Table of Contents

Schedule webpage: http://ims.nus.edu.sg/events/2018/wstat/wk2.php

Tutorial on Large Scale Networks

Wanjie Wang

National University of Singapore

ABSTRACT

In the last 20 years, network structures arise in various fields, such as social media, biology, and technology. This trend has motivated the statistical analysis in networks. This tutorial will give a brief overview of some recent works, including:

1. Introduction to networks. Overview different types of networks. Discuss the scientific impact of networks, and frontier problems of interest.

2. Models/methods. Briefly overview some popular models, with a focus on the stochastic block model and its variants. Review some state of the art methods on community detection and membership estimation, under stochastic block model.

3. Application to a real data example. The results on a statistician's coauthorship network will be discussed. The data is based on about 60,000 published papers in 20 top statistics journals in 1980-2015.

4. A short introduction to topic models.

Differential Inclusion Method in High Dimensional Statistics

Yuan Yao

The Hong Kong University of Science & Technology, Hong Kong

ABSTRACT

Boosting, as gradient descent method, is arguably the 'best off-the-shelf' methods in machine learning. Here a novel Boosting-type algorithm is proposed based on restricted gradient descent whose underlying dynamics are governed by differential inclusions. In particular, we present an iterative regularization path with structural sparsity where the parameter is sparse under some linear transforms, based on the Linearized Bregman Iteration or sparse mirror descent. Despite its simplicity, it outperforms the popular (generalised) Lasso in both theory and experiments. A theory of path consistency is presented that equipped with a proper early stopping, it may achieve model selection consistency under a family of Irrepresentable Conditions which can be weaker than the necessary and sufficient condition for generalized Lasso. The utility and benefit of the algorithm are illustrated by applications on sparse variable selection, learning graphical models, partial order ranking, and Alzheimer's disease detection via neuroimaging.

Visualising high-dimensional spaces with application to particle physics models

Dianne Cook[a], Ursula Laa[b], Michael Kipp[a], and German Valencia[b]

[a] *Econometrics and Business Statistics, Monash University, Australia*
[b] *Physics, Monash University, Australia*

ABSTRACT

The grand tour, guided tour and manual tour [1] are used to sample projections of high-dimensional space. These are excellent for understanding the overall shape of structures, and differences between groups, in the space. The grand tour algorithm consists of a sequence of projections of the object space $\mathbb{R}^p$ onto a viewing space $\mathbb{R}^d$. Often $d = 2$, i.e., the object is mapped onto planes, but technically d might be 1, 2, 3, or larger. To be precise, denote $G(d, p)$ to be the Grassmann manifold of d-dimensional planes in $\mathbb{R}^p$. Defining it as $G(d, p)$, ensures that within-plane rotation is removed from the sequence of views.

This talk will discuss new work in adapting tour methods to examine multi-parameter models from particle physics. As experiments in physics get more sophisticated and precise, one is commonly confronted with the problem of interpreting the results in terms of models with multiple parameters. The difficulty of these comparisons between predictions and measurements increases enormously with the dimensionality of the parameter space. In response to these challenges, experimental collaborations have, for the most part, developed tools appropriate for their specific needs. The average theorist, on the other hand, still resorts to simply dropping all but one or two of the relevant parameters. Although this is a very useful way to guide analysis, it misses possible multivariate dependencies. Results obtained in this manner often produce many cluttered plots that are difficult to interpret or understand. Tours can help to examine the results in more than two dimensions.

The new developments include ways to conduct guided tours to search for mismatches in distribution in high-dimensions, and new software for conducting tours in R [4].

References

[1] Buja, A., D. Cook, D. Asimov, and C. Hurley (2005). Computational Methods for High-Dimensional Rotations in Data Visualization. In C. R. Rao, E. J.

Wegman, and J. L. Solka (Eds.), *Handbook of Statistics: Data Mining and Visualization*, pp. 391–413. Elsevier/North Holland.

[2] Cook, D. and D. Swayne (2007). *Interactive and Dynamic Graphics for Data Analysis with examples using R and GGobi*. New York: Springer. With contributions from Buja, A., Temple Lang, D., Hofmann, H., Wickham, H. and Lawrence, M. and additional data, R code and demo movies at `http://www.ggobi.org`.

[3] Wickham, H., D. Cook, H. Hofmann, and A. Buja (2011). tourr: An r package for exploring multivariate data with projections. *Journal of Statistical Software 40*(2), `http://www.jstatsoft.org/v40`.

