Emerging New Topics in Functional Data Analysis

10 Jul 2023–21 Jul 2023, Institute for Mathematical Sciences, National University of Singapore

Organizing Committee

- Jialiang Li (National University of Singapore)
- Yehua Li (University of California, Riverside)
- David Nott (National University of Singapore)
- Jian Qing Shi (Southern University of Science and Technology and Newcastle University)
- Wanjie Wang (National University of Singapore)
- Jin-Ting Zhang (National University of Singapore)

Abstract Book

Speakers

Al Alawi, Maryam 4
Chen, Rong 4
Delaige, Aurore 5
Greven, Sonja 5
Hsing, Tailen 6
Jiang, Xuejun 6
Knei, Alois 7
Lee, Kuang-Yao 7
Li, Bing 8
Li, Cheng 8
Li, Yaguang 9
Li, Yehua 9
Lin, Zhenhua 10
Luo, Shan 10
Müller, Hans-Georg 11
Nott, David 11
Qu, Annie 12
Shang, Hanlin 12
Shi, Jian Qing 13
Tan, Siew Li Linda 13
Wang, Di 14
Wang, Jane-Ling 14
Wang, Lan 15
Wang, Wanjie 15
Xiang, Liming 16
Yang, Daewon 16
Yao, Qiwei 17
Zhang, Anderson 17
Zhang, Anru 18
Zhang, Jin-Ting 18

Zhou, Hang 19

Al Alawi, Maryam

Sultan Qaboos University

Title: A New Functional Data Clustering Technique Based on Spectral Clustering and Downsampling

Abstract:

We present a new framework for clustering functional data along with a new paradigm for performing model selection based on downsampling. Our clustering framework is a generalisatiion of the spectral clustering approach and is flexible enough to exploit higher order features of curves, including derivatives. Extensive comparative studies with existing methods show a clear advantage of our approach over existing functional data analysis clustering approaches. Additionally, we present a new paradigm for model selection, by introducing the technique of downsampling, which allows us to create lower resolution replicates of the observed curves. These replicates can then be used to provide insight into the tuning parameters for the specific clustering techniques. The usefulness of the proposed methods is illustrated through simulations and applications to real-life datasets.

Chen, Rong

Rutgers University

Title: Functional Quantile Autoregression

Abstract:

Conditional distribution or conditional quantiles provides important information in time series analysis. In this talk, we propose a new class of time series models, the functional quantile autoregression (FQAR) models, in which the conditional distribution of the observation at the current time point is affected by its past distributional information, and expressed as a functional of the past conditional quantile functions. The models can capture systematic influences of the past distributional information on the current conditional distribution, and therefore constitute a significant extension of traditional time series models in which the effect of conditioning information is confined to only a few selected characteristics of the past distribution. The FQAR models are similar to functional autoregression models. However, different from conventional functional time series models which are based on functionally observed data, the proposed model and method study functional dynamics in traditional time series data. In fact the FQAR models share some similar features to the GARCH model. However, instead of only focusing on the dynamic relationship of the second moments (conditional variance), we look at the relationship between the current conditional distribution to the past conditional distribution in a functional autoregressive form. Stationary conditions of the model are provided. We propose a sieve estimator for the model. Under stationarity conditions, the FQAR model converges and facilitates an estimation that does not depend on the initial quantile function. Identification is investigated. The asymptotic properties of the estimators are derived. A Monte Carlo experiment is conducted to investigate the finite sample performance of the proposed estimator and an empirical application to stock return time series highlights the proposed method.

Delaige, Aurore
University of Melbourne

Title: Analysing fragments of functional data

Abstract:

We consider functional data that are only observed on parts of their domains, in the form of fragments. We present methods for analysing these incomplete functional data and consider the problem of recovering unobserved parts of the curves. Several approaches can be used, for example using Markov assumptions or using a covariance completion procedure. We discuss and compare several approaches to this problem.

Greven, Sonja
Humboldt-Universität zu Berlin

Title: Statistical methods for curves, shapes and forms

Abstract:

When comparing the shape of, say, a specific outline marked on medical images across different patients or of cells across different biophysical simulations, the concrete coordinate system used for recording is often arbitrary and not of interest: the shape neither depends on positioning in space, nor on orientation or size. Analogously, the outline can be mathematically represented via a parameterized curve, but the particular parameterization of the outline curve is often not of interest, just its image.

