



香港中文大學統計與數據科學系

Department of
Statistics and Data Science
THE CHINESE UNIVERSITY OF HONG KONG

Symposium on Statistics and Risk Management 2025

cum Celebration of the Renaming of the Department of Statistics
and Data Science at The Chinese University of Hong Kong



20-21
December 2025 (Sat – Sun)

LT1B, Cheng Yu Tung Building,
The Chinese University of Hong Kong

Distinguished Speakers:

20 December 2025 AM –

Professor Jun LIU

Tsinghua University / Harvard University

Professor Heping ZHANG

Yale University

Professor Fang YAO

Peking University

20 December 2025 PM –

Professor Todd OGDEN

Columbia University

Professor Yuxin CHEN

The Wharton School,
The University of Pennsylvania

21 December 2025 AM –

Professor Xin GUO

University of California, Berkeley

Professor Henry LAM

Columbia University

21 December 2025 PM –

Professor Xuegong ZHANG

Tsinghua University

Professor Qi LONG

University of Pennsylvania

Organiser:

Department of Statistics and Data Science,
The Chinese University of Hong Kong

Registration Deadline:

5 December 2025



Enquiries: statistics_sym@sta.cuhk.edu.hk



<https://www.sta.cuhk.edu.hk/symposium/2025/>



Saturday, December 20, 2025 (Hong Kong Time)

08:00 – 08:30 Registration + Photo Taking at the Backdrop

08:30 – 08:55 Opening Ceremony

08:55 – 09:00 Group Photo

Session Chair: Professor Xiaodan Fan (The Chinese University of Hong Kong)

09:00 – 09:45 Keynote Lecture

Professor Jun Liu (*Tsinghua University / Harvard University*)
AI Deployment and Statistical Thinking

Session Chair: Professor Tony Sit (The Chinese University of Hong Kong)

09:45 – 10:45 Industrial Talks

Mr. Chi Ho Chan (*Census and Statistics Department, HKSAR*)
Smart Statistics: C&SD's Digital Transformation in the AI Era

Dr. Pengcheng Chen (*Sina Corporation*)
Statistical Decision in Financial Risk Management

Mr. Ted Yu (*The Hongkong and Shanghai Banking Corporation Limited*)
Leveraging AI and Data Science for Risk Management in Banking

10:45 – 11:05 *Coffee Break*

Session Chair: Professor Huichen Zhu (The Chinese University of Hong Kong)

11:05 – 11:50 Keynote Lecture

Professor Heping Zhang (*Yale University*)
Unadorned Statistics in the Light of AI

11:50 – 12:35 Keynote Lecture

Professor Fang Yao (*Peking University*)
Deep Semiparametric Partial Differential Equation Models

12:35 – 14:30 *Lunch (By Invitation)*

Session Chair: Professor Gen Li (The Chinese University of Hong Kong)

14:30 – 15:15 Keynote Lecture

Professor Todd Ogden (*Columbia University*)
Functional Metric Learning for Classification Problems

15:15 – 16:00 Keynote Lecture

Professor Yuxin Chen (*The Wharton School, The University of Pennsylvania*)
Transformers Meet In-Context Learning: A Universal Approximation Theory

16:00 – 16:20 *Coffee Break*

Session Chair: Professor Gen Li (The Chinese University of Hong Kong)

16:20 – 16:50 **Professor Jie Ding** (*University of Minnesota*)
Designing Intelligent AI: Insights from Human Cognition

16:50 – 17:20 **Professor Jingyuan Liu** (*Xiamen University*)
LLM-Powered Deep Panel Modeling

17:20 – 17:50 **Professor Gen Li** (*The Chinese University of Hong Kong*)
Faster Convergence and Acceleration for Diffusion-Based Generative Models

18:30 *Banquet (By invitation)*

Symposium on Statistics and Risk Management 2025

Sunday, December 21, 2025 (Hong Kong Time)

Session Chair: Professor Jiacheng Zhang (The Chinese University of Hong Kong)

09:00 – 09:45 Keynote Lecture

Professor Xin Guo (*University of California, Berkeley*)
From LLM to RL and Diffusion Models, via (rough) Differential Equations

09:45 – 10:30 Keynote Lecture

Professor Henry Lam (*Columbia University*)
Bootstrap with Very Few Resamples: An Integrative View on Data and Monte Carlo Uncertainties

