Model-Free Inference for Characterizing Protein Mutations through a Coevolutionary Lens

Fan Yang[1], **Zhao Ren[1]**, Wen Zhou[2], Robert Jernigan[3]

[1]Department of Statistics, University of Pittsburgh,

[2]Department of Biostatistics, New York University,

[3]Department of Biochemistry, Biophysics and Molecular Biology, Iowa State University

ABSTRACT

Multiple sequence alignment (MSA) data play a crucial role in the study of protein mutations, with contact prediction being a notable application. Existing methods are often model-based or algorithmic and typically do not incorporate statistical inference to quantify the uncertainty of the prediction outcomes. To address this, we propose a novel framework that transforms the task of contact prediction into a statistical testing problem. Our approach is motivated by the partial correlation for continuous random variables. With one-hot encoding of MSA data, we are able to construct a partial correlation graph for multivariate categorical variables. In this framework, two connected nodes in the graph indicate that the corresponding positions on the protein form a contact. A new spectrum-based test statistic is introduced to test whether two positions are partially correlated. Moreover, the new framework enables the identification of amino acid combinations that contribute to the correlation within the identified contacts, an important but largely unexplored aspect of protein mutations. Numerical experiments demonstrate that our proposed method is valid in terms of controlling Type I errors and powerful in general. Real data applications on various protein families further validate the practical utility of our approach in coevolution and mutation analysis.

Keywords: Multivariate categorical data; Partial correlation; Precision matrix; Multiple sequence alignment; Protein mutation

Back to Sessions List

Microbial Causal Mediation Analysis under Spatially Correlated Exposure

Sooran Kim, Chan Wang, Soyoung Kwak, Jiyoung Ahn, **Huilin Li**

New York University, Grossman School of Medicine, Department of Population Health

ABSTRACT

Recent research suggests that the environmental exposure has been linked with the human microbiome, which can contribute to human health and disease. In particular, air pollution exposure plays a critical role in cancer presentation and prognosis. Understanding the complex interplay among disease status, microbiome abundances, and environmental exposure is therefore essential. Environmental exposures, such as air pollution, are often spatially autocorrelated, posing significant challenges for statistical analysis. In this work, we explore the associations among three key factors—disease status, environmental exposures, and microbiome abundances—while incorporating the spatial dependence of environmental exposures and the intrinsic correlation within microbiome.

Key words: Microbiome, causal mediation, spatial correlation

Back to Sessions List

Testing Composite Null Hypotheses with High-Dimensional Dependent Data: A Computationally Scalable FDR-Controlling Procedure

Pengfei Lyu, Xianyang Zhang and **Hongyuan Cao**

Duke University

Texas A & M University

Florida State University

ABSTRACT

Testing composite null hypotheses is fundamental to many scientific applications, including mediation and replicability analyses, and becomes particularly challenging in high-throughput settings involving tens of thousands of features. Existing high dimensional composite null hypotheses testing often ignores the dependence structure among features, leading to overly conservative or liberal results. To address this limitation, we develop a four-state hidden Markov model (HMM) for bivariate p-value sequences arising from two-study replicability analysis. This model captures local dependence among features and accommodates study-specific heterogeneity. Based on the HMM, we propose a multiple testing procedure that asymptotically controls the false discovery rate (FDR). Extending this framework to more than two studies is computationally intensive, with complexity growing exponentially in the number of studies $n.$ To address this scalability issue, we introduce a novel e-value framework that reduces computational complexity to quadratic in $n,$ while preserving asymptotic FDR control. Extensive simulations demonstrate that our method achieves higher power than existing approaches at comparable FDR levels. When applied to genome-wide association studies (GWAS), the proposed approach identifies novel biological findings that are missed by current methods.

Keywords: Composite null hypotheses, e-values, false discovery rate, hidden Markov model, high dimension, non-parametric maximum likelihood estimation

Back to Sessions List

D-CDLF: Decomposition of Common and Distinctive Latent Factors for Multi-view High-Dimensional Data

Hai Shu[1], Hongtu Zhu[2]

[1]*Department of Biostatistics, New York University*

[2]*Department of Biostatistics, The University of North Carolina at Chapel Hill*

ABSTRACT

Modern biomedical studies often collect multi-view data, that is, multiple types of data measured on the same set of objects. A typical approach to the joint analysis of multiple high-dimensional data views is to decompose each view's data matrix into three parts: a low-rank common-source matrix generated by common latent factors of all data views, a low-rank distinctive-source matrix generated by distinctive latent factors of the corresponding data view, and an additive noise matrix. Existing decomposition methods often focus on the uncorrelatedness between the common latent factors and distinctive latent factors, but inadequately address the equally necessary uncorrelatedness between distinctive latent factors from different data views. We propose a novel decomposition method, called Decomposition of Common and Distinctive Latent Factors (D-CDLF), to effectively achieve both types of uncorrelatedness. Consistent estimators of our D-CDLF method are established, demonstrating reasonably good finite-sample numerical performance. The superiority of D-CDLF over state-of-the-art methods is corroborated by simulations and the analysis of imaging and genomic data from the Alzheimer's Disease Neuroimaging Initiative.

Keywords: Canonical correlation analysis; Common and distinctive latent factors