复杂时间序列分析：高维性、变点、预报与因果性

Shcedule for complex time series analysis: high-dimensionality, change-points, forecasting, and causality, January 3-January 7, 2024

Time \ Date	Wednesday (January 3)	Thursday (January 4)	Friday (January 5)	Saturday (January 6)	Sunday (January 7)
7:30-8:30	*Breakfast (7:30-8:30) Opening Ceremony(8:30-8:40)*	*Breakfast*			
Chair	Xinghao Qiao	Dong Li	Qiwei Yao	Yuhao Wang	Zhaoxing Gao
8:40-9:20	Guodong Li	Wenyang Zhang	Peng Ding	Peng Ding	Ke Zhu
9:20-10:00	Weining Wang	Simone Giannerini	Yuhao Wang	Hanzhong Liu	Feiyu Jiang
10:00-10:30	*Coffee Break*	*Group Photo Coffee Break*	*Coffee Break*		
Chair	Qin Fang	Alexander Kreiss	Michael Eichler	Baojun Dou	Bin Guo
10:30-11:10	Flavio Ziegelmann	Xianyang Zhang	Daniel Pena	Alexander Kreiss	Zhaoxing Gao
11:10-11:50	Hanlin Shang	Guillem Rigaill	Zhou Zhou	Xiaojun Song	Di Wang
12:00-13:30	*Lunch*				
Chair	Dan Yang	Xiaojun Song	Free Discussion 13:30-17:00	Lilun Du	Bo Zhang
14:00-14:40	Yundong Tu	Zhao Chen		Zaytsev Alexey	Yutong Wang
14:40-15:20	Qin Fang	Dan Yang		Pengkun Yang	Zihan Wang
15:20-15:50	*Coffee Break*			*Coffee Break*	
Chair	Feiyu Jiang	Rongmao Zhang		Pengkun Yang	Cheng Yu
15:50-16:30	Greta Goracci	Cheng Yu		Dan Wang	Michael Eichler
16:30-17:10	Yaxing Yang	Baojun Dou		Jiebo Song	Yuxin Tao
17:30-19:00	*Dinner*		Banquet 18:00-20:00	*Dinner*	

Titles and Abstracts

An efficient tensor regression for high-dimensional data

Guodong Li
The University of Hong Kong, Hong Kong SAR

Most currently used tensor regression models for high-dimensional data are based on Tucker decomposition, which has good properties but loses its efficiency in compressing tensors very quickly as the order of tensors increases, say greater than four or five. However, for the simplest tensor autoregression in handling time series data, its coefficient tensor already has the order of six. This paper revises a newly proposed tensor train (TT) decomposition and then applies it to tensor regression such that a nice statistical interpretation can be obtained. The new tensor regression can well match the data with hierarchical structures, and it even can lead to a better interpretation for the data with factorial structures, which are supposed to be better fitted by models with Tucker decomposition. More importantly, the new tensor regression can be easily applied to the case with higher order tensors since TT decomposition can compress the coefficient tensors much more efficiently. The methodology is also extended to tensor autoregression for time series data, and nonasymptotic properties are derived for the ordinary least squares estimations of both tensor regression and autoregression. A new algorithm is introduced to search for estimators, and its theoretical justification is also discussed. Theoretical and computational properties of the proposed methodology are verified by simulation studies, and the advantages over existing methods are illustrated by two real examples.

Inference on derivatives of high dimensional regression function with deep neural network

Weining Wang
University of Groningen, Netherlands

We present a significance test for any given variable in nonparametric regression with many variables via estimating derivatives of a nonparametric function. The test is based on the moment generating function of the partial derivative of an estimator of the regression function, where the estimator is a deep neural network whose structure is allowed to become more complex as the sample size grows. This test finds applications in model specification and variable screening for high-dimensional data. To render our test applicable to high-dimensional inputs, whose dimensions can also increase with sample size, we make the assumption that the observed high-dimensional predictors can effectively serve as proxies for certain latent, lower-dimensional predictors that are actually involved in the regression function. Additionally, we finely adjust the regression function estimator, enabling us to achieve the desired asymptotic normality under the null hypothesis, as well as consistency for any fixed scenarios and certain local alternatives.

Multivariate functional time series: dimension reduction and risk forecasting

Flavio Ziegelmann
Federal University of Rio Grande do Sul, Brazil

In modern days, the accurate prediction and forecasting of risk measures, such as Value at Risk (VaR) and Expected Shortfall (ES), is an essential task for asset market managers. When calculating risk measures, an essential step, for most approaches, is to estimate the probability density function (PDF) of asset returns. A daily sequence of intraday return densities of p assets, denoted by $\mathbf{Y}_t, t = 1, \ldots, n$, can be seen as a p-dimensional functional time series. If p is large ($\mathbf{Y}_t$ is high dimensional), then one has to perform a two-way dimension reduction: in the high dimensional vector and in the infinite dimensional curves. Guo, Qiao and Wang (2023) introduce an approach that can be employed for this two-way dimension reduction. Here we propose a modification of their work by combining their Functional Factor Model (FFM) with the univariate Dynamic Functional Principal Components Analysis (DFPCA) method by Bathia, Yao and Ziegelmann (2010) as a two way reduction approach. On top of that we add the following steps: i) construct the functional time series from standardised intraday returns of p assets, obtained via a non-functional highfrequency ARMA-GARCH model, obtaining estimated daily PDF curves; ii) employ the non-functional model from step i) along with the predicted curves to simulate (via a MC Rejection Sampling method) new high-frequency predicted data; iii) use the simulated data from step ii) to forecast future daily risk measures as in Horta and Ziegelmann (2018).

