HKUST 2025 Workshop on Statistics and Data Science

Date: August 8, 2025

Organized by: Department of Mathematics, HKUST

All the committee members are listed in the alphabetical order.

Organizers:

Xinzhou Guo, Department of Mathematics, HKUST

Shiqing Ling, Department of Mathematics, HKUST

Rong Tang, Department of Mathematics, HKUST

Dong Xia, Department of Mathematics, HKUST

Can Yang, Department of Mathematics, HKUST

Logistics Committee:

Biying Hu, Department of Mathematics, HKUST

Ziyue Tan, Department of Mathematics, HKUST

Hongrui Wang, Department of Mathematics, HKUST

Can Yang, Department of Mathematics, HKUST

Workshop Chairs:

Congyuan Duan, Department of Mathematics, HKUST

Ziyue Tan, Department of Mathematics, HKUST

Rui Li, Department of Mathematics, HKUST

Xiaomeng Wan, Department of Mathematics, HKUST

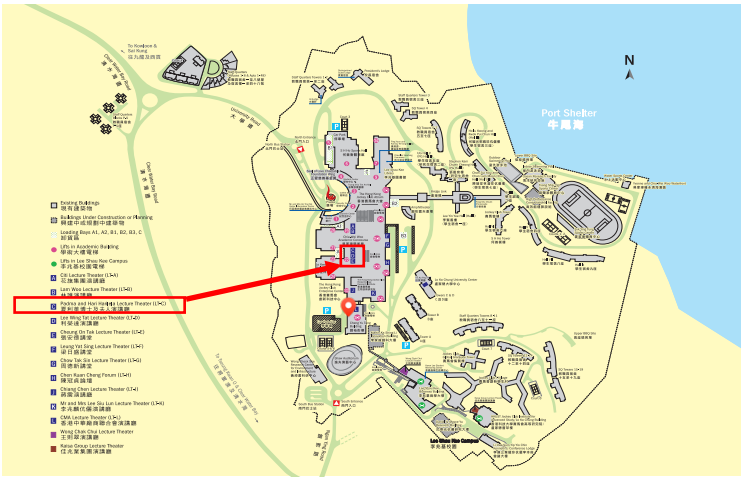
Jiarui Zhang, Department of Mathematics, HKUST

Supporting Staff:

Maggie Hui, Department of Mathematics, HKUST

Noreen Lee, Department of Mathematics, HKUST

Venue: LT-C, HKUST



Time	Talk Title	Host
9:00-9:10	**Welcome to HKUST!** *Shiqing Ling, HKUST* *Dong Xia, HKUST*	**Can Yang**
9:10-9:30	**On the Optimality of Robust Inference on the Mean Outcome under Optimal Treatment Regime** *Xinzhou Guo, HKUST*	**Congyuan Duan**
9:30-9:50	**Learning summary statistics for approximate Bayesian computation** *Rong Tang, HKUST*	
9:50-10:10	**Learning to Bid in Non-Stationary Repeated First-Price Auctions** *Zihao Hu, HKUST*	
10:10-10:40	**Coffee break**	
10:40-11:00	**Deciphering proteins in Alzheimer's disease: A new Mendelian randomization method integrated with AlphaFold3 for 3D structure prediction** *Zhonghua Liu, Columbia University*	**Ziyue Tan**
11:00-11:20	**Identifying genetic risk variants for autism spectrum disorder: a statistical framework with error rate control** *Yi Yang, CityU*	
11:20-11:40	**traceCB: Trans-ancestry cell-type-specific eQTLs mapping by integrating scRNA-seq and bulk data** *Mingxuan Cai, CityU*	
	Lunch at Uni Biostro (By invitation only)	
13:30-13:50	**Faster Convergence and Acceleration for Diffusion-Based Generative Models** *Gen Li, CUHK*	**Rui Liu**
13:50-14:10	**Stochastic Gradient Methods: Bias, Stability and Generalization** *Yunwen Lei, HKU*	
14:10-14:30	**Statistical Inference for Differentially Private Stochastic Gradient Descent** *Zhanrui Cai, HKU*	
14:30-15:00	**Coffee Break**	
15:00-15:20	**Integrative Analysis and Regulatory Inference in Spatial Multi-Omics Data**	**Xiaomeng Wan**