10:30 – 10:50 *Coffee Break*

Session Chair: Professor Jiacheng Zhang (The Chinese University of Hong Kong)

10:50 – 11:20 **Professor Chi Seng Pun** (*Nanyang Technological University*)
Backward Stochastic Volterra Integral Equations with General Nonlinearities

11:20 – 11:50 **Professor Mathieu Lauriere** (*New York University Shanghai*)
An Efficient On-Policy Deep Learning Framework for Stochastic Optimal Control

11:50 – 12:20 **Professor Jiacheng Zhang** (*The Chinese University of Hong Kong*)
Continuous-time Mean Field Games: A Primal-dual Characterization

12:20 – 14:30 *Lunch (By Invitation)*

Session Chair: Professor Yingxin Lin (The Chinese University of Hong Kong)

14:30 – 15:15 Keynote Lecture

Professor Xuegong Zhang (*Tsinghua University*)
Large AI Cellular Models: Promises and Challenges

15:15 – 16:00 Keynote Lecture

Professor Qi Long (*University of Pennsylvania*)
Responsible Statistical and AI/ML Methods for Harnessing the Power of Electronic Health Records

16:00 – 16:20 *Coffee Break*

Session Chair: Professor Yingxin Lin (The Chinese University of Hong Kong)

16:20 – 16:50 **Professor Linbo Wang** (*University of Toronto*)
Fighting Noise with Noise: Causal Inference with Many Candidate Instruments

16:50 – 17:20 **Professor Qiongshi Lu** (*University of Wisconsin–Madison*)
The Blurred Line Between Genes and Environments: Insights from GWAS of Family Members' Phenotypes

17:20 – 17:50 **Professor Yingxin Lin** (*The Chinese University of Hong Kong*)
Controlling False Discoveries after Clustering via Data Splitting for Spatial Domain Marker Detection

17:50 – 17:55 Closing Remarks
Professor Xinyuan Song (The Chinese University of Hong Kong)



Chi-ho CHAN
*Census and
Statistics
Department,
HKSAR*

Smart Statistics: C&SD's Digital Transformation in the AI Era

The AI era demands a fundamental evolution in the production of official statistics. This presentation shares the Census and Statistics Department's (C&SD) digital transformation journey since 2018, which is redefining our operational capabilities.

We will introduce our key data science strategies, designed to foster innovation at scale while maintaining governance. The talk will highlight tangible AI use cases, including Deep Learning models for trade anomaly detection and commodity coding, and Large Language Models (LLMs) for analysing large volumes of financial documents.

Underpinning every innovation is an unwavering commitment to quality. We will outline the Quality Assurance Mechanisms governing our AI projects, ensuring they meet the high standards of accuracy and reliability essential for official statistics.

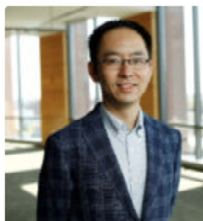


Pengcheng CHEN
*Sina Digital
Technology Co.*

Statistical Decision in Financial Risk Management

With the development of big data technology and statistics, their applications in various fields have become increasingly widespread. Financial risk management, especially in the credit risk management, represents a forefront of statistics and big data applications. Compared to the other industries, the data of this field is higher integration, and the benefits of technological advancements are more easily measured. Financial risk management has evolved from initially using basic models such as logistic regression to progressively applying multiple techniques such as Transfer-Learning, Pareto-Optimization, and Dynamic Decision-making algorithm to manage risk throughout the entire credit lifecycle of users.

Statistical decision in financial risk management refers to the process of using statistical models and decision science, based on historical data and business scenario characteristics, to make quantitative judgments and optimal choices regarding uncertainties in key areas such as credit approval, in-process monitoring, risk pricing, and fraud detection. The key endpoint is to find the balance between "risk" and "returns". In this report, we will share some real problems encountered in practice and corresponding solutions.