Detection and estimation of structural breaks in high-dimensional functional time series

Han-Lin Shang
Macquarie University, Australia

In this paper, we consider detecting and estimating breaks in heterogeneous 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 statistic 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 heterogeneous 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 post-clustering break point estimate. Monte-Carlo simulation studies and empirical applications show that the proposed estimation and testing techniques have satisfactory performance in finite samples.

A tale of two types of structural instabilities in high-dimensional factor models

Yundong Tu,
Peking University, China

With the increasing availability of large data sets in economics and finance, the large factor model has become one of the most important tools to achieve dimension reduction in the statistical and econometric analysis. To capture the instability caused by economic condition shifts or policy reforms, factor models with structural breaks in the factor loadings are accordingly developed. On the other hand, recurring regime shifts that relate to higher frequency recurring fluctuation arise in situation where ``history repeats", and are conveniently described by threshold factor models, which allow recurring regime shifts in the factor loadings according to the magnitude of a

(continuous) threshold variable. In practice, it is often difficult to decide whether structural break or threshold effect, or both types of instabilities one should employ to portray the observed data. This talk shall discuss how to model each type of instability in factor analysis separately first, and then provide a solution to distinguish the two categories in a model that simultaneously allow both types of structural instabilities. The proposed models are estimated by machine learning techniques such as group Lasso, backward elimination algorithms and information criterion-based model selection methods. The associated asymptotic properties are established and are corroborated by finite sample simulation results and empirical examples. This talk is based on joint projects with Chenchen Ma, who is currently a Ph.D. candidate at the Center of Statistical Science, Peking University.

Autoregressive networks with dependent edges

Qin Fang
The University of Sydney, Australia

We propose a general AR(m) network model for dynamic network processes in which the edges change depending on the previous m observed networks. We investigate the estimation consistency (MLE) and asymptotic normality of the general framework, and further propose three examples of such networks to depict more explicitly some stylized features often observed in real network data. Our simulation results indicate that the proposed models are capable to simulate and to reflect some observed interesting dynamic network phenomena, and the general estimation method works well.

Robust estimation for Threshold Autoregressive Moving-Average models

Greta Goracci
Free University of Bozen-Bolzano, Italy

Threshold autoregressive moving-average (TARMA) models extend the popular TAR model and are among the few parametric time series specifications to include a moving average in a non-linear setting. The state dependent reactions to shocks is particularly appealing in Economics and Finance. However, no theory is currently available when the data present heavy tails or anomalous observations. Here we provide the first theoretical framework for robust M-estimation for TARMA models and study its practical relevance. Under mild conditions, we show that the robust estimator for the threshold parameter is super-consistent, while the estimators for autoregressive and moving-average parameters are strongly consistent and asymptotically normal. The Monte Carlo study shows that the M-estimator is superior, in terms of both bias and variance, to the least squares estimator, which can be heavily affected by outliers. The findings suggest that robust M-estimation should be generally preferred to the least squares method. We apply our methodology to a set of commodity price time series; the robust TARMA fit presents smaller standard errors and superior forecasting accuracy. The results support the hypothesis of a two-regime non-linearity characterised by slow expansions and fast contractions.

A multiple-regime threshold GARCH(1,1) model: structure and its estimation

Yaxing Yang
Xiamen University, China

This paper studies a multiple-regime threshold GARCH (1,1) (MTGARCH) model. The sufficient and necessary condition is obtained for a strictly stationary and ergodic solution to the model. It turns out that this condition only depends on the two extreme regimes and the space of parameters satisfying this condition is possibly unbounded. The sufficient and necessary condition of its invertibility condition is also obtained. Based on this, the paper further studies the quasi-maximum likelihood estimator (QMLE) of parameters in the MTGARCH model when threshold parameters are unknown. Under only stationarity and invertibility conditions which include the case with the infinite-variance MTGARCH process, it is shown that the estimated thresholds are n-consistent and asymptotically independent, each of which converges weakly to a functional of a two-sided compound Poisson process. The QMLE of the remaining parameters are root n consistent and asymptotically multivariate normal. Monte Carlo simulation studies are carried out to assess the performance of our procedure in finite samples and an empirical example is given to illustrate the usefulness of MTGARCH models.

