriented statistician trying to compete in the fast-paced and connected "data science" age. I will describe how I plan to avoid similar experiences, while hopefully doing useful work and at the same time remaining gainfully employed.

Xiaoping Shi, Thompson Rivers University

Smoke detection from hyperspectral image data

Hyperspectral remote sensing images acquired from Earth-orbiting satellites hold great promise in helping monitor and catalogue substances emitted into the atmosphere. Even though some methods are successful for smoke detection, there is still a need for more research on these challenging problems: 1) understanding the impact on image by the different wavelengths; 2) collecting continuous images for accurate image segmentation; 3) extending the smoke identification problem to other pollutant detection problem. We will propose two methods for smoke detection.