We present statistical methods for datasets where an observational unit is the shape of a curve, defined as equivalence class a) over the shape invariances translation, rotation and scale and/or b) over re-parameterization, geometrically incorporating the invariances into the analysis. In particular, we develop approaches to compute means and distances for such data – useful also for e.g. classification or clustering – as well as regression methods, allowing for shapes to be observed as landmarks and/or potentially sparsely measured curves.

Hsing, Tailen

University of Michigan

Title: A functional-data perspective in spatial data analysis

Abstract:

More and more spatiotemporal data nowadays can be viewed as functional data. The first part of the talk focuses on the Argo data, which is a modern oceanography dataset that provides unprecedented global coverage of temperature and salinity measurements in the upper 2,000 meters of depth of the ocean. We discuss a functional kriging approach to predict temperature and salinity as a smooth function of depth, as well as a functional co-kriging approach of predicting oxygen concentration based on temperature and salinity data. The second part of the talk considers the nonparametric estimation of the spectral density of a spatial functional process. As is common in spatial statistics, the process is assumed to be observed at irregularly-spaced spatial locations. We propose a lag-window estimator and discuss its asymptotic properties, including optimality results assuming that the spectral density belongs to a certain class of functions.

Jiang, Xuejun

Southern University of Science and Technology

Title: Estimation and inference for ultra-high dimensional quasi-likelihood models based on data splitting

Abstract:

In this article, we develop a valid weighted estimation and inference framework for ultra-high dimensional quasi-likelihood models. The weighted estimator is obtained by minimizing the variance function. We split the full data into two subsets, conduct the model selection on one subset and compute the maximum quasi-likelihood estimator on the other subset. Then we aggregate the two estimators with the optimal weighted matrixes to form the final weighted estimator. With the weighted estimator, we construct the confidence intervals for the group components of the regression vector, and the Wald test for the linear structure of the group components. Theoretically, we establish the asymptotic normality of the weighted estimator, and the asymptotic χ^2-distribution of the corresponding Wald test without assuming model selection consistency. Advantages of the proposed tests are highlighted via theoretical and empirical comparison to some competitive tests, which guarantees that our proposed estimation and inference framework is locally optimal. In addition, when the selection consistency is achieved, we prove that the proposed Wald test is asymptotically identically distributed as the oracle tests in the sense that it knows the support of regression vector. Extensive simulations demonstrate more favorable finite sample performance of the proposed tests. An application to the breast cancer data illustrates the use of our proposed methodology.

Knei, Alois
University of Bonn

Title: Combining Concurrent and Functional Linear Regression

Abstract:

In many functional data applications, the domain of a function represents some time scale. In this context, a useful function-on-function linear regression model is the historical linear model, which explains the value of a response function at a given time by the values of a regressor function before that time. An alternative approach consists in the so-called concurrent model which at each time points assumes a direct linear dependence of the response on the current value of the predictor function. We propose a more general model which incorporates a concurrent as well as an historical component. Identifiability is studied in detail. An asymptotic theory is derived which provides rates of converge for all functional model parameters. The usefulness of the approach is illustrated by a real data application.

Lee, Kuang-Yao
Temple University

Title: Functional Directed Acyclic Graphs

Abstract:

In this work, we introduce a new method to estimate a directed acyclic graph (DAG) from multivariate functional data. We build on the notion of faithfulness that relates a DAG with a set of conditional independences among the random functions. We develop two linear operators, the conditional covariance operator and the partial correlation operator, to characterize and evaluate the conditional independence. Based on these operators, we adapt and extend the PC-algorithm to estimate the functional directed graph, so that the computation time depends on the sparsity rather than the full size of the graph. We study the asymptotic properties of the two operators, derive their uniform convergence rates, and establish the uniform consistency of the estimated graph, all of which are obtained while allowing the graph size to diverge to infinity with the sample size. We demonstrate the efficacy of our method through both simulations and an application to a time-course proteomic dataset.

(This is joint work with Lexin Li (UC Berkeley) and Bing Li (Penn State)).