Yuxin CHEN
*The Wharton
School, The
University of
Pennsylvania*

Transformers Meet In-Context Learning: A Universal Approximation Theory

Large language models are capable of in-context learning, the ability to perform new tasks at test time using a handful of input-output examples, without parameter updates. We develop a universal approximation theory to elucidate how transformers enable in-context learning. For a general class of functions (each representing a distinct task), we demonstrate how to construct a transformer that, without any further weight updates, can predict based on a few noisy in-context examples with vanishingly small risk. Unlike prior work that frames transformers as approximators of optimization algorithms (e.g., gradient descent) for statistical learning tasks, we integrate Barron's universal function approximation theory with the algorithm approximator viewpoint. Our approach yields approximation guarantees that are not constrained by the effectiveness of the optimization algorithms being mimicked, extending far beyond convex problems like linear regression. The key is to show that (i) any target function can be nearly linearly represented, with small ℓ_1 -norm, over a set of universal features, and (ii) a transformer can be constructed to find the linear representation -- akin to solving Lasso -- at test time.

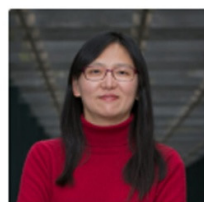
This is joint work with Gen Li, Yuchen Jiao, Yu Huang, and Yuting Wei.



Jie DING
*University of
Minnesota*

Designing Intelligent AI: Insights from Human Cognition

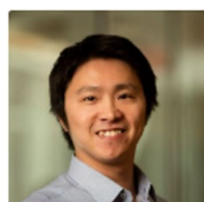
Modern AI systems are evolving from passive tools to autonomous agents capable of reasoning, learning, and collaboration. This talk explores emerging research directions in generative AI and foundational principles inspired by human cognition: continuous learning and adaptation, effective knowledge transfer, and multi-objective decision making. The discussion aims to stimulate thoughts on developing domain-specific AI that can operate reliably in complex, real-world environments.



Xin GUO
*University of
California,
Berkeley*

From LLM to RL and Diffusion Models, via (rough) Differential Equations

Transfer learning is a machine learning technique that leverages knowledge acquired in one domain to enhance performance on a related task. It plays a central role in the success of large language models (LLMs) such as GPT and BERT, which leverage pretraining to enable broad generalization across downstream applications. In this talk, I will discuss how reinforcement learning (RL), and in particular continuous time RL, can benefit from transfer learning principles. I will present convergence results formulated through stability analysis for stochastic control systems, using rough differential equation techniques. Finally, I will show how this analysis yields a natural corollary establishing robustness guarantees for a class of score-based generative diffusion models.



Henry LAM
*Columbia
University*

Bootstrap with Very Few Resamples: An Integrative View on Data and Monte Carlo Uncertainties

While the bootstrap is demonstrably powerful in quantifying statistical uncertainty, its implementation could incur substantial computation costs in expensive models due to repeated resampling. We present a bootstrap implementation that can use a number of resamples as low as one, while maintaining valid coverage guarantees. Our approach is empowered by an integrative view on the statistical noise from data and the Monte Carlo noise in the resampling execution, in contrast to their separate handling in the conventional bootstrap framework. We leverage our idea to efficiently quantify uncertainties in several problems, including neural network training, stochastic programming and gradient descent, and problems in simulation modeling that comprises both epistemic and aleatory uncertainties.



**Mathieu
LAURIERE**
*New York
University
Shanghai*

An Efficient On-Policy Deep Learning Framework for Stochastic Optimal Control

We present a novel on-policy algorithm for solving stochastic optimal control (SOC) problems. By leveraging the Girsanov theorem, our method directly computes on-policy gradients of the SOC objective without expensive backpropagation through stochastic differential equations or adjoint problem solutions. This approach significantly accelerates the optimization of neural network control policies while scaling efficiently to high-dimensional problems and long time horizons. We evaluate our method on classical SOC benchmarks as well as applications to sampling from unnormalized distributions via Schrödinger-Föllmer processes and fine-tuning pre-trained diffusion models. Experimental results demonstrate substantial improvements in both computational speed and memory efficiency compared to existing approaches. Joint work with Mengjian Hua and Eric Vanden-Eijnden.

Link to the paper: <https://openreview.net/forum?id=sv5PiLZbUr>



Gen LI
*The Chinese
University of
Hong Kong*

Faster Convergence and Acceleration for Diffusion-Based Generative Models

Diffusion models, which generate new data instances by learning to reverse a Markov diffusion process from noise, have become a cornerstone in contemporary generative modeling. While their practical power has now been widely recognized, the theoretical underpinnings remain underdeveloped. Particularly, despite the recent surge of interest in accelerating sampling speed, convergence theory for these acceleration techniques remains limited. In this talk, I will first introduce an acceleration sampling scheme for stochastic samplers that provably improves the iteration complexity under minimal assumptions. The second part focuses on diffusion-based language models, whose ability to generate tokens in parallel significantly accelerates sampling relative to traditional autoregressive methods. Adopting an information-theoretic lens, we establish a sharp convergence theory for diffusion language models, thereby providing the first rigorous justification of both their efficiency and fundamental limits.