	Zhixiang Lin, CUHK	
15:20-15:40	**Data integration of atlas-scale single-cell multi-modal data** *Yingxin Lin, CUHK*	
15:40-16:00	**Dissecting multicellular cell niches from spatial transcriptomics data with NicheScope** *Jiashun Xiao, Sun Yat-Sen University*	
16:00-16:30	**Coffee Break**	
16:30-16:50	**Hybridred Censored Quantile Regression Forest** *Huichen Zhu, CUHK*	**Jiarui Zhang**
16:50-17:10	**Generalized Median of Means Principle for Bayesian Inference** *Shunan Yao, HKBU*	
17:10-17:30	**When Tukey meets Chauvenet: a new boxplot criterion for outlier detection** *Tiejun Tong, HKBU*	
	Dinner at China Garden **(南北小厨, by invitation only)**	

1. On the Optimality of Robust Inference on the Mean Outcome under Optimal Treatment Regime

Xinzhou Guo

The Hong Kong University of Science and Technology

Abstract: When an optimal treatment regime (OTR) is considered, we need to address the question of how good the OTR is in a valid and efficient way. The classical statistical inference applied to the mean outcome under the OTR, assuming the OTR is the same as the estimated OTR, might be biased when the regularity assumption that the OTR is unique is violated. Although several methods have been proposed to allow nonregularity in inference on the mean outcome under the OTR, the optimality of such inference is unclear due to challenges in deriving semiparametric efficiency bounds under potential nonregularity. In this paper, we address the bias issue induced by potential nonregularity via adaptive smoothing over the estimated OTR and develop a valid inference procedure on the mean outcome under the OTR regardless of whether the regularity assumption is satisfied or not. We establish the optimality of the proposed method by deriving a lower bound of the asymptotic variance for the robust asymptotically linear unbiased estimator to the mean outcome under the OTR and showing that our proposed estimator achieves the variance lower bound. The considered class of the estimator is general and includes the efficient regular estimator and the current state-of-the-art approach allowing nonregularity, and the derived lower bound of the asymptotic variance can be viewed as an extension of the classical semiparametric theory for OTR to a more general scenario allowing nonregularity. The merit of the proposed method is demonstrated by re-analyzing the ACTG 175 trial.

Bio: Xinzhou Guo is an Assistant Professor in the Department of Mathematics at the Hong Kong University of Science and Technology. He received his B.S. in Applied Mathematics from Peking University and Ph.D. in Statistics from the University of Michigan. Prior to joining HKUST in 2021, he did a postdoc at Harvard University. His main research interests are subgroup analysis, resampling methods, precision medicine and regulatory decision-making.

2. Learning summary statistics for approximate Bayesian computation

Rong Tang

The Hong Kong University of Science and Technology

Abstract: The challenge of performing Bayesian inference in models where likelihood functions are di icult to evaluate but sampling is straightforward has driven the development of likelihood-free methods such as approximate Bayesian computation (ABC). A key element in ABC is the use of summary statistics to reduce the dimensionality of the data, thereby avoiding the curse of dimensionality in nonparametric conditional density estimation as the observed data size grows. We propose a general framework for learning informative summary statistics and the subsequent posterior inference based on the summaries. The proposed approach

involves jointly training the summary statistic and an auxiliary function that models the posterior distribution given the summaries, by minimizing proxies of suitable expected discrepancies between the true posterior and the posterior given the summaries. In addition, a more refined posterior approximations for specific datasets can be obtained by integrating this approach with MCMC methods

Bio: Rong Tang is an assistant professor in the Department of Mathematics at the Hong Kong University of Science and Technology (HKUST). She received her Ph.D. in Statistics from the University of Illinois at Urbana-champaign (UIUC). Her research interests include Bayesian inference, Machine learning theory and Complexity of sampling.