Li, Bing

Penn State University

Title: Nonlinear function-on-function regression by RKHS

Abstract:

We propose a nonlinear function-on-function regression model where both the covariate and the response are random functions. The nonlinear regression is carried out in two steps: we first construct Hilbert spaces to accommodate the functional covariate and the functional response, and then build a second-layer Hilbert space for the covariate to capture nonlinearity. The second-layer space is assumed to be a reproducing kernel Hilbert space, which is generated by a positive definite kernel determined by the inner product of the first-layer Hilbert space for X--this structure is known as the nested Hilbert spaces. We develop estimation procedures to implement the proposed method, which allows the functional data to be observed at different time points for different subjects. Furthermore, we establish the convergence rate of our estimator as well as the weak convergence of the predicted response in the Hilbert space. Numerical studies including both simulations and a data application are conducted to investigate the performance of our estimator in finite sample.

Li, Cheng

National University of Singapore

Title: Bayesian fixed-domain asymptotics for covariance parameters in spatial Gaussian process regression models

Abstract:

Gaussian process models typically contain finite dimensional parameters in the covariance function that need to be estimated from the data. We study the Bayesian fixed-domain asymptotics for the covariance parameters in spatial Gaussian process regression models with an isotropic Matern covariance function, which has many applications in spatial statistics. For the model without nugget, we show that when the dimension of the domain is less than or equal to three, the microergodic parameter and the range parameter are asymptotically independent in the posterior. While the posterior of the microergodic parameter is asymptotically close in total variation distance to a normal distribution with shrinking variance, the posterior distribution of the range parameter does not converge to any point mass distribution in general. For the model with nugget, we derive new evidence lower bound and consistent higher-order quadratic variation estimators, which lead to explicit posterior contraction rates for both the microergodic parameter and the nugget parameter. We further study the asymptotic efficiency and convergence rates of Bayesian kriging prediction. All the new theoretical results are verified in numerical experiments and real data analysis.

Li, Yaguang
University of Science and Technology of China

Title: On Functional Processes with Multiple Discontinuities

Abstract:

We consider the problem of estimating multiple change points for a functional data process. There are numerous examples in science and finance in which the process of interest may be subject to some sudden changes in the mean. The process data that are not in a close vicinity of any change point can be analyzed by the usual nonparametric smoothing methods. However, the data close to change points and contain the most pertinent information of structural breaks need to be handled with special care. This paper considers a half-kernel approach that addresses the inference of the total number, locations, and jump sizes of the changes. Convergence rates and asymptotic distributional results for the proposed procedures are thoroughly investigated. Simulations are conducted to examine the performance of the approach, and a number of real data sets are analyzed to provide an illustration.

Li, Yehua
University of California, Riverside

Title: Bayesian Spatially Varying Coefficient Models with Functional Predictors

Abstract:

Reliable prediction for crop yield is crucial for economic planning, food security monitoring, and agricultural risk management. This study aims to develop a crop yield forecasting model at large spatial scales using meteorological variables closely related to crop growth. The influence of climate patterns on agricultural productivity can be spatially inhomogeneous due to local soil and environmental conditions. We propose a Bayesian spatially varying functional model (BSVFM) to predict county-level corn yield for five Midwestern states, based on annual precipitation and daily maximum and minimum temperature trajectories modeled as multivariate functional predictors. The proposed model accommodates spatial correlation and measurement errors of functional predictors, and respects the spatially heterogeneous relationship between the response and associated predictors by allowing the functional coefficients to vary over space. The model also incorporates a Bayesian variable selection device to further expand its capacity to accommodate spatial heterogeneity. The proposed method is demonstrated to outperform other highly competitive methods in corn yield prediction, owing to the flexibility of allowing spatial heterogeneity with spatially varying coefficients in our model. Our study provides further insights into understanding the impact of climate change on crop yield.

Lin, Zhenhua
National University of Singapore

Title: Statistical Inference for Functional Data via Bootstrapping

Abstract:

Statistical inference is of fundamental importance and yet challenging in functional data analysis. In response to the challenge, a set of powerful bootstrap-based procedures are developed for constructing simultaneous confidence bands for the mean function and coefficient functions in the varying coefficient model and for hypothesis testing related to the slope function in functional linear regression. Validity and consistency of the proposed procedures are established, and convergence rates are derived. The proposed procedures are shown to enjoy excellent numeric performance especially when the sample size is limited while the signal is relatively weak.

