香港中文大學統計學系
Department of Statistics
THE CHINESE UNIVERSITY OF HONG KONG

# Symposium on Data Science and Risk Analytics 2023

CUHK 60th Anniversary Alumni Homecoming



Date: 9 December 2023 (Sat)
Time: 8:55am – 5:55pm
Venue: LT6, Lady Shaw Building, CUHK

Organizing Committee
Department of Statistics
CUHK

## Invited Speakers

FENG Xiangnan
Fudan University

GAO Lan
The University of Tennessee Knoxville

HU Jie
Xiamen University

KANG Kai
Sun Yat-sen University

LI Han
Shenzhen University

LIU Songhao
Southern University of Science and Technology

SHEN Guohao
The Hong Kong Polytechnic University

SHI Jiasheng
The Chinese University of Hong Kong, Shenzhen

SONG Fangda
The Chinese University of Hong Kong, Shenzhen

YIN Jie
HSBC Hong Kong

ZHANG Zhuosong
Southern University of Science and Technology

Symposium website and
Programmme Information
https://www.sta.cuhk.edu.hk/symposium/2023/

statdept@cuhk.edu.hk

(852) - 3943 7931

**Saturday, December 9, 2023 (Hong Kong Time)**

08:55 – 09:00 Xinyuan Song (The Chinese University of Hong Kong)

Opening Remarks

09:00 – 09:30 Session Chair: Junhui Wang (The Chinese University of Hong Kong)

Jiasheng Shi (The Chinese University of Hong Kong, Shenzhen)
Time-Since-Infection Model for Hospitalization and Incidence Data

09:30 – 10:00 Session Chair: Xinyuan Song (The Chinese University of Hong Kong)

Xiangnan Feng (Fudan University)
Variational Bayesian Analysis of Nonhomogeneous Hidden Markov Models with Long and Ultra-long Sequences

10:00 – 10:30 Kai Kang (Sun Yat-sen University)
Blockwise Mixed Membership Model for Discovering the Clinical Heterogeneity of Parkinson's Disease

10:30 – 11:00 *Coffee Break*

11:00 – 11:30 Session Chair: Yuanyuan Lin (The Chinese University of Hong Kong)

Guohao Shen (The Hong Kong Polytechnic University)
Estimation of Non-crossing Quantile Regression Process with Deep ReQU Neural Networks

11:30 – 12:00 Session Chair: Zhixiang Lin (The Chinese University of Hong Kong)

Fangda Song (The Chinese University of Hong Kong, Shenzhen)
Survival Mixed Membership Blockmodel

12:00 – 12:15 Photo Session

12:30 – 14:30 *Lunch (By Invitation Only)*

14:30 – 15:00 Session Chair: Xiaodan Fan (The Chinese University of Hong Kong)

Jie Hu (Xiamen University)
Inferring Cell Division Mode and Population Size Based on Temporal Cell Proportion Data

15:00 – 15:30 Han Li (Shenzhen University)
A Divided Mallows Model for Ranked Data Aggregation

15:30 – 15:50 *Coffee Break*

15:50 – 16:20 Session Chair: Xiao Fang (The Chinese University of Hong Kong)

Lan Gao (The University of Tennessee Knoxville)
ARK: Robust Knockoffs Inference with Coupling

16:20 – 16:50 Zhuosong Zhang (Southern University of Science and Technology)
Dense Multigraphon-Valued Stochastic Processes and Edge-Changing Dynamics in the Configuration Model

16:50 – 17:20 Songhao Liu (Southern University of Science and Technology)
High-dimensional Central Limit Theorems by Stein's Method in the Degenerate Case

17:20 – 17:50 Session Chair: Tony Sit (The Chinese University of Hong Kong)

Jie Yin (HSBC Hong Kong)
4/2 Rough and Smooth

17:50 – 17:55 Junhui Wang (The Chinese University of Hong Kong)