Estimation of low-rank high-dimensional multivariate linear models for multi- response data

Wenyang Zhang
University of Macau, Macau SAR

In this talk, I will focus on low rank high-dimensional multivariate linear models (LRMLM) for high-dimensional multi-response data. I will present an intuitively appealing estimation approach together with an implementation algorithm. I will show the asymptotic properties of the estimation method to justify the estimation procedure theoretically. Intensive simulation study results will be presented to demonstrate the performance of the proposed method when the sample size is finite, and a comparison will be made with some popular methods from the literature. I will show the proposed estimator outperforms all of the alternative methods under various circumstances. Finally, I will apply the LRMLM together with the proposed estimation to analyze an environmental dataset and predict concentrations of PM2.5 at the locations concerned. I will illustrate how the proposed method provides more accurate predictions than the alternative approaches.

Consistent and efficient model selection with possible misspecification for vec- tor time series

Simone Giannerini
University of Bologna, Italy

The Misspecification-Resistant Information Criterion (MRIC) proposed in H.-L. Hsu, C.-K. Ing, H. Tong:*On model selection from a finite family of possibly misspecified time series models* . The Annals of Statistics. 47 (2), 1061--1087 (2019), is a model selection criterion for univariate parametric time series that enjoys both the property of consistency and asymptotic efficiency. Its appealing properties make it an ideal tool for time series model selection but, to date, only the univariate response case has been studied. In this article we extend the MRIC to the multivariate time series case. We obtain an asymptotic expression for the mean squared prediction error matrix, we define the vectorial MRIC, and prove the consistency of its method-of-moments estimator. Moreover, we prove its asymptotic efficiency. We discuss the conditions of applicability of the vectorial MRIC for possibly misspecified vector autoregressive models with exogenous variables (VARX) and present a fully worked out example that highlights the need to provide a model

selection criterion for multivariate time series that accounts for misspecification. Finally, we discuss the usage of the criterion in the high-dimensional setting.

High-dimensional change-point detection using generalized homogeneity metrics

Xianyang Zhang
Texas A&M University, USA

Change-point detection has been a classical problem in statistics and econometrics. This work focuses on the problem of detecting abrupt distributional changes in the data-generating distribution of a sequence of high-dimensional observations, beyond the first two moments. This has remained a substantially less explored problem in the existing literature, especially in the high-dimensional context, compared to detecting changes in the mean or the covariance structure. We develop a nonparametric methodology to (i) detect an unknown number of change-points in an independent sequence of high-dimensional observations and (ii) test for the significance of the estimated change-point locations. Our approach essentially rests upon nonparametric tests for the homogeneity of two high-dimensional distributions. We construct a single change-point location estimator via defining a cumulative sum process in an embedded Hilbert space. As the key theoretical innovation, we rigorously derive its limiting null distribution and prove its consistency under the high dimension medium sample size (HDMSS) framework. Subsequently we combine our statistic with the idea of wild binary segmentation to recursively estimate and test for multiple change-point locations. The superior performance of our methodology compared to other existing procedures is illustrated via extensive simulation studies as well as over stock prices data observed during the period of the global financial crisis in the United States.

Online multivariate changepoint detection: leveraging links with computa- tional geometry

Guillem Rigaill
National Research Institute for Agriculture, Food and Environment (INRAE), France

The increasing volume of data streams poses significant computational challenges for detecting changepoints online. Likelihood-based methods are effective, but their straightforward implementation becomes impractical online. We develop two online algorithms that exactly calculate the likelihood ratio test for a single changepoint in p-dimensional data streams by leveraging fascinating connections with computational geometry. Our first algorithm is straightforward and empirically quasi-linear. The second is more complex but provably quasi-linear: $\mathcal{O}(n\log(n)^{p+1})$ for $_n$data points. Through simulations, we illustrate, that they are fast and allow us to process millions of points within a matter of minutes up to $p = 5$.

Inference on functional-coefficient double AR model

Zhao Chen
Fudan University, China

The double autoregressive model (DAR) is an advanced statistical model that has gained significant attention in the field of time series analysis. However, the simple linear structure of the autoregressive component fails to adequately capture the variability and complexity of real-world data. To address this limitation, we propose a nonlinear structure and explore nonparametric

inference for the functional-coefficient double autoregressive (FDAR) model. We first consider the stationary conditions of the FDAR model and subsequently develop profile local quasi-maximum exponential likelihood (PL-QMELE) estimators. The corresponding asymptotic properties are established. Furthermore, we present hypothesis tests for functional coefficients and conditional heteroskedasticity respectively. The simulation study demonstrate that the estimator satisfactorily exhibits the expected asymptotic properties. Moreover, the FDAR model outperforms the DAR model in capturing the nonlinear structure of the data. To illustrate our findings, we provide an example involving the Shanghai Stock Exchange Index.

Factor models for high-dimensional tensor time series

Dan Yang
The University of Hong Kong, Hong Kong SAR

Large tensor (multi-dimensional array) data routinely appear nowadays in a wide range of applications, due to modern data collection capabilities. Often such observations are taken over time, forming tensor time series. In this article we present a factor model approach to the analysis of high-dimensional dynamic tensor time series and multi-category dynamic transport networks. This article presents two estimation procedures along with their theoretical properties and simulation results. We present two applications to illustrate the model and its interpretations.