Luo, Shan
Shanghai Jiaotong University

Title: A Portmanteau Local Feature Discrimination Approach to the Classification with High-dimensional Matrix-variate Data

Abstract:

Matrix-variate data arise in many scientific fields such as face recognition,medical imaging, etc. Matrix data contain important structure information which can be ruined by vectorization. Methods incorporating the structure information into analysis have significant advantages over vectorization approaches. In this article, we consider the problem of two-class classification with high-dimensional matrix-variate data, and propose a novel portmanteau-local-feature discrimination (PLFD) method. This method first identifies local discrimination features of the matrix variate and then pools them together to construct a discrimination rule. We investigated the theoretical properties of the PLFD method and established its asymptotic optimality. We carried out extensive numerical studies including simulation and real data analysis to compare this method with other methods available in the literature, which demonstrate that the PLFD method has a great advantage over the other methods in terms of misclassification rate.

Müller, Hans-Georg
University of California, Davis

Title: Modeling Distributional Time Series

Abstract:

Distributional data are commonly encountered. Such data pose specific challenges as densities do not form a vector space and need to be estimated from samples that they generate. For one-dimensional distributions the Wasserstein space has become popular due to its inherent connection with optimal transport and the good statistical performance of the Wasserstein metric in applications. This has motivated the development of Wasserstein regression and other extrinsic approaches for the autoregressive modeling of distributional time series. An intrinsic approach based on a transport algebra provides an alternative, leading to an autoregressive transport model. Another alternative is to abandon the Wasserstein metric and instead work with the Fisher-Rao metric for distributions, which corresponds to the geodesic distance on the Hilbert sphere. This motivates a study of spherical autoregressive models, which can be based on rotations and are also applicable for some types of non-distributional data.

Nott, David
National University of Singapore

Title: Gaussian variational approximation for high-dimensional state space models

Abstract:

We consider a Gaussian variational approximation of the posterior density in high-dimensional state space models. The number of parameters in the covariance matrix of the variational approximation grows as the square of the number of model parameters, so it is necessary to find simple yet effective parametrisations of the covariance structure when the number of model parameters is large. We approximate the joint posterior density of the state vectors by a dynamic factor model, having Markovian time dependence and a factor covariance structure for the states. This gives a reduced description of the dependence structure for the states, as well as a temporal conditional independence structure similar to that in the true posterior. We illustrate the methodology on a spatio-temporal model for the spread of the Eurasian collared-dove across North America. Our approach compares favorably to a recently proposed ensemble Kalman filter method for approximate inference in high-dimensional hierarchical spatio-temporal models. This is joint work with Matias Quiroz and Robert Kohn.

Qu, Annie
University of California Irvine

Title: Individualized Dynamic Model for Multi-resolutional Data

Abstract:

Mobile health has emerged as a major success in tracking individual health status, due to the popularity and power of smartphones and wearable devices. This has also brought great challenges in handling heterogeneous, multi-resolution data which arise ubiquitously in mobile health due to irregular multivariate measurements collected from individuals. In this talk, we propose an individualized dynamic latent factor model for irregular multi-resolution time series data to interpolate unsampled measurements of time series with low resolution. One major advantage of the proposed method is the capability to integrate multiple irregular time series and multiple subjects by mapping the multi-resolution data to the latent space. In addition, the proposed individualized dynamic latent factor model is applicable to capturing heterogeneous longitudinal information through individualized dynamic latent factors. In theory, we provide the interpolation error bound of the proposed estimator and derive the convergence rate with non-parametric approximation methods. Both the simulation studies and the application to smartwatch data demonstrate the superior performance of the proposed method compared to existing methods.

Shang, Hanlin
Macquarie University

Title: Detecting structural breaks in high-dimensional functional time series

Abstract:

In this paper, we consider detecting breaks in heterogenous mean functions of high-dimensional functional time series which are allowed to be cross-sectionally correlated and temporally dependent. A new test statistic combining the functional CUSUM quantity and power enhancement component is proposed with asymptotic null distribution theory comparable to the conventional CUSUM theory derived for a single functional time series. In particular, the extra power enhancement component enlarges the region where the proposed test has power, and results in stable power performance when breaks are sparse in the alternative hypothesis. Furthermore, we impose a latent group structure on the subjects with heterogenous break points and introduce an easy-to-implement clustering algorithm with an information criterion to consistently estimate the unknown group number and membership. The estimated group structure can subsequently improve the convergence property of the homogenous break point estimate (within the group). Monte-Carlo simulation studies and an empirical application show that the proposed estimation and testing techniques have satisfactory performance in finite samples.