Closing Remarks

18:15 *Dinner (By Invitation Only)*

**FENG Xiangnan**
*Fudan University*

## Variational Bayesian Analysis of Nonhomogeneous Hidden Markov Models with Long and Ultra-long Sequences

Nonhomogeneous hidden Markov models (NHMMs) are useful in modeling sequential and autocorrelated data. Bayesian approaches, particularly Markov chain Monte Carlo (MCMC) methods, are principal statistical inference tools for NHMMs. However, MCMC sampling is computationally demanding, especially for long observation sequences. We develop a variational Bayes (VB) method for NHMMs, which utilizes a structured variational family of Gaussian distributions with factorized covariance matrices to approximate target posteriors, combining a forward-backward algorithm and stochastic gradient ascent in estimation. To improve efficiency and handle ultra-long sequences, we further propose a subsequence VB (SVB) method that works on subsamples. The SVB method exploits the memory decay property of NHMMs and uses buffers to control for bias caused by breaking sequential dependence from subsampling. We highlight that the local nonhomogeneity of NHMMs substantially affects the required buffer lengths and propose the use of local Lyapunov exponents that characterize local memory decay rates of NHMMs and adaptively determine buffer lengths. Our methods are validated in simulation studies and in modeling ultra-long sequences of customers' telecom records to uncover the relationship between their mobile Internet usage behaviors and conventional telecommunication behaviors.

**GAO Lan**
*The University of Tennessee Knoxville*

## ARK: Robust Knockoffs Inference with Coupling

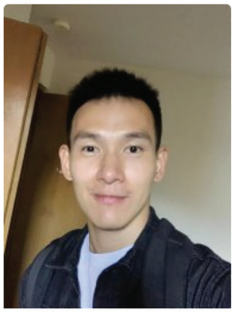
We investigate the robustness of the model-X knockoffs framework with respect to the misspecified or estimated feature distribution. We achieve such a goal by theoretically studying the feature selection performance of a practically implemented knockoffs algorithm, which we name as the approximate knockoffs (ARK) procedure, under the measures of the false discovery rate (FDR) and family wise error rate (FWER). The approximate knockoffs procedure differs from the model-X knockoffs procedure only in that the former uses the misspecified or estimated feature distribution. A key technique in our theoretical analyses is to couple the approximate knockoffs procedure with the model-X knockoffs procedure so that random variables in these two procedures can be close in realizations. We prove that if such coupled model-X knockoffs procedure exists, the approximate knockoffs procedure can achieve the asymptotic FDR or FWER control at the target level. We showcase three specific constructions of such coupled model-X knockoff variables, verifying their existence and justifying the robustness of the model-X knockoffs framework.

**HU Jie**
*Xiamen University*

## Inferring Cell Division Mode and Population Size Based on Temporal Cell Proportion Data

Analyzing the dynamic shifts in cell type proportions over time provides crucial insights into the mechanisms of cell division and growth. In this presentation, we will concentrate on a dynamic model of phenotypic plasticity and delve into the methods for detecting cell division modes and estimating cell population size by utilizing temporal data from Fluorescence-Activated Cell Sorting. By employing a Bayesian statistical framework, both symmetric and asymmetric division probabilities are deduced, demonstrating our approach's applicability in uncovering cancer cell de-differentiation, using SW620 colon cancer cell lines as a case study. Furthermore, two distinct techniques for estimating cell population size based on temporal proportion data are also proposed.

This is a joint work with Yihong Gu.



**KANG Kai**
*Sun Yat-sen University*

## Blockwise Mixed Membership Model for Discovering the Clinical Heterogeneity of Parkinson's Disease

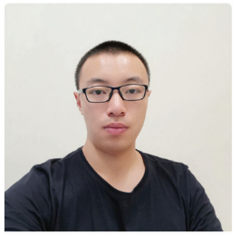
Current diagnostic landscape for Parkinson's disease (PD) faces formidable challenges due to the heterogeneous nature of disease course, including that (i) the disease progression varies hugely between patients, (ii) various types of motor and nonmotor symptoms exist, and (iii) the time to develop those clinical symptoms differs significantly. To tackle these complexities, we propose a novel blockwise mixed membership model (BM$^3$) to systematically unveil between-patient, between-symptom, and between-time clinical heterogeneity within PD. The key idea behind BM$^3$, which is fundamentally different from conventional mixed membership models, is to partition multivariate longitudinal observed variables into distinct blocks, enabling variables within each block to share a common latent membership while allowing different latent memberships across blocks. Consequently, the heterogeneous PD-related symptoms across time are divided into clinically homogeneous blocks consisting of correlated symptoms and time-dependent visits. Moreover, BM$^3$ assign each patient a subject-specific vector characterizing partial membership across latent clusters, thus enabling the intricate description of individualized disease progression. We provide both theoretical and empirical justification for the identifiability and posterior consistency of the unknown blocking structures and model parameters. By applying BM$^3$ to the Parkinson's Progression Markers Initiative (PPMI) data, we advance our comprehension of PD heterogeneity, paving the way for the development of more precise and targeted therapies to benefit patients.