Large covariance matrix estimation with factor-assisted variable clustering

Cheng Yu
Tsinghua University, China

In the field of large covariance matrix estimation, several methods have been developed based on the factor models, assuming the existence of a few common factors that can explain the co-movement of asset pricing. However, many studies have demonstrated the presence of cross-sectional correlation between assets after removing the common factors. To account for this effect, we propose an approximate observable factor model with latent cluster structure, along with a three-step estimator to accurately estimate the large covariance matrix for high-dimensional time series. The rates of convergence of the residual covariance with latent cluster structure and the whole large covariance matrix are studied under various norms. Additionally, we introduce a novel ratio-based criteria for determining the latent cluster structure, which can achieve clustering consistency with probability approaching to one. The asymptotic results are supported by simulation studies, and we demonstrate the practical application of our approach through real data analysis on minimal variance portfolio allocation.

High-dimensional spatio-temporal autoregressive models for matrix-valued time series

Baojun Dou
City University of Hong Kong, Hong Kong SAR

TThis paper considers the modelling of spatio-temporal matrix time series data, which is motivated by the prediction of daily trading volume curve for various assets, served as the key input to certain execution algorithms, volume-weighted average price (VWAP) in particular, used by major execution brokers to process large buy or sell orders on behalf of their institutional

clients. Two subclasses of models will be considered. The first originates from spatial econometric conventions by pre-determining the weight matrices. The second sub-class sidesteps the complexities of predetermined weight matrices, allowing these matrices to be unknown parameters with certain banded sparse structures derived from practical applications. Due to the innate endogeneity, we apply the iterated least squares method based on Yule-Walker equations for estimation. Theories and asymptotic results for the proposed methods are established for both fixed and high dimensions. The proposed methodology is further illustrated using both simulated and real data sets. This is a joint work with Qiwei Yao (LSE), Jinyuan Chang (SWUFE), Jing He (SWUFE) and Sudhir Tiwari (CITIC Securities).

Causal inference in network experiments: regression-based analysis and design- based properties

Peng Ding
University of California, Berkeley, USA

Investigating interference or spillover effects among units is a central task in many social science problems. Network experiments are powerful tools for this task, which avoids endogeneity by randomly assigning treatments to units over networks. However, it is non-trivial to analyze network experiments properly without imposing strong modeling assumptions. Previously, many researchers have proposed sophisticated point estimators and standard errors for causal effects under network experiments. We further show that regression-based point estimators and standard errors can have strong theoretical guarantees if the regression functions and robust standard errors are carefully specified to accommodate the interference patterns under network experiments. We first recall a well-known result that the Hajek estimator is numerically identical to the coefficient from the weighted-least-squares fit based on the inverse probability of the exposure mapping. Moreover, we demonstrate that the regression-based approach offers three notable advantages: its ease of implementation, the ability to derive standard errors through the same weighted-least-squares fit, and the capacity to integrate covariates into the analysis, thereby enhancing estimation efficiency. Furthermore, we analyze the asymptotic bias of the regression-based network-robust standard errors. Recognizing that the covariance estimator can be anti-conservative, we propose an adjusted covariance estimator to improve the empirical coverage rates. Although we focus on regression-based point estimators and standard errors, our theory holds under the design-based framework, which assumes that the randomness comes solely from the design of network experiments and allows for arbitrary misspecification of the regression models.

Long-term causal inference under persistent confounding via data combination

Yuhao Wang
Tsinghua University, China

We study the identification and estimation of long-term treatment effects when both experimental and observational data are available. Since the long-term outcome is observed only after a long delay, it is not measured in the experimental data, but only recorded in the observational data. However, both types of data include observations of some short-term outcomes. In this paper, we uniquely tackle the challenge of persistent unmeasured confounders, i.e., some unmeasured confounders that can simultaneously affect the treatment, short-term outcomes and the long-term outcome, noting that they invalidate identification strategies in previous literature. To address this challenge, we exploit the sequential structure of multiple short-term outcomes, and

develop three novel identification strategies for the average long-term treatment effect. We further propose three corresponding estimators and prove their asymptotic consistency and asymptotic normality. We finally apply our methods to estimate the effect of a job training program on long-term employment using semisynthetic data. We numerically show that our proposals outperform existing methods that fail to handle persistent confounders.

Detecting outliers in high dimensional time series by dynamic factor models

Daniel Pena
Universidad Carlos III de Madrid, Spain

A procedure to detect outliers in a large collection of time series is presented. The high-dimensional setting is handled by assuming that the time series have been generated by a dynamic factor model and that outliers can appear either in the latent factors or in the idiosyncratic noise. The factor outliers affect all or many of the time series whereas the idiosyncratic outliers affect only a few, or just one of the observed time series. These two types of outliers can be fairly well detected by projecting the series on the factor and idiosyncratic spaces constructed from robust estimates of the factor loading matrix. We propose an efficient procedure based on these linear transformations for detecting outliers. The behavior of the procedure is illustrated with simulations and the analysis of a real data example.