3. Learning to Bid in Non-Stationary Repeated First-Price Auctions

Zihao Hu

The Hong Kong University of Science and Technology

Abstract: First-price auctions have recently gained significant traction in digital advertising markets, exemplified by Google's transition from second-price to first-price auctions. Unlike in second-price auctions, where bidding one's private valuation is a dominant strategy, determining an optimal bidding strategy in first-price auctions is more complex. From a learning perspective, the learner (a specific bidder) can interact with the environment (other bidders, i.e., opponents) sequentially to infer their behaviors. Existing research often assumes specific environmental conditions and benchmarks performance against the best fixed policy (static benchmark). While this approach ensures strong learning guarantees, the static benchmark can deviate significantly from the optimal strategy in environments with even mild non-stationarity. To address such scenarios, a dynamic benchmark—representing the sum of the highest achievable rewards at each time step—offers a more suitable objective. However, achieving no-regret learning with respect to the dynamic benchmark requires additional constraints. By inspecting reward functions in online first-price auctions, we introduce two metrics to quantify the regularity of the sequence of opponents' highest bids, which serve as measures of non-stationarity. We provide a minimax-optimal characterization of the dynamic regret for the class of sequences of opponents' highest bids that satisfy either of these regularity constraints. Our main technical tool is the Optimistic Mirror Descent (OMD) framework with a novel optimism configuration, which is well-suited for achieving minimax-optimal dynamic regret rates in this context.

Bio: Dr. Zihao Hu is a Research Assistant Professor in the Department of Mathematics at the Hong Kong University of Science and Technology. He earned his Ph.D. in Machine Learning from Georgia Tech under the supervision of Prof. Jacob Abernethy. Hu's research spans operations research and mathematical optimization, with a focus on online advertising auctions and Riemannian optimization. His work has appeared in top venues such as COLT, NeurIPS, and AISTATS. He collaborates closely with Prof. Yuan Yao, Prof. Jiheng Zhang, and Prof.

Zhengyuan Zhou.

4. **Deciphering proteins in Alzheimer's disease: A new Mendelian randomization method integrated with AlphaFold3 for 3D structure prediction**

Zhonghua Liu

Columbia University

Abstract: Hidden confounding biases hinder identifying causal protein biomarkers for Alzheimer's disease in non-randomized studies. While Mendelian randomization (MR) can mitigate these biases using protein quantitative trait loci (pQTLs) as instrumental variables, some pQTLs violate core assumptions, leading to biased conclusions. To address this, we propose MR-SPI, a novel MR method that selects valid pQTL instruments using Leo Tolstoy's Anna Karenina principle and performs robust post-selection inference. Integrating MR-SPI with AlphaFold3, we developed a computational pipeline to identify causal protein biomarkers and predict 3D structural changes. Applied to genome-wide proteomics data from 54,306 UK Biobank participants and 455,258 subjects (71,880 cases and 383,378 controls) for a genome-wide association study of Alzheimer's disease, we identified seven proteins (TREM2, PILRB, PILRA, EPHA1, CD33, RET, and CD55) with structural alterations due to missense mutations. These findings offer insights into the etiology and potential drug targets for Alzheimer's disease.

Bio: Dr. Zhonghua Liu is an Assistant Professor of Biostatistics at Columbia University and a member of the Data Science Institute. His research focuses on robust and scalable causal inference, semiparametric theory, and statistical machine learning, with applications in genomics, proteomics, and precision medicine. He has published in leading statistics journals including JASA, Biometrika, Biometrics, and JRSS-B, as well as top interdisciplinary venues such as Cell Genomics, Nature Computational Science, and premier ML/AI conferences like NeurIPS and ICML. His work integrates rigorous statistical methodology with modern machine learning to advance causal discovery in complex biomedical data.