[4] R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. `https://www.R-project.org/`

Nature-Inspired metaheuristic algorithms for finding optimal designs for high dimensional problems

Ray-Bing Chen[a], Ping Yang Chen[a], Guanghao Qi[b], Weng Kee Wong[c], and Zizhao Zhao[c]

[a] *Department of Statistics, National Cheng Kung University, Taiwan*
[b] *Department of Biostatistics, Johns Hopkins University, USA*
[c] *Department of Biostatistics, UCLA, USA*

ABSTRACT

Algorithms are practical ways to find optimal experimental designs. Most published work in the literature concerns optimal design problems for statistical models with a small number of factors. With big data, there are increasingly more design problems with many factors and current algorithms do not work well. I discuss nature-inspired metaheuristic algorithms potential for tackling high dimensional design problems and demonstrate their flexibility and efficiencies for finding several types of optimal designs for nonlinear models, including multiple-objective optimal designs and standardized maximin optimal designs for which the design criterion is non-differentiable and may involve three or more layers of optimization. If time permits, I will discuss specific applications to find standardized maximin optimal designs for enzyme kinetic models, adaptive designs for Phase II clinical trials, and optimal designs for generalized linear models with mixed factors.

Estimation and Classification in a Two-component Mixture Model with Stochastic Dominance

Jingjing Wu

University of Calgary, Canada

ABSTRACT

In this research, we studied a two-component nonparametric mixture model with stochastic dominance constraint, a model arises naturally from genetic studies. Our interest lies in both the estimation of mixing proportion and classification. For this model, we proposed and studies a nonparametric estimation based on cumulative distribution functions and a MLE through multinomial approximation. In order to incorporate nicely the stochastic dominance constraint, we introduced a semiparametric model for which we proposed and investigated both MLE and minimum Hellinger distance estimation (MHDE). We also proposed a hypothesis testing to test the validity of the semiparametric model. For the proposed methods, we investigated both their asymptotic properties such as consistency and asymptotic normality and their finite-sample performance through simulation studies and real data analysis.

Optimal Shrinkage Estimation in Heteroscedastic Hierarchical Models: beyond Gaussian

SAMUEL KOU

National University of Singapore and Harvard University, USA

ABSTRACT

Hierarchical models are powerful statistical tools widely used in scientific and engineering applications. The homoscedastic (equal variance) case has been extensively studied, and it is well known that shrinkage estimates, the James-Stein estimate in particular, offer nice theoretical (e.g., risk) properties. The heteroscedastic (the unequal variance) case, on the other hand, has received less attention, even though it frequently appears in real applications. It is not clear of how to construct ”optimal” shrinkage estimate. In this talk, we study this problem. In particular, we consider hierarchical linear models and models beyond Gaussian. We introduce a class of shrinkage estimates, constructed by minimizing an unbiased risk statistic. We show that this class is asymptotically optimal in the heteroscedastic case. We apply the estimates to real examples and observe competitive numerical results.

Key words: quadratic variance function, exponential family, location-scale family, linear regression, hierarchical linear model, asymptotic optimality

Brain network topological changes in neuropsychiatric disorders: from voxels to networks

JUAN (HELEN) ZHOU

Duke-NUS Medical School

ABSTRACT

Multimodal neuroimaging assays have demonstrated its usefulness in understanding large-scale neural networks in health and disease. Emerging network-sensitive neuroimaging techniques have allowed researchers to demonstrate that the spatial patterning of each neurodegenerative disease relates closely to a distinct functional intrinsic connectivity network (ICN), mapped in the healthy brain with task-free or "resting-state" functional magnetic resonance imaging (fMRI). This talk will describe our recent work on large-scale brain functional network topological changes using graph theory in persons at-risk for psychosis and its relevance with clinical conversion. Moreover, longitudinal inter-network connectivity and loss of functional segregation and its relationship with cognitive decline in healthy elderly will be highlighted. Lastly, we will discuss our recent work on linking dynamic functional connectivity states with vigilance fluctuations. Further developed, multimodal connectivity signatures may help us reveal disease mechanism and predict or track disease progression.

Multivariate Variable Selection: Beamforming-based Approach

JIAN ZHANG

University of Kent, UK

ABSTRACT

This article extends the idea of principal component analysis to multivariate variable selection for multivariate regression models, where regression coefficients of multiple responses on each predictor are treated as values derived from a random variable. The proposed method, called principal variable analysis, aims at select predictor variables with relatively higher coefficient variations.

The basic premise behind the proposal is to scan through a predictor variable space with a series of forward filters named null-beamformers; each is tailored to a particular region in the space and resistant to interference effects originating from other regions. The new approach attempts to explore the maximum amount of variation in the data with a small number of principal variables. Applying the proposal to simulated data and real cancer drug data, we show that it substantially outperforms the existing methods in terms of sensitivity and specificity. An asymptotic theory on selection consistency is established under some regularity conditions. We also extend the approach to the setting of multivariate nonparametric additive model. This talk is based on some unpublished joint work with my students Elaheh Oftadeh and Hui Ding.