Shi, Jian Qing
Southern University of Science and Technology
Newcastle University

Title: Wrapped Gaussian Process Functional Regression Model for Batch Data on Riemannian Manifold

Abstract:

This talk will discuss concurrent functional regression models for batch data on Riemannian manifolds by estimating both mean structure and covariance structure simultaneously. The response variable is considered to follow a wrapped Gaussian process. Nonlinear relationship between manifold-valued response variables and multiple Euclidean covariates can be captured by this model in which the covariates could be either functional or scalar. Numerical results with both simulated data and real data will be presented to show the performance of the model.

Tan, Siew Li Linda
National University of Singapore

Title: Efficient data augmentation techniques for some classes of state space models

Abstract:

Data augmentation improves the convergence of iterative algorithms, such as the EM algorithm and Gibbs sampler by introducing carefully designed latent variables. In this article, we first propose a data augmentation scheme for the first-order autoregression plus noise model, where optimal values of working parameters introduced for recentering and rescaling of the latent states, can be derived analytically by minimizing the fraction of missing information in the EM algorithm. The proposed data augmentation scheme is then utilized to design efficient Markov chain Monte Carlo (MCMC) algorithms for Bayesian inference of some non-Gaussian and nonlinear state space models, via a mixture of normals approximation coupled with a block-specific reparametrization strategy. Applications on simulated and benchmark real datasets indicate that the proposed MCMC sampler can yield improvements in simulation efficiency compared with centering, noncentering and even the ancillarity-sufficiency interweaving strategy.

Wang, Di
Shanghai Jiaotong University

Title: High-Dimensional Vector Autoregression with Common Response and Predictor Factors

Abstract:

Interpreting the reduced-rank vector autoregressive (VAR) model of order one as a supervised factor model, where two factor modelings are simultaneously conducted to response and predictor spaces, this talk will introduce a new model, called vector autoregression with common response and predictor factors, to explore further the common structure between the response and predictors of a high-dimensional time series. The new model can provide better physical interpretations and improve estimation efficiency. In conjunction with the tensor operation, the model can easily be extended to any finite-order VAR models. A regularization-based method is considered for the high-dimensional estimation with the gradient descent algorithm, and its computational and statistical convergence guarantees are established. Simulation experiments confirm our theoretical findings, and a macroeconomic application showcases the appealing properties of the proposed model in structural analysis and forecasting. This talk is based on the joint work with Xiaoyu Zhang, Guodong Li, and Ruey S. Tsay.

Wang, Jane-Ling
University of California, Davis

Title: The trouble with sparse functional data

Abstract:

Functional data are random functions on an interval, e.g. [0, 1], hence they can be viewed as stochastic processes. They have become increasingly common due to advances in modern technology to collect and store such data. In reality these random functions can only be measured at discrete time grids and the measurement schedule may vary among subjects. Depending on the sampling frequency, functional data are collected either intensively or sparsely, which affects both methodology and theory. Furthermore, the data may contain noise, a.k.a. measurement errors. Among the various sampling plans, sparsely observed functional data that feature only a few measurements per subject are the most challenging type to deal with, both in methodology and theory. Such sparsely observed functional data are ubiquitous in longitudinal studies and require special handling. Although the challenges for mean and covariance estimation and many regression settings have been addressed for such sparse functional data, testing independence for multivariate functional data and testing the equality of distributions between two independent samples of functional data will be shown not to be feasible for sparse functional data. Instead, we show what is feasible for each problem and propose solutions for them. Both solutions involve non-standard methods to tackle the irregular sampling plan of sparse functional data and the associated measurement errors. Theory and numerical results will also be presented for both problems.