5. **Identifying genetic risk variants for autism spectrum disorder: a statistical framework with error rate control**

Yi Yang

City University of Hong Kong

Abstract: Autism spectrum disorder (ASD) is an early-onset neurodevelopmental disorder with a strong heritable component driven by multiple genetic risk variants. A conventional method for identifying these risk variants is to use family-based association tests on trio data comprised of affected children and their parents. However, conventional methods often have low power when applied to trio data, primarily due to the difficulty of obtaining large sample sizes. To overcome the low power of conventional methods, we propose statistical methods that increase power by integrating a knockoff framework and leveraging external case-control data.

Moreover, our methods rigorously control the false discovery rate to guarantee that a high proportion of identified variants are true risk factors for ASD. Through simulations and real-data applications to multiple ASD study cohorts, we demonstrate that our methods can identify genetic risk variants missed by conventional methods.

Bio: Dr. Yi Yang is an Assistant Professor in the Department of Biostatistics at City University of Hong Kong (CityU). He received his PhD and MS in biostatistics from the University of Minnesota and BEng in management information systems from Zhejiang University. Prior to joining CityU, he served as a postdoctoral research scientist in the Department of Biostatistics at Columbia University. Dr. Yang's research focuses on variable selection methods for high-dimensional data using knockoff statistics and Bayesian hierarchical models. He has developed a number of statistical methods and packages to identify genetic risk variants for human diseases including autism, osteosarcoma, Crohn's disease, and Kawasaki disease. His research is supported by the Research Grants Council of Hong Kong Early Career Scheme.

6. traceCB: Trans-ancestry cell-type-specific eQTLs mapping by integrating scRNA-seq and bulk data

Mingxuan Cai

City University of Hong Kong

Abstract: The emergence of expression quantitative trait loci (eQTLs) studies offered a unique opportunity to connect GWAS variations to gene expression in relevant biological conditions. Despite the promise, two major challenges remain for existing eQTL studies. First, many eQTL effects are cell-type-specific. Traditional eQTL analyses often rely on bulk RNA sequencing (bulk RNA-seq) data, obscuring the cell-type-specific genetic effects. While the recent advancements in single-cell RNA sequencing (scRNA-seq) allow for deeper investigation at the cell-type level, they are limited by increased technical noise and smaller sample sizes. Second, current eQTL findings are predominantly based on European samples, making it difficult to extrapolate these results to non-European ancestries and hindering the interpretation of GWAS findings in diverse populations. We introduce a unified framework for trans-ancestry cell-type-specific eQTLs (ct-eQTLs) mapping by integrating summary statistics from bulk RNA-seq and scRNA-seq datasets (traceCB). TraceCB leverages several unique features. First, it boosts the statistical power by leveraging a larger scRNA-seq data from the European population while accounting for ancestral heterogeneities. Second, using the scRNA-seq data as a bridge, it further improves ct-eQTL mapping by integrating a large bulk RNA-seq data (e.g., the GTEx cohort). Third, unlike existing meta-analysis methods, it effectively accounts for the heterogeneous eQTL effects across populations and cell types, yielding well-calibrated p-values. Fourth, it only requires summary statistics of eQTL studies as its input, making it widely applicable to various tissues, cell types, and populations.We demonstrate the effectiveness of traceCB through extensive simulations and analyses of real datasets, including

multiple sc-eQTL datasets from peripheral blood mononuclear cells and bulk eQTL data from the GTEx and eQTLGen consortia. Our results show that traceCB achieves a substantial gain in the statistical power for detecting ct-eQTLs compared to existing methods. Colocalization analysis of traceCB output and GWAS data reveals novel cell-type-specific regulatory mechanisms, enhancing our understanding of the genetic basis of complex traits in the African and East Asian populations at a cellular resolution.