Yingxin LIN
*The Chinese
University of
Hong Kong*

Controlling False Discoveries after Clustering via Data Splitting for Spatial Domain Marker Detection

Spatial omics technologies are transforming biomedical research by enabling genome-wide measurement of molecular activity while preserving the spatial context within tissues. These advances create unprecedented opportunities to uncover cell-cell interactions, tissue organization, and disease mechanisms. A crucial step in realizing this potential is identifying spatial domain markers, which are essential for defining tissue architecture and understanding disease progression. However, using the same data for both clustering and marker detection creates the problem of "double dipping", which can lead to inflated false discoveries, particularly when domain boundaries are poorly defined. To address this challenge, we develop SpaDS and SpaMDS, data splitting based approaches for differential expression testing after clustering for spatial omics, enabling robust spatial domain marker discovery with controlled false discovery rate and high power. Through extensive simulations and analyses of spatial omics datasets from diverse technologies, we demonstrate that the data-splitting methods are easy to implement, adaptable to existing spatial omics data analysis pipelines, and often outperform other approaches under weak signals and high feature correlations.



Jingyuan LIU
Xiamen
University

LLM-Powered Deep Panel Modeling

Panel modeling for economic dynamics is crucial for timely and effective policymaking. However, it typically relies only on low-frequency, high-cost surveys and macroeconomic variables, thus often fails to capture rapid market fluctuations and leads to inaccurate predictions. In this paper, we propose a new framework that integrates large language model (LLM) analyses and social media narratives to enhance the prediction power of dynamic panel modeling. Through narrative corpus constructed from social media data, we introduce a prompt-based GPT model and a series of fine-tuned BERT models to generate high-frequency LLM-induced surrogates for the economic indices of interest. A novel joint modeling strategy is then advocated to transfer the information from these surrogates to enhance the prediction power for the targeted economic indices. To solve the joint objectives, we further develop a new deep panel learning procedure with region-wise homogeneity pursuit, which has its own significance in panel data analysis literature. In addition, conformal-based panel prediction intervals are provided to quantify the uncertainty of the LLM-powered prediction. Empirical and theoretical analyses demonstrate that our approach significantly reduces short-term forecasting errors and more effectively captures abrupt inflationary shifts compared to traditional econometric models.



Jun LIU
Tsinghua
University /
Harvard
University

AI Deployment and Statistical Thinking

Driven by big data, technology developments, and statistical ideas, artificial intelligence (AI) has witnessed remarkable development in recent years. However, how to effectively implement and deepen AI across various application domains remains a major challenge for the next decade. Statistics focuses on designing experiments, collecting data, and extracting insights from noisy data to guide decision-making and make predictions while quantifying associated uncertainties; at its core, it employs probability theory as a foundation to build generative and predictive models, enhancing our understanding of data and cognition. These methodological and philosophical perspectives of statistical thinking serve as one of the key driving forces behind the current AI revolution and represent crucial means to further root AI technologies across industries. Fundamental statistical concepts and methods—such as randomization, cross-validation, shrinkage estimation, regularization, bias-variance tradeoff, causal inference, and the Bayesian theory—remain essential guides for handling noisy data, advancing machine learning and AI, and are critical in assessing uncertainty and building interpretable models. We illustrate the power of integrating statistical thinking with AI through several examples.



Qi LONG
University of
Pennsylvania

Responsible Statistical and AI/ML Methods for Harnessing the Power of Electronic Health Records

Rich electronic health records (EHR) data offer remarkable opportunities in advancing precision medicine, but they also present daunting analytical challenges. Multi-modal data in EHR that are recorded at irregular time intervals with varying frequencies include structured data such as labs and vitals, codified data such as diagnosis and procedure codes, and unstructured data such as clinical notes and pathology reports. They are typically incomplete and fraught with other errors and biases. Such data issues, if not adequately addressed, would lead to biased results (Getzen et al. 2023, JBI). In this talk, I will share my lab's work on developing statistical and ML models including agentic AI for harnessing EHR data for intelligent precision medicine (Orcutt et al., 2025, Nature Medicine) and responsible statistical and large language models (LLMs) based methods for addressing the aforementioned challenges associated with EHR data (Zhang et al., 2025, Commun. Med.; Consoli et al., 2025, JAMIA). Since LLMs are themselves plagued by various risks and biases, I will also briefly discuss our research on developing rigorous methods for mitigating pitfalls and risks of LLMs (Xiao et al. 2025a JASA and 2025b, ICML; Li et al. 2025a AoS and 2025b JRSSB).