Simultaneous sieve inference for time-inhomogeneous nonlinear time series regression

Zhou Zhou
University of Toronto, Canada

In this talk, we consider the time-inhomogeneous nonlinear time series regression for a general class of locally stationary time series. On one hand, we propose sieve nonparametric estimators for the time-varying regression functions. On the other hand, we develop a unified simultaneous inferential theory which can be used to conduct both structural and exact form testings on the functions. Our proposed statistics are powerful under locally weak alternatives. We also propose a multiplier bootstrapping procedure for practical implementation. Our methodology and theory do not require any structural assumptions on the regression functions and we also allow the functions to be supported in an unbounded domain. We also establish a Gaussian approximation result for affine and quadratic forms for high dimensional locally stationary time series, which can be of independent interest. Numerical sim- ulations and a real financial data analysis are provided to support our results.

Flexible sensitivity analysis for causal inference in observational studies subject to unmeasured confounding

Peng Ding
University of California, Berkeley, USA

Causal inference with observational studies often suffers from unmeasured confounding, yielding biased estimators based on the unconfoundedness assumption. Sensitivity analysis assesses how the causal conclusions change with respect to different degrees of unmeasured confounding. Most existing sensitivity analysis methods work well for specific types of estimation or testing

strategies. We propose a flexible sensitivity analysis framework that can deal with commonly-used inverse probability weighting, outcome regression, and doubly robust estimators simultaneously. It is based on the well-known parametrization of the selection bias as comparisons of the observed and counterfactual outcomes conditional on observed covariates. It is attractive for practical use because it only requires simple modifications of the standard estimators. Moreover, it naturally extends to many other causal inference settings, including the average treatment effect on the treated units and studies with survival outcomes. We also develop an R package saci that implements our sensitivity analysis estimators.

Rerandomization and covariate adjustment in split-plot designs

Hanzhong Liu
Tsinghua University, China

The split-plot design arises from agricultural sciences with experimental units, also known as subplots, nested within groups known as whole plots. It assigns the whole-plot intervention by a cluster randomization at the whole-plot level and assigns the subplot intervention by a stratified randomization at the subplot level. The randomization mechanism guarantees covariate balance on average at both the whole-plot and subplot levels, and ensures consistent inference of the average treatment effects by the Horvitz--Thompson and Hajek estimators. However, covariate imbalance often occurs in finite samples and subjects subsequent inference to possibly large variability and conditional bias. Rerandomization is widely used in the design stage of randomized experiments to improve covariate balance. The existing literature on rerandomization nevertheless focuses on designs with treatments assigned at either the unit or the group level, but not both, leaving the corresponding theory for rerandomization in split-plot designs an open problem. To fill the gap, we propose two strategies for conducting rerandomization in split-plot designs based on the Mahalanobis distance and establish the corresponding design-based theory. We show that rerandomization can improve the asymptotic efficiency of the Horvitz--Thompson and Hajek estimators. Moreover, we propose two covariate adjustment methods in the analysis stage, which can further improve the asymptotic efficiency when combined with rerandomization. The validity and improved efficiency of the proposed methods are demonstrated through numerical studies.

Testing for global covariate effects in dynamic interaction event networks

Alexander Kreiss
Leipzig University, Germany

In statistical network analysis it is common to observe so called interaction data. Such data is characterized by actors forming the vertices and interacting along edges of the network, where edges are randomly formed and dissolved over the observation horizon. In addition covariates are observed and the goal is to model the impact of the covariates on the interactions. We distinguish two types of covariates: global, system-wide covariates (i.e. covariates taking the same value for all individuals, such as seasonality) and local, dyadic covariates modeling interactions between two individuals in the network. In the talk we will firstly discuss existing parametric and non-parametric models using counting process to model such data. Then, secondly, we will extended those models to allow for comparing a completely parametric model and a model that is parametric only in the local covariates but has a global non-parametric time component. This allows, for instance, to test whether global time dynamics can be explained by simple global covariates like weather, seasonality etc. The procedure is applied to a bike-sharing network by

using weather and weekdays as global covariates and distances between the bike stations as local covariates.

Significance testing in nonparametric autoregression

Xiaojun Song
Peking University, China

In this paper, we propose significance tests in nonparametric autoregression. Under the null, forecast of any nonlinear autoregression of order p is unaffected by considering any extra lagged value. A necessary and sufficient condition, which forms a basis for the tests, is that the residuals of the p-th order nonparametric autoregression are uncorrelated with any measurable function of the lagged variables. The test statistic is based on Fourier transform of the autocorrelation function of the nonparametric residuals and functions of the lagged values. The tests are implemented with the assistance of a bootstrap technique. We illustrate the practical performance of the test by means of simulations and an empirical application.

How to break the curse of dimensionality for change point detection: empirical evidence from neural networks

Zaytsev Alexey
Skoltech, Russia

The ‘curse of dimensionality’ is evident for change point detection methods, as they often rely on kernel methods and requires projection to work for high dimension observations. However, little is known, on how to design a good kernel or a good projection. We adopt deep neural networks as a remedy in such problems. Our results show, that this approach works well in supervised and unsupervised scenario. We present experiments for change point detection in video and human activity recognition, adding to the discussion on the benchmarks for change point detection methods in high dimension scenario.