Bio: Dr. Cai is an Assistant Professor at Department of Biostatistics, City University of Hong Kong. He obtained his PhD degree from The Hong Kong University of Science and Technology in 2022. His broad area of interest lies in statistical machine learning and data science with applications in genetics and genomics data analysis.

7. Faster Convergence and Acceleration for Diffusion-Based Generative Models

Gen Li

The Chinese University of Hong Kong

Abstract: 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.

Bio: Gen Li is currently an assistant professor in the Department of Statistics and Data Science at the Chinese University of Hong Kong. He received the Ph.D. in the Department of Electronic Engineering at Tsinghua University in 2021, and received the bachelor's degree from the Department of Electronic Engineering and Department of Mathematics at Tsinghua University in 2016. His research interests include diffusion based generative model, reinforcement learning, high-dimensional statistics, machine learning.

8. Stochastic Gradient Methods: Bias, Stability and Generalization

Yunwei Lei

The University of Hong Kong

Abstract: Recent developments of stochastic optimization often suggest biased gradient estimators to improve either the robustness, communication efficiency or computational speed. Representative biased stochastic gradient methods (BSGMs) include Zeroth-order stochastic gradient descent (SGD), Clipped-SGD and SGD with delayed gradients. In this talk, we present the first framework to study the stability and generalization of BSGMs for convex and smooth problems. We apply our general result to develop the first stability bound for Zeroth-order SGD with reasonable step size sequences, and the first stability bound for Clipped-SGD.

Bio: Yunwen Lei is currently an Assistant Professor at the Department of Mathematics, The University of Hong Kong. His main research interests include machine learning, statistical learning theory and stochastic optimization.

9. Statistical Inference for Differentially Private Stochastic Gradient Descent

Zhanrui Cai

The University of Hong Kong

Abstract: Privacy preservation in machine learning, particularly through Differentially Private Stochastic Gradient Descent (DP-SGD), is critical for sensitive data analysis. However, existing statistical inference methods for SGD predominantly focus on cyclic subsampling, while DP-SGD requires randomized subsampling. This paper first bridges this gap by establishing the asymptotic properties of SGD under the randomized rule and extending these results to DP-SGD. For the output of DP-SGD, we show that the asymptotic variance decomposes into statistical, sampling, and privacy-induced components. Two methods are proposed for constructing valid confidence intervals: the plug-in method and the random scaling method. We also perform extensive numerical analysis, which shows that the proposed confidence intervals achieve nominal coverage rates while maintaining privacy.

Bio: Zhanrui Cai is a tenure-track assistant professor in the Faculty of Business and Economics at the HKU Business School. His research interests include statistical inference, differential privacy, etc.

10. Integrative Analysis and Regulatory Inference in Spatial Multi-Omics Data

Zhixiang Lin

The Chinese University of Hong Kong

Abstract: New spatial multi-omics technologies, which jointly profile transcriptome and epigenome/prot markers for the same tissue section, have expanded the frontiers of spatial techniques. Here we introduce MultiGATE, which utilizes a two-level graph attention auto-encoder to integrate the multi-modality and spatial information in spatial multi-omics data. The key feature of MultiGATE is that it simultaneously performs embedding of the spatial pixels and infers the cross-modality regulatory relationship, which allows deeper data integration and provides insights on transcriptional regulation. We evaluated the performance of MultiGATE

on spatial multi-omics datasets obtained from different tissues and platforms. Through effectively integrating spatial multi-omics data, MultiGATE both enhances the extraction of latent embeddings of the pixels and boosts the inference of transcriptional regulation for cross-modality genomic features.

Bio: Zhixiang Lin is an associate professor at the Chinese University of Hong Kong. He received his B.S. from Tsinghua University and Ph.D. from Yale University.