**LI Han**
*Shenzhen University*

## A Divided Mallows Model for Ranked Data Aggregation

In our work, we study the rank aggregation problem, which aims to find a consensus ranking by aggregating multiple ranking lists. To tackle the problem probabilistically, we formulate an elaborate ranking model by generalizing the traditional Mallows model. The original model assumes an uniform pair preference structure, which imposes a strict condition on the data. We attempt to relax this condition, and propose a new model that allows the pair preference varies structurally. Our model is quite flexible and has a closed form expression for complete rankings as well as top-k rankings. We investigate several useful theoretical properties of the model and propose efficient algorithms to infer the model structure and parameters. Through extensive simulation studies and real applications, the new model is demonstrated to have satisfactory performance in different scenarios.

This is a joint work with Haijin HE and Xiaodan FAN.



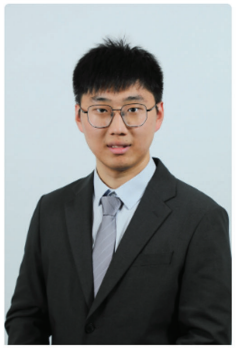
**LIU Songhao**
*Southern University of Science and Technology*

## High-dimensional Central Limit Theorems by Stein's Method in the Degenerate Case

In the literature of high-dimensional central limit theorems, there is a gap between results for general limiting correlation matrix $\Sigma$ and the strongly non-degenerate case.

For the general case where $\Sigma$ may be degenerate, under certain light-tail conditions, when approximating a normalized sum of $n$ independent random vectors by the Gaussian distribution $N(0,\Sigma)$ in multivariate Kolmogorov distance, the best-known error rate has been $O(n^{-1/4})$, subject to logarithmic factors of the dimension. For the strongly non-degenerate case, that is, when the minimum eigenvalue of $\Sigma$ is bounded away from 0, the error rate can be improved to $O(n^{-1/2})$ up to a $\log n$ factor. In this paper, we show that the $O(n^{-1/2})$ rate up to a $\log n$ factor can still be achieved in the degenerate case, provided that the minimum eigenvalue of the limiting correlation matrix of any three components is bounded away from 0.

We prove our main results using Stein's method in conjunction with previously unexplored inequalities for the integral of the first three derivatives of the standard Gaussian density over convex polytopes. These inequalities were previously known only for hyperrectangles. Our proof demonstrates the connection between the three-components condition and the third moment Berry--Esseen bound.



**SHEN Guohao**
*The Hong Kong Polytechnic University*

## Estimation of Non-crossing Quantile Regression Process with Deep ReQU Neural Networks

We propose a penalized nonparametric approach to estimating the quantile regression process (QRP) in a non-separable model using rectifier quadratic unit (ReQU) activated deep neural networks and introduce a novel penalty function to encourage non-crossing of quantile regression curves. We establish the non-asymptotic excess risk bounds for the estimated QRP and derive the mean integrated squared error for the estimated QRP under mild smoothness and regularity conditions. To establish these non-asymptotic risk and estimation error bounds, we also develop a new error bound for approximating C^s smooth functions with s>0 and their derivatives using ReQU activated neural networks. This is a new approximation result for ReQU networks and is of independent interest and may be useful in other problems. Our numerical experiments demonstrate that the proposed method is competitive with or outperforms existing methods, including those using reproducing kernels and random forests, for nonparametric quantile regression.