Wang, Lan
University of Miami

Title: Doubly Robust Quantile Off-policy Evaluation for Dynamic Data

Abstract:

We propose a doubly-robust estimation and inference procedure for quantile off-policy evaluation in sequential decision-making and study its asymptotic properties. In particular, we propose utilizing state-of-the-art deep conditional generative learning methods to handle parameter-dependent nuisance function estimation. We demonstrate the advantages of this proposed estimator through both simulations and a real-world dataset from a short-video platform. In particular, we find that our proposed estimator outperforms classical OPE estimators for the mean in settings with heavy-tailed reward distributions. (Joint work with Yang Xu, Chengchun Shi, Shikai Luo, and Rui Song)

Wang, Wanjie
National University of Singapore

Title: Temporal Ordering and Manifold Recovery on Noisy Data

Abstract:

The analysis of proteins and biological macromolecules is of great interest today, with the development of single-particle cryo-electron microscopy (cryo-EM). The observation Y_i follows $X_{t_i} + Noise$ because of the motions of the molecule. Since the motions repeated a hidden pattern, the ordering of t_i does not have the same ordering with i. Hence, a proper ordering of Y_i will largely improve the recovery of the functional $X(t)$.

In our work, we present a spectral method on the Laplacian matrix to order Y_i. We first reduce the noise in Y by taking the top eigenvectors of Y. Let Z be the matrix formed by these eigenvectors and we find the ordering of rows in Z. To do it, we first build the Gaussian kernel matrix on Z and then set L_Z to be the Laplacian of the kernel matrix. Ordering the second smallest eigenvector of Z will give the correct ordering of Y. We have set up the theoretical results to show consistency.

Xiang, Liming

Nanyang Technological University

Title: Conditional Quasi-likelihood Inference for Mean Residual Life Regression with Clustered Failure Time Data

Abstract:

In the analysis of clustered failure time data, frailty models based on the Cox proportional hazards model have been extensively studied by using frailty with a prespecified distribution to address the potential correlation of data within clusters. In this paper, we propose to analyze such correlated data under a frailty proportional mean residual life regression model using a novel conditional quasi-likelihood inference procedure. The proposed method utilizes stochastic process and the inverse probability of censoring weighting (IPCW) to form quasi-scores for regression parameters. Conditional inference based on the penalized quasi-likelihood is then developed to address within-cluster correlation without specifying the frailty distribution, bringing the method closer to what suffices for real-world applications. By incorporating the Buckley-James estimator of the censoring time survival function in the IPCW, our method further allows for dependent censoring. We establish asymptotic properties of the proposed estimator and evaluate its finite sample performance via simulation studies. An application to the data from a multi-institutional breast cancer study is finally presented as an illustration.

Yang, Daewon

Chungnam National University

Title: Nonparametric Bayesian covariate-dependent multivariate functional clustering: an application to time-series data for multiple air pollutants

Abstract:
Air pollution is a major threat to public health. Understanding the spatial distribution of air pollution concentration is of great interest to government or local authorities, as it informs about target areas for implementing policies for air quality management. Cluster analysis has been popularly used to identify groups of locations with similar profiles of average levels of multiple air pollutants, efficiently summarising the spatial pattern. This study aimed to cluster locations based on the seasonal patterns of multiple air pollutants incorporating the location-specific characteristics such as socio-economic indicators. For this purpose, we proposed a novel non-parametric Bayesian sparse latent factor model for covariate-dependent multivariate functional clustering. Furthermore, we extend this model to conduct clustering with temporal dependency. The proposed methods are illustrated through a simulation study and applied to time-series data for daily mean concentrations of ozone (O_3), nitrogen dioxide (NO_2), and fine particulate matter ($PM_{2.5}$) collected for 25 cities in Canada in 1986-2015.

Yao, Qiwei
London School of Economics

Title: Probabilistic Forecasting for Daily Electricity Loads and Quantiles for Curve-to-Curve Regression

Abstract:

Probabilistic forecasting of electricity load curves is of fundamental importance for effective scheduling and decision making in the increasingly volatile and competitive energy markets. We propose a novel approach to construct probabilistic predictors for curves (PPC), which leads to a new definition of quantiles in the context of curve-to-curve linear regression. There are three types of PPC: a predict set, a predictive band and a predictive quantile, and all of them are defined at a pre-specified nominal probability level. In the simulation study, the PPC achieve promising coverage probabilities under a variety of data generating mechanisms. When applying to one day ahead forecasting for the French daily electricity load curves, PPC outperform several state-of-the-art predictive methods in terms of forecasting accuracy, coverage rate and average length of the predictive bands. For example, PPC achieve up to 2.8-fold of the coverage rate with much smaller average length of the predictive bands. The predictive quantile curves provide insightful information which is highly relevant to hedging risks in electricity supply management.