11. Data integration of atlas-scale single-cell multi-modal data

Yingxin Lin

The Chinese University of Hong Kong

Abstract: Single-cell technologies offer unprecedented opportunities to dissect gene regulatory mechanisms in context-specific ways. The recent rise of multi-sample, multi-condition, and multi-cohort single-cell studies enables researchers to explore diverse cellular states across a range of biological contexts. While individual studies provide valuable insights, the effective integration of large-scale, heterogeneous datasets holds the key to uncovering regulatory programs that would otherwise remain hidden. Despite progress in integrating multi-modal data from single tissues, the scale and complexity of cell atlas–level data continue to pose significant challenges. In this talk, I will present scalable, interpretable, and biologically informed algorithms we have developed to address these challenges, including transfer learning and statistical methods that advance single-cell data science and help translate high-dimensional biomedical data into biological insights.

Bio: I am a Tenure Track Assistant Professor in Department of Statistics and Data Science, The Chinese University of Hong Kong (CUHK). Previously, I was a Postdoctoral Associate in the Department of Biostatistics at the Yale School of Public Health, advised by Prof. Hongyu Zhao. I completed my PhD in Statistics under the supervision of Prof. Jean Yang, A/Prof. Rachel Wang and A/Prof. John Ormerod at the University of Sydney in 2022. In the broad areas of statistics, data science and bioinformatics, the central theme of my research is to formulate the data science challenges motivated by biomedical and biotechnological data into computational problems and tackle them by developing novel statistical and machine learning methodologies as well as analytical workflows to enable new scientific discoveries in various diseases.

12. Dissecting multicellular cell niches from spatial transcriptomics data with NicheScope

Jiashun Xiao

Sun Yat-sen University

Abstract: Tissues are composed of diverse local microenvironments, or cell niches, formed by the coordinated interactions of multiple cell types. Deciphering the spatial organization and

functional properties of these niches is essential for understanding tissue physiology and disease mechanisms, yet remains challenging with current computational approaches. Here, we present NicheScope, a robust and scalable framework for transcriptome-wide identification and characterization of multicellular cell niches from spatial transcriptomics data. NicheScope systematically models the association between an index cell type's gene expression and its spatial neighborhood, enabling unbiased detection of niches defined by specific combinations of cell types and gene expression programs. We demonstrate the versatility and reproducibility of NicheScope across multiple spatial transcriptomics platforms and tissue types, including normal lymph node, non-small cell lung cancer, and head and neck squamous cell carcinoma. NicheScope reveals established tissue structures, clinically relevant microenvironments, and shared and condition-specific niches under multiple conditions. Our results establish NicheScope as a powerful tool for dissecting the spatial and functional organization of complex tissues, providing new insights into multicellular coordination in health and disease.

Bio: Dr. Xiao Jiashun is an associate professor at the School of Public Health of Sun Yat-sen University. He obtained a Bachelor's degree in Bioinformatics from Southern University of Science and Technology (SUSTech) in 2017, followed by a PhD in Mathematics from the Hong Kong University of Science and Technology（HKUST）in 2022, with a focus on data science, applied statistics, and bioinformatics. He leads several research projects, including the Young Scientists Fund of the National Natural Science Foundation of China, Guangdong Natural Science Foundation General Project, etc. He has published multiple high-impact papers as first or corresponding author (including co-authorship) in top journals such as Nature Communications and The American Journal of Human Genetics. His current research interests include AI for Science, multi-omics data integration and analysis, and precision medicine with personalized disease prediction.