Keywords: Agentic AI; AI/ML; Electronic Health Records; Large Language Models; Precision Medicine.



Qiongshi LU
University of
Wisconsin-
Madison

The Blurred Line Between Genes and Environments: Insights from GWAS of Family Members' Phenotypes

Genome-wide association study (GWAS) methodologies have become quite standard for complex trait genetic research. Today, a modern GWAS typically correlates a phenotype with tens of millions of genetic variants in large cohorts of millions of individuals to reveal genotype-phenotype associations. However, this seemingly standard approach can give largely biased and/or confounded results in various applications. In this talk, I will discuss a new study design which associates genetic data of a cohort with their family members' phenotypes. That is, the genotypic and phenotypic variables in the GWAS are collected from different individuals. Through three separate applications, focusing on offspring, parental, and spousal phenotypes, I will discuss several challenges and new insights in genetic nurture, ascertainment bias, and assortative mating. The phenotypes discussed in this talk will include socioeconomic outcomes, neurodegenerative disease risk, as well as human partner choice.



Todd OGDEN
Columbia
University

Functional Metric Learning for Classification Problems

Metric-based classification methods, such as K-nearest neighbors, are widely used, and their success often depends on the choice of distance measure. While metric learning has been shown to improve classification in multivariate settings, it faces challenges in high-dimensional spaces and when data are noisy. In many modern applications, including EEG signals and brain images, data are more naturally represented as functional data, which exhibit inherent smoothness, continuity, and structural dependence. Motivated by this, we propose a functional metric learning framework for classification and explore its performance through simulation studies and an application to EEG data.

This is joint work with Shanghong XIE, University of South Carolina.



Chi Seng PUN
Nanyang
Technological
University

Backward Stochastic Volterra Integral Equations with General Nonlinearities

We establish existence, uniqueness, and regularity results for multi-dimensional backward stochastic Volterra integral equations (BSVIEs) with generators that may be random and exhibit nonlinear dependence on the solution, its martingale integrand, and their diagonal processes. Our approach leverages Malliavin calculus, offering a novel framework to address the analytical challenges posed by diagonal terms, in contrast to traditional techniques. We further derive a probabilistic representation for classical solutions to corresponding semi-linear partial differential equations via adapted solutions of BSVIEs. As an application, we study the dynamically optimal mean-variance portfolios under stochastic market models, where the myopic and intertemporal hedging demands are characterized by the diagonal components of some BSVIE solutions.



Linbo WANG
*University of
Toronto*

Fighting Noise with Noise: Causal Inference with Many Candidate Instruments

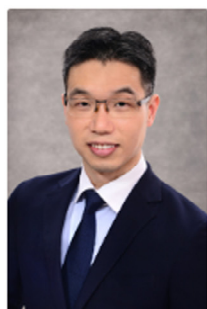
Instrumental variable methods provide useful tools for inferring causal effects in the presence of unmeasured confounding. To apply these methods with large-scale data sets, a major challenge is to find valid instruments from a possibly large candidate set. In practice, most of the candidate instruments are often not relevant for studying a particular exposure of interest. Moreover, not all relevant candidate instruments are valid as they may directly influence the outcome of interest. In this article, we propose a data-driven method for causal inference with many candidate instruments that addresses these two challenges simultaneously. A key component of our proposal involves using pseudo variables, known to be irrelevant, to remove variables from the original set that exhibit spurious correlations with the exposure. Synthetic data analyses show that the proposed method performs favourably compared to existing methods. We apply our method to a Mendelian randomization study estimating the effect of obesity on health-related quality of life.