Two phases of scaling laws for nearest neighbor classifiers

Pengkun Yang
Tsinghua University, China

A scaling law refers to the observation that the test performance of a model improves as the number of training data increases. A fast scaling law implies that one can solve machine learning problems by simply boosting the data and the model sizes. Yet, in many cases, the benefit of adding more data can be negligible. In this work, we study the rate of scaling laws of nearest neighbor classifiers. We show that a scaling law can have two phases: in the first phase, the generalization error depends polynomially on the data dimension and decreases fast; whereas in the second phase, the error depends exponentially on the data dimension and decreases slowly. Our analysis highlights the complexity of the data distribution in determining the generalization error. When the data distributes benignly, our result suggests that nearest neighbor classifier can achieve a generalization error that depends polynomially, instead of exponentially, on the data dimension.

Portfolio selection through functional linear regression

Dan Wang
New York University Shanghai, China

For portfolio management, Markowitz's mean-variance (MV) theory provides the method to construct optimal portfolios. However, in a market with huge amounts of available assets and rapidly changing dynamics, the allocation weights obtained by MV theory are imprecise, which leads to portfolios far inferior to the optimal ones. To improve the precision, existing literature suggests choosing a subgroup from a large number of assets for portfolio selection. In this paper, we provide a new approach to facilitate selection decisions and improve assets allocation. Instead of following the market index as the benchmark, we create a new dynamic index which outperforms the market averages and select assets by tracking it. To capture market dynamics and overcome high-dimensional challenges, we propose to use a functional linear regression model to extract more information from intraday high frequency trading data, and to adopt functional sure independence screening (FIS) to select assets. Then we construct portfolios with assets from the Chinese stock market from 2013 to 2016, which includes both bull and bear markets. Compared to the bench- mark market average, the new portfolio allocation based on only 6-month data, has better accumulative returns, maximum drawdown, and Sharpe ratio. We also provide theoretical and numerical studies about the FIS method.

Application of time series prediction in Beijing healthcare insurance

Jiebo Song
Beijing Institute of Mathematical Sciences and Applications (BIMSA), China

Through the analysis of historical time-series data of medical insurance records from healthcare institutions in Beijing, and considering public health events, Medicare policies, and the actual development of healthcare institutions, we have developed the Beijing Healthcare Insurance Security - Global Budget Index (BJ- GBI) model. According to the characteristics of historical data, we employ three distinct types of time series prediction models for these institutions, namely ARIMA, Prophet and the Average Growth Rate Assignment Model. This methodology assists the Beijing Municipal Medical Insurance Bureau in forecasting the total amount of medical insurance funds for the next year and provides decision support for global budget management.

Inference for panel ARMA-GARCH model

Ke Zhu
The University of Hong Kong, Hong Kong SAR

We propose a panel ARMA-GARCH model to capture the dynamics of a large panel of data. For this model, we provide a two-step estimation to estimate the ARMA parameters and GARCH parameters separately. Under some regular conditions, we show that all of the proposed estimators are asymptotically to normal with the convergence rate $(NT)^{-1/2}$, and they have the asymptotic biases when both N and T diverge to infinity at the same rate. Particularly, we find that the asymptotic biases can come from the fixed effect, estimation effect, and unobservable initial values. To correct the bias, we further propose the bias-corrected version of estimators by using either analytical asymptotics or jackknife method. Our asymptotic results are based on a new

central limit theorem for the linear-quadratic form in either mixing or martingale difference sequence. Simulations and one real example are given to demonstrate the usefulness of our panel ARMA-GARCH model.

Model-free change point detection using modern classifiers

Feiyu Jiang
Fudan University, China

In contemporary data analysis, it is increasingly common to work with complex datasets that are non-stationary. These datasets typically extend beyond the classical low-dimensional Euclidean space, making it challenging to detect shifts in their distribution without relying on strong structural assumptions. In this paper, we introduce a novel offline change-point detection method that leverages modern classifiers developed in the machine learning community. With suitable data split- ting, the test statistic is constructed through the sequential computation of the Area Under the Curve (AUC) of a classifier, which is trained on data segments before and after a potential change-point. It is shown that the resulting AUC process attains its maxima at the true change-point location, which then serves the goal of change- point estimation. Our proposed method enjoys several appealing advantages. It is entirely nonparametric, highly versatile, quite flexible, and does not require stringent assumptions about the nature of the data or the distributional shift. Theoretically, we derive the limiting pivotal distribution of the proposed test statistics under null, as well as the asymptotical behaviors under both local and fixed alternatives. The weak consistency of the change-point estimator is also provided. To improve the finite sample performance of the test, we introduce a conditional version of the permutation test to assess significance. Extensive simulation studies and the analysis of two real-world datasets illustrate the superior performance of our approach compared to existing model-free change-point detection methods.