**SHI Jiasheng**
*The Chinese University of Hong Kong, Shenzhen*

## Time-Since-Infection Model for Hospitalization and Incidence Data

The Time since Infection (TSI) models have gained widespread popularity and acclaim for their performance during the COVID-19 pandemic, establishing them as an increasingly popular choice for modeling infectious disease transmission due to their practicality, flexibility and ability to address complex disease control questions. However, a notable limitation of TSI models is their primary reliance on incidence data. In existing TSI models, even when hospitalization data are available, they have not been designed to estimate disease transmission or to predict disease-related hospitalizations — metrics crucial for understanding the trajectory of a pandemic and for hospital resource planning. Furthermore, their dependence on reported infection data makes them vulnerable to variations in data quality. In this study, we advance TSI models by integrating hospitalization data, a critical component for a comprehensive understanding of infectious diseases. This integration marks a significant step forward in infectious disease modeling using TSI models. Our improvement enables hospitalization nowcasting, reduces bias in incidence data, and connects TSI models with other infectious disease models. We introduce hospitalization propensity parameters to model incidence and hospitalization counts jointly. We use a composite likelihood function to accommodate complex data structures and an MCEM algorithm to effectively estimate model parameters. We apply our method to COVID-19 data and estimate disease transmission dynamics, assess risk factor impacts, and calculate hospitalization propensities. Our novel TSI model offers a fresh perspective on using hospitalization data to enhance the understanding of disease dynamics and support public health efforts.

**SONG Fangda**
*The Chinese University of Hong Kong, Shenzhen*

## Survival Mixed Membership Blockmodel

Whenever we send a message via a channel such as E-mail, Facebook, WhatsApp, WeChat, or LinkedIn, we care about the response rate—the probability that our message will receive a response—and the response time—how long it will take to receive a reply. Recent studies have made considerable efforts to model the sending behaviors of messages in social networks with point processes. However, statistical research on modeling response rates and response times on social networks is still lacking. Compared with sending behaviors, which are often determined by the sender's characteristics, response rates and response times further depend on the relationship between the sender and the receiver. Here, we develop a survival mixed membership blockmodel (SMMB) that integrates semiparametric cure rate models with a mixed membership stochastic blockmodel to analyze time-to-event data observed for node pairs in a social network, and we are able to prove its model identifiability without the pure node assumption. We develop a Markov chain Monte Carlo algorithm to conduct posterior inference and select the number of social clusters in the network according to the conditional deviance information criterion. The application of the SMMB to the Enron E-mail corpus offers novel insights into the company's organization and power relations. Supplementary materials for this article are available online.

**YIN Jie**
*HSBC Hong Kong*

## 4/2 Rough and Smooth

Conflicting opinions on rough volatility motivate us to propose a convex combination of the rough Heston (rough 1/2) and smooth 3/2 models to create a novel 4/2 rough and smooth (4/2RS) volatility model. This parsimonious two-factor model captures many stylized facts from empirical studies and flexibly provides realistic variance distributions and rich autocorrelation structures. For instance, it generates an elasticity of variance (EV) of the variance process that is consistent with empirical estimates in the literature and captures the volatility roughness of short-term options. Moreover, the proposed model allows an analytical formula for the characteristic function of option pricing. Even with a very small weight on the rough 1/2 component of the model, the empirical analysis of short-term option data still identifies a roughness level similar to that identified by the rough Heston model. That is, the 4/2RS model features separate calibration of roughness and EV.

This is a joint work with Tingjin Yan, Ling Wang, and Hoi Ying Wong.

**ZHANG Zhuosong**
*Southern University of Science and Technology*

# Dense Multigraphon-Valued Stochastic Processes and Edge-Changing Dynamics in the Configuration Model

Time-evolving random graph models have appeared and have been studied in various fields of research over the past decades. However, the rigorous mathematical treatment of large graphs and their limits at the process-level is still in its infancy. In this article, we adapt the approach of Athreya, den Hollander and Röllin (2021) to the setting of multigraphs and multigraphons, introduced by Kolossváry and Ráth (2011). We then generalise the work of Ráth (2012) and Ráth and Szakács (2012), who analysed edge-flipping dynamics on the configuration model — in contrast to their work, we establish weak convergence at the process-level, and by allowing removal and addition of edges, these limits are non-deterministic.

This is a joint work with Adrian Röllin.

2023


2023

*Organizing Committee:*

Department of Statistics, The Chinese University of Hong Kong

*Email:*

statistics_sym@sta.cuhk.edu.hk

*Phone:*

852-3943 7932 (Ms. Wendy TANG)
852-3943 7952 (Ms. Yanny NG)

*Correspondence:*

Rm 119, Lady Shaw Building
Department of Statistics
The Chinese University of Hong Kong
Shatin, N.T.
Hong Kong