Fang YAO
Peking University

Deep Semiparametric Partial Differential Equation Models

In many scientific fields, the generation and evolution of data are governed by partial differential equations (PDEs) which are typically informed by established physical laws at the macroscopic level to describe general and predictable dynamics. However, some complex influences may not be fully captured by these laws at the microscopic level due to limited scientific understanding. This work proposes a unified framework to model, estimate, and infer the mechanisms underlying data dynamics. We introduce a general semiparametric PDE (SemiPDE) model that combines interpretable mechanisms based on physical laws with flexible data-driven components to account for unknown effects. The physical mechanisms enhance the SemiPDE model's stability and interpretability, while the data-driven components improve adaptivity to complex real-world scenarios. A deep profiling M-estimation approach is proposed to decouple the solutions of PDEs in the estimation procedure, leveraging both the accuracy of numerical methods for solving PDEs and the expressive power of neural networks. For the first time, we establish a semiparametric inference method and theory for deep M-estimation, considering both training dynamics and complex PDE models. We analyze how the PDE structure affects the convergence rate of the nonparametric estimator, and consequently, the parametric efficiency and inference procedure enable the identification of interpretable mechanisms governing data dynamics. Simulated and real-world examples demonstrate the effectiveness of the proposed methodology and support the theoretical findings.



Ted YU
*The Hongkong
and Shanghai
Banking
Corporation
Limited*

Leveraging AI and Data Science for Risk Management in Banking

This talk explores the application of AI and data science for real-time counterparty credit risk management in banking, highlighting how it overcomes the limitations of traditional add-on methods and improves the accuracy without heavy computation. The presentation also discusses other AI use cases in banking.

The presentation is a joint work prepared in collaboration with Jane Yin and Anson Lam.



Heping ZHANG
Yale University

Unadorned Statistics in the Light of AI

Regression, clustering, and sequential analysis are fundamental techniques in statistics. Today, these same concepts are often relabeled as supervised learning, unsupervised learning, deep learning, reinforcement learning, or, more broadly, artificial intelligence. In this talk, I will present several of our statistical methods, developed in response to real-world applications, including the analysis of high-dimensional data for building-related occupant syndromes, inference of risk factors with uncertain frequencies from haplotype data, and residual diagnostics for generalized linear models. By revisiting these examples, I will highlight the essential ideas and techniques that our approaches share with modern AI methods. My goal is to reflect on why our statistical methods appear so “unadorned,” and to ask whether—and how—we might close the gap in how statistics and AI are recognized and valued.



Jiacheng ZHANG
The Chinese University of Hong Kong

Continuous-time Mean Field Games: A Primal-dual Characterization

This paper establishes a primal-dual formulation for continuous-time mean field games (MFGs) and provides a complete analytical characterization of the set of all Nash equilibria (NEs). We first show that for any given mean field flow, the representative player's control problem with *measurable coefficients* is equivalent to a linear program over the space of occupation measures. We then establish the dual formulation of this linear program as a maximization problem over smooth subsolutions of the associated Hamilton-Jacobi-Bellman (HJB) equation, which plays a fundamental role in characterizing NEs of MFGs. Finally, a complete characterization of all NEs for MFGs is established by the strong duality between the linear program and its dual problem. This strong duality is obtained by studying the solvability of the dual problem, and in particular through analyzing the regularity of the associated HJB equation.

Compared with existing approaches for MFGs, the primal-dual formulation and its NE characterization do not require the convexity of the associated Hamiltonian or the uniqueness of its optimizer, and remain applicable when the HJB equation lacks classical or even continuous solutions.



Xuegong ZHANG
Tsinghua University

Large AI Cellular Models: Promises and Challenges

For decades data analysis and modeling tasks in functional genomics have been featured by the major challenge of small samples in high-dimensional space. The development of high-throughput single-cell sequencing technologies have brought the field into real big data era: Suddenly, we can have hundreds of millions of cells of thousands or higher dimensionality. Inspired by the Transformer-based large language models, large cellular models showed great promise as AI virtual cells that can predict cellular responses to various types of perturbations. This talk will share our practices along this line, and discuss challenges in experiment design and benchmarking of such models from the view of statistics and data sciences.

Memo

Contact Us

Organizing Committee:

Department of Statistics and Data Science, The Chinese University of Hong Kong

Email:

statistics_sym@sta.cuhk.edu.hk

Phone:

852-3943 7952 (Ms. Yanny NG)

Correspondence:

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

Symposium Venue

LT 1B, Cheng Yu Tung Building, The Chinese University of Hong Kong