13. Hybridred Censored Quantile Regression Forest

Huichen Zhu

The Chinese University of Hong Kong

Abstract: Understanding heterogeneous treatment responses is crucial for advancing precision medicine. Various factors, including patient demographics, genetic predispositions, and environmental exposures, can significantly impact treatment responses. Additionally, individuals may respond differently to the same treatment across the outcome distribution. When outcomes are censored—a common occurrence in health applications—traditional mean-based methods often fall short of capturing the true variability in patient responses. In this paper, we introduce a novel forest framework designed to assess heterogeneous quantile treatment effects in the context of time-to-event outcomes. We propose a new partitioning procedure specifically targeted at this objective. A post-forest estimation procedure enables us to effectively address the challenges of censoring and heterogeneity, both in covariate space and

across the outcome distribution. We establish the validity of our method by providing comprehensive large-sample theory that demonstrates the consistency and asymptotic normality of our estimates. Extensive simulation studies validate the efficacy and stability of the Hybrid Censored Quantile Regression Forest (HCQRF). We apply HCQRF to a clinical trial of colorectal cancer, where the treatment effect estimates derived from our framework provide valuable insights, and the variable importance results yield meaningful information.

Bio: Got my PhD. degree from the department of biostatistics in Columbia University. Research mainly focus on quantile regression, random forest, machine learning, heterogeneous treatment effect and etc.

14. Generalized Median of Means Principle for Bayesian Inference

Shunan Yao

Hong Kong Baptist University

Abstract: The topic of robustness is experiencing a resurgence of interest in the statistical and machine learning communities. In particular, robust algorithms making use of the so-called median of means estimator were shown to satisfy strong performance guarantees for many problems, including estimation of the mean, covariance structure as well as linear regression. In this work, we propose an extension of the median of means principle to the Bayesian framework, leading to the notion of the robust posterior distribution. In particular, we (a) quantify robustness of this posterior to outliers, (b) show that it satisfies a version of the Bernstein-von Mises theorem that connects Bayesian credible sets to the traditional confidence intervals, and (c) demonstrate that our approach performs well in applications.

Bio: Dr. Shunan Yao is an assistant professor from the math department at Hong Kong Baptist University.

15. When Tukey meets Chauvenet: a new boxplot criterion for outlier detection

Tiejun Tong

Hong Kong Baptist University

Abstract: The box-and-whisker plot, introduced by Tukey (1977), is one of the most popular graphical methods in descriptive statistics. On the other hand, however, Tukey's boxplot is free of sample size, yielding the so-called "one-size-fits-all" fences for outlier detection. Although improvements on the sample size adjusted boxplots do exist in the literature, most of them are either not easy to implement or lack justification. As another common rule for outlier detection, Chauvenet's criterion uses the sample mean and standard derivation to perform the test, but it is often sensitive to the included outliers and hence is not robust. In this paper, by combining Tukey's boxplot and Chauvenet's criterion, we introduce a new boxplot, namely the Chauvenet-type boxplot, with the fence coefficient determined by an exact control of the outside rate per observation. Our new outlier criterion not only maintains the simplicity of the boxplot from a

practical perspective, but also serves as a robust Chauvenet's criterion. Simulation study and a real data analysis on the civil service pay adjustment in Hong Kong demonstrate that the Chauvenet-type boxplot performs extremely well regardless of the sample size, and can therefore be highly recommended for practical use to replace both Tukey's boxplot and Chauvenet's criterion. Lastly, to increase the visibility of the work, a user-friendly R package named 'ChauBoxplot' has also been officially released on CRAN.

Bio: Tiejun Tong is a Professor in the Department of Mathematics at Hong Kong Baptist University. He received his Ph.D. from the University of California, Santa Barbara in 2005. From 2005 to 2007, he conducted postdoctoral research at Yale University. Between 2007 and 2010, he was an Assistant Professor at the University of Colorado Boulder. Since 2010, he has been with the Department of Mathematics at Hong Kong Baptist University.

His main research interests include nonparametric regression models, high-dimensional data analysis, meta-analysis, and evidence-based medicine. He has published over 100 research papers in internationally renowned journals such as JASA, Biometrika, Statistical Science, JMLR, and Nature Communications. Among these are several ESI Hot Papers and ESI Highly Cited Papers, with his most cited paper receiving over 9,200 citations.