Zhang, Anderson
University of Pennsylvania

Title: Spectral Methods for Learning from Pairwise Comparisons

Abstract:

The task of learning from pairwise comparisons is essential in various fields such as recommendation systems, sports, psychology, and social choices. Spectral methods have become increasingly popular for analyzing pairwise comparison data due to their simplicity and computational efficiency. In the first part of the talk, we focus on ranking n players using partial pairwise comparison data under the Bradley-Terry-Luce model. We analyze the Rank Centrality algorithm, a popular spectral method, and derive non-asymptotic expansions that quantify its entrywise uncertainty. This allows us to test the superiority of one player over another and construct confidence intervals for individual ranks. In the second part of the talk, we examine the problem of permutation synchronization, where permutation matrices are latent objects and their pairwise measurements are noisy and randomly missing. We propose a novel spectral method that can recover the permutations optimally and outperforms existing methods. To establish the optimality of our proposed method, we conduct a detailed spectral analysis and derive an exponential error bound that matches the minimax rate.

Zhang, Anru
Duke University

Title: Functional tensor SVD

Abstract:

In this talk, we present functional tensor SVD, a new framework for reducing dimension in tensors with one functional mode and multiple tabular modes. It addresses the challenge of high-order longitudinal data analysis. The model assumes the data as a discrete-time approximation of a low-rank functional tensor. Using tensor algebra and RKHS theory, we propose a RKHS-based iterative method with spectral initialization. The method accurately estimates the singular vectors and functions of the low-rank structure in the data. The error bounds of the algorithm are also established under mild assumptions. Experiments on both simulated and real data in longitudinal microbiome data analysis showcase its superiority.

Zhang, Jin-Ting
National University of Singapore

Title: A fast and accurate kernel-based independence test with applications to functional and high-dimensional data

Abstract:

Testing the dependency between two random variables is an important inference problem in statistics since many statistical procedures rely on the assumption that the two samples are independent. To test whether two samples are independent, a so-called HSIC (Hilbert-Schmidt Independence Criterion)-based test has been proposed. Its null distribution is approximated either by permutation or a Gamma approximation. Unfortunately, the permutation-based test is very time-consuming and the Gamma-approximation-based test does not work well for high-dimensional data. In this paper, a new HSIC-based test is proposed. Its asymptotic null and alternative distributions are established. It is shown that the proposed test is root-n consistent. A three-cumulant matched chi-squared approximation is adopted to approximate the null distribution of the test statistic. By choosing a proper reproducing kernel, the proposed test can be applied to many different types of data including multivariate, high-dimensional, and functional data. Three simulation studies and two real data applications show that in terms of level accuracy, power, and computational cost, the proposed test outperforms several existing tests for multivariate, high-dimensional, and functional data.

Zhou, Hang
University of California, Davis

Title: Theory of functional principal components analysis for discretely observed data

Abstract:

Functional data analysis is an important research field in statistics which treats data as random functions drawn from some infinite-dimensional functional space, and functional principal component analysis (FPCA) based on eigen-decomposition plays a central role for data reduction and representation. After nearly three decades of research, there remains a key problem unsolved, namely, the perturbation analysis of covariance operator for diverging number of eigencomponents obtained from noisy and discretely observed data. This is fundamental for studying models and methods based on FPCA, while there has not been substantial progress since Peter Hall et. al. 2006 result for a fixed number of eigenfunction estimates. In this work, we aim to establish a unified theory for this problem, deriving the moment bounds of eigenfunctions and asymptotic distributions of eigenvalues for a wide range of sampling schemes. We also propose a double truncation technique to derive the uniform convergence in time domain of estimated eigenfunctions. The technical arguments in this work are useful for handling the perturbation series with noisy and discretely observed data and can be applied in models or those involving inverse problems based on FPCA as regularization, such as functional linear regression.