Supervised dynamic PCA: linear dynamic forecasting with many predictors

Zhaoxing Gao
Zhejiang University, China

This paper proposes a novel dynamic forecasting method using a new supervised Principal Component Analysis (PCA) when a large number of predictors are available. The new supervised PCA provides an effective way to bridge the gap between predictors and the target variable of interest by scaling and combining the predictors and their lagged values, resulting in an effective dynamic forecasting. Unlike the traditional diffusion-index approach, which does not learn the relationships between the predictors and the target variable before conducting PCA, we first rescale each predictor according to their significance in forecasting the targeted variable in a dynamic fashion, and a PCA is then applied to a re-scaled and additive panel, which establishes a connection between the predictability of the PCA factors and the target variable. We also propose to use penalized methods such as the LASSO to select the significant factors that have superior predictive power over the others. Theoretically, we show that our estimators are consistent and outperform the traditional methods in prediction under some mild conditions. We conduct extensive simulations to verify that the proposed method produces satisfactory forecasting results and outperforms most of the existing methods using the traditional PCA. An example of predicting U.S. macroeconomic variables using a large number of predictors showcases that our method fares better than most of the existing ones in applications.

High-dimensional vector autoregression with common response and predictor factors

Di Wang
Shanghai Jiao Tong University, China

The reduced-rank vector autoregressive (VAR) model can be interpreted as a supervised factor model, where two factor modelings are simultaneously applied to response and predictor spaces. This talk introduces a new model, called vector autoregression with common response and predictor factors, to explore further the common structure between the response and predictors in the VAR framework. 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 model. 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. For data with pervasive cross-sectional dependence, a transformation for responses is developed to alleviate the diverging eigenvalue effect. Moreover, we consider additional sparsity structure in factor loading for the case of ultra-high dimension. 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.

Autoregressive networks: sparsity and degree heterogeneity

Yutong Wang
London School of Economics, UK

We studied a type of dynamic network model where edges between each pair of nodes change over time independently. In this model, the connecting probability and disconnecting probability are allowed to go to 0 as the number of nodes goes to infinity. We proposed a two-step estimation strategy, where the first step is MLE and the second step is an M-estimation followed by the uniqueness of the solution. To evaluate the convergence rate of the estimator, we construct a concentration via the 'Cantor set construction' trick for strongly correlated stochastic processes.

Calibrating weights for improved estimation in factor models for high-dimensional time series

Zihan Wang
Tsinghua University, China

This paper revisits the estimation of latent factor models for time series in high dimensions. A novel calibrating weight matrix has been proposed from a reduced-rank regression viewpoint to improve the estimation performance, especially when the factor strength is weak. An information criterion has been proposed to determine the tuning parameter. Besides, a spectral-guided estimation with tapering weights, which allows for serial correlations in idiosyncratic errors, has been introduced to utilize the information contained in temporal dependence. Our proposals can be widely applied in various complex factor models, including matrix-valued, functional and tensor factor models. Asymptotic theory and finite-sample performance of the estimated quantities have been provided.

Causal inference from multivariate time series: principles and problems

Michael Eichler
Maastricht University, Netherlands

This paper revisits the estimation of latent factor models for time series in high dimensions. A novel calibrating weight matrix has been proposed from a reduced-rank regression viewpoint to improve the estimation performance, especially when the factor strength is weak. An information criterion has been proposed to determine the tuning parameter. Besides, a spectral-guided estimation with tapering weights, which allows for serial correlations in idiosyncratic errors, has been introduced to utilize the information contained in temporal dependence. Our proposals can be widely applied in various complex factor models, including matrix-valued, functional and tensor factor models.Asymptotic theory and finite-sample performance of the estimated quantities have been provided.

Homogeneity pursuit in ranking inferences based on pairwise comparison data

Yuxin Tao
Tsinghua University, China

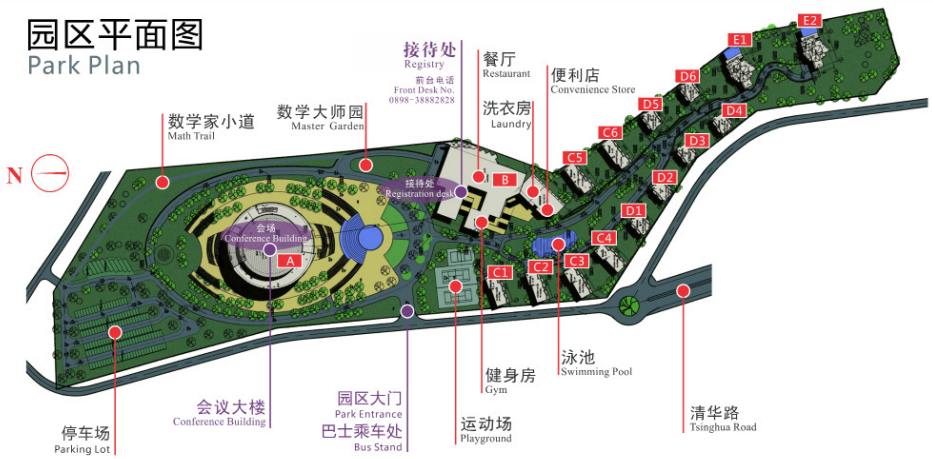
The Bradley-Terry-Luce (BTL) model is one of the most celebrated models for ranking inferences based on pairwise comparison data, which associates individuals with latent preference scores and produces ranks. An important question that arises is the uncertainty quantification for ranks. It is natural to think that ranks for two individuals are not trustworthy if there is only a subtle difference in their preference scores. In this paper, we explore the homogeneity of scores in the BTL model, which assumes that individuals cluster into groups with the same preference scores. We introduce the clustering algorithm in regression via data-driven segmentation (CARDS) penalty into the likelihood function, which can rigorously and automatically separate parameters and uncover group structure. Statistical properties of two versions of CARDS are analyzed. As a result, we achieve a faster convergence rate and sharper confidence intervals for the maximum likelihood estimation (MLE) of preference scores, providing insight into the power of exploring low-dimensional structure in a high-dimensional setting. We analyze real data examples, including NBA basketball ranking and journal ranking, to highlight the improved prediction performance and interpretation ability of our method.