References

[1] Zhang J. and Elaheh Oftadeh (2017), *Multivariate variable selection by means of null-beamforming*, KAR, University of Kent.

[2] Zhang J. and Liu C. (2015), *On linearly constrained minimum variance beamforming*, *Journal of Machine Learning Research*, **16**, 2099-2145.

Functional Data Analysis Using a Topological Summary Statistic: The Smooth Euler Characteristic Transform

Anthea Monod

Columbia University, USA

ABSTRACT

We introduce a novel statistic, the smooth Euler characteristic transform (SECT), which is designed to integrate shape information into regression models by representing shapes and surfaces as a collection of curves. Its construction is based on theory from topological data analysis (TDA). Due to its well-defined i nner p roduct structure, the SECT can be used in a wider range of functional and nonparametric modeling approaches than other previously proposed topological summary statistics. We provide mathematical properties of this statistic, notably, its injectivity, which is an implication for statistical sufficiency.

We illustrate the utility of the SECT in a radiomics context by showing that the topological quantification of t umors, assayed by magnetic resonance imaging (MRI), are better predictors of clinical outcomes in patients with glioblastoma multiforme (GBM). We show that topological features of tumors captured by the SECT alone explain more of the variance in patient survival than gene expression, volumetric features, and morphometric features.

References

[1] Crawford, L., Monod, A., Chen, A. X., Mukherjee, S. and Rabadán, R. (2017). Functional Data Analysis using a Topological Summary Statistic: the Smooth Euler Characteristic Transform. *arXiv:1611.06818.*

[2] Turner, K., Mukherjee, S. and Boyer, D. (2014). Persistent homology transform for modeling shapes and surfaces. *Inf. Inference*, **3**, 310–344.

[3] Monod, A., Kališnik Verovšek, S., Patiño-Galindo, J. Á. and Crawford, L. (2017). Tropical Sufficient Statistics for Persistent Homology. *arXiv:1709.02647.*

Pairwise Covariates-Adjusted Block Model for Community Detection

Yang Feng

Columbia University, USA

ABSTRACT

One of the most fundamental problems in network study is community detection. The stochastic block model (SBM) is one widely used model for network data with different estimation methods developed with their community detection consistency results unveiled. However, the SBM is restricted by the strong assumption that all nodes in the same community are stochastically equivalent, which may not be suitable for practical applications. We introduce pairwise covariates-adjusted stochastic block model (PCABM), a generalization of SBM that incorporates pairwise covariate information. In our model, the pairwise covariates can be constructed using any bivariate function of the corresponding covariates of the pair of nodes considered. We study the maximum likelihood estimators of the coefficients for the covariates as well as the community assignments. It is shown that both the coefficient estimates of the covariates and the community assignments are consistent under typical sparsity conditions. Spectral clustering with adjustment (SCWA) is introduced to efficiently solve PCABM. Under certain conditions, we derive the error bound of community estimation under SCWA and show that it is community detection consistent. PCABM compares favorably with SBM or degree-corrected stochastic block model (DCBM) under a wide range of simulated and real networks when covariate information is accessible. This is a joint work with Sihan Huang.

Optimal Covariance Change Point Detection in High Dimension

Daren Wang[1], Yi Yu[2], and Alessandro Rinaldo[1]

[1] *Carnegie Mellon University, USA*
[2] *University of Bristol, UK*

ABSTRACT

In this paper, we study covariance change point detection problem in high dimension. Specifically, we assume that the time series $X_i \in \mathbb{R}^p$, $i = 1, \ldots, n$ are independent p dimensional sub-Gaussian random vectors and that the corresponding covariance matrices $\{\Sigma_i\}_{i=1}^n$ are stationary within segments and only change at certain time points. Our generic model setting allows p grows with n and we do not place any additional structural assumptions on the covariance matrices. We introduce algorithms based on binary segmentation (e.g. Vostrikova, 1981) and wild binary segmentation (Fryzlewicz, 2014) and establish the consistency results under suitable conditions. To improve the detection performance in high dimension, we propose Wild Binary Segmentation through Random Projection (WBSRP). We show that WBSRP can optimally estimate the location of the change points. Our analysis also reveals a phase transition effect based on our generic model assumption and to the best of our knowledge, this type of results have not been established in elsewhere in the change point detection literature.

References

Fryzlewicz, P. (2014). Wild binary segmentation for multiple change-point detection. *The Annals of Statistics*, **42**, 2243–2281.

Vostrikova, L. (1981). Detection of the disorder in multidimensional random-processes. *Doklady Akademii Nauk SSSR*, **259**, 270–274.