Trans-Ancestry Cell-Type-Specific eQTLs Mapping by Integrating scRNA-seq and Bulk Data

Wenxin Jiang[1], **Mingxuan Cai[1,*]**

[1]*Department of Biostatistics, City University of Hong Kong*

ABSTRACT

Genome-wide association studies (GWAS) have successfully identified numerous genetic variants associated with complex traits and diseases, primarily located in noncoding regions. The emergence of expression quantitative trait loci (eQTLs) studies offered a unique opportunity to connect these 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, which measures the average gene expression levels across heterogeneous cell types and states, 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.

Here, 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). This approach not only controls type I errors at nominal levels but also enhances the statistical power for identifying ct-eQTLs in underrepresented populations. 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 power gain in ct-eQTL mapping compared to existing methods.

Colocalization analysis of traceCB output and GWAS data reveals novel cell-type-specific regulatory mechanisms, elucidating the genetic basis of complex traits in the African and East Asian populations at a cellular resolution.

Keywords: single-cell eQTL; Statistical Genetics; GWAS

Back to Sessions List

Online Stochastic Optimization with Offline Data

Jize Xie, **Yi Chen**, Rachel Q. Zhang

Department of Industrial Engineering and Decision Analytics

Hong Kong University of Science and Technology

ABSTRACT

For online decision-making under uncertainty, data arrives in a streaming manner and decisions must be made sequentially. As a prominent modeling tool, online stochastic optimization finds broad applications, with the online stochastic gradient descent (OSGD) algorithm being a standard solution. While much existing research focuses on the online process solely, offline data are often available ahead, offering opportunities to enhance the online decision-making's quality. In this work, we investigate this problem by integrating offline data into the standard OSGD algorithm. We first propose an algorithm called OSGDwO that leverages offline information for a better online process initialization and stochastic gradient variance reduction. When offline and online data are identically distributed, by introducing a novel analytical framework, we demonstrate that the OSGDwO algorithm attains regrets decaying at the rate of O(1/N), where N is the offline sample size. To mitigate the risk due to a potential shift between offline and online distributions, we further develop a robustified algorithm by incorporating sequential hypothesis tests. This algorithm adaptively determines whether to utilize offline data, achieving the near-optimal performance guarantee regardless of the distributional shift's magnitude. Extensive simulation studies demonstrate the improved and robust performances of our algorithms. Finally, we apply our algorithms to two operations management problems: auction bidding and inventory management, illustrating both the practicality and extensibility of our approaches.

Keywords: stochastic optimization, sequential analysis, regret analysis

Adaptive Bayesian Optimization with Consistent Smoothness Estimation and Hyperparameters Exploration

Youxin Wang[1], **Saifei Sun[2]**, Shouri Hu[3]

[1,3]*School of Mathematical Sciences, University of Electronic Science and Technology of China.*

[2]*Department of Biostatistics, City University of Hong Kong.*

ABSTRACT

We present a novel algorithm, Adaptive Matérn Kernel Bayesian Optimization (AMKBO), designed to address the challenges of hyperparameter uncertainty in Gaussian Process (GP)-based Bayesian Optimization. The proposed method accelerates convergence by consistently estimating the smoothness parameter of the Matérn covariance kernel, adjusting the length-scale hyperparameter adaptively, and incorporating an opposition-based learning mechanism into acquisition function optimization. Specifically, AMKBO generates initial sampling points using random curves and constructs an estimator for the smoothness parameter of the covariance kernel, allowing the kernel to model functions with varying degrees of smoothness. During optimization, a refined length-scale adjustment strategy is employed to achieve a more balanced trade-off between exploration and exploitation, while preventing excessively aggressive exploration. In parallel, the opposition-based learning mechanism enhances the search coverage and computational efficiency in acquisition function optimization. Experimental results on synthetic benchmark functions and real-world problems demonstrate that AMKBO outperforms existing methods by achieving faster convergence and avoiding local optima compared to existing methods.

Keywords: Gaussian processes, Bayesian Optimization, hyperparameter adaptation, Matérn covariance kernel, smoothness estimation, opposition-based learning.

A Preferential Latent Space Model for Text Networks

Maoyu Zhang[1], **Biao Cai[2],** Dong Li[3], Xiaoyue Niu[4], Jingfei Zhang[1]

[1]*Goizueta Business School, Emory University, Atlanta, GA*

[2]*Department of Decision Analytics and Operations, City University of Hong Kong, Hong Kong*

[3]*Department of Statistics and Data Science, Tsinghua University, Beijing, China*

[4]*Department of Statistics, Pennsylvania State University, University Park, PA*

ABSTRACT

Network data enriched with textual information, referred to as text networks, arise in a wide range of applications, including email communications, scientific collaborations, and legal contracts. In such settings, both the structure of interactions (i.e., who connects with whom) and their content (i.e., what is communicated) are useful for understanding network relations. Traditional network analyses often focus only on the structure of the network and discard the rich textual information, resulting in an incomplete or inaccurate view of interactions. In this paper, we introduce a new modeling approach that incorporates texts into the analysis of networks using topic-aware text embedding, representing the text network as a generalized multi-layer network where each layer corresponds to a topic extracted from the data. We develop a new and flexible latent space network model that captures how node-topic preferences directly modulate edge formation, and establish identifiability conditions for the proposed model. We tackle model estimation with a projected gradient descent algorithm, and further discuss its theoretical properties. The efficacy of our proposed method is demonstrated through simulations and an analysis of an email network.

Keywords: text networks; latent space model; non-convex optimization; text analysis; topic-aware embedding.