Welcome to TSIMF

The facilities of TSIMF are built on a 23-acre land surrounded by pristine environment at Phoenix Hill of Phoenix Township. The total square footage of all the facilities is over 29,000 square meter that includes state-of-the-art conference facilities (over 10,000 square meter) to hold many international workshops simultaneously, two reading rooms of library, a guest house (over 10,000 square meter) and the associated catering facilities, a large swimming pool, gym and sports court and other recreational facilities.

Management Center of Tsinghua Sanya International Forum is responsible for the construction, operation, management and service of TSIMF. The mission of TSIMF is to become a base for scientific innovations, and for nurturing of innovative human resource; through the interaction between leading mathematicians and core research groups in pure mathematics, applied mathematics, statistics, theoretical physics, applied physics, theoretical biology and other relating disciplines, TSIMF will provide a platform for exploring new directions, developing new methods, nurturing mathematical talents, and working to raise the level of mathematical research in China.

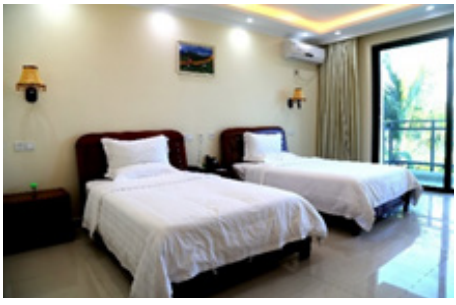
About Facilities



Registration

Conference booklets, room keys and name badges for all participants will be distributed at the front desk. Please take good care of your name badge. It is also your meal card and entrance ticket for all events.

Guest Room



All the rooms are equipped with: free Wi-Fi(Password:tsimf123), TV, air conditioning and other utilities.

Family rooms are also equipped with kitchen and refrigerator.

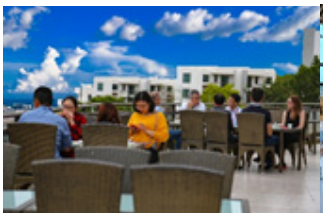
Library

TSIMF library is available during the conference and can be accessed by using your room card. There is no need to sign out books but we ask that you kindly return any borrowed books to the book cart in library before your departure.
Opening Hours: 09:00am-22:00pm

In order to give readers a better understanding of the contributions made by the Fields Medalists, the library of Tsinghua Sanya International Mathematics Forum (TSIMF) instituted the Special Collection of Fields Medalists as permanent collection of the library to serve the mathematical researchers and readers.
So far, there are 234 books from 47 authors in the Special Collection of Fields Medalists of TSIMF library. They are on display in room A220. The participants are welcome to visit.

Restaurant

All the meals are provided in the restaurant (Building B1) according to the time schedule.



Laundry

The self-service laundry room is located in the Building 1 (B1).

Opening Hours: 24 hours

Gym

The gym is located in the Building 1 (B1), opposite to the reception hall. The gym provides various fitness equipment, as well as pool tables, tennis tables etc.

Playground

Playground is located on the east of the central gate. There you can play basketball, tennis and badminton. Meanwhile, you can borrow table tennis, basketball, tennis balls and badminton at the reception desk.

Swimming Pool

Please note that there are no lifeguards. We will not be responsible for any accidents or injuries. In case of any injury or any other emergency, please call the reception hall at +86-898-38882828.

Free Shuttle Bus Service at TSIMF

We provide free shuttle bus for participants and you are always welcome to take our shuttle bus, all you need to do is wave your hands to stop the bus.

Destinations: Conference Building, Reception Room, Restaurant, Swimming Pool, Hotel etc.

Contact Information of Administration Staff

Location of Conference Affairs Office: Room 104, Building A
Tel: 0086-898-38263896

Conference Manager: Shouxi He 何守喜
Tel:0086-186-8980-2225
Email: hesx@tsimf.cn

Location of Accommodation Affairs Office: Room 200, Building B1
Tel:0086-898-38882828
Accommodation Manager: Ms. Li YE 叶莉
Tel: 0086-139-7679-8300
Email: yeli@tsimf.cn

Director Assistant of TSIMF
Kai CUI 崔凯
Tel/Wechat: 0086- 136-1120-7077
Email :cuikai@tsimf.cn

Director of TSIMF
Prof.Xuan GAO 高瑄
Tel: 0086-186-0893-0631
Email: gaoxuan@tsinghua.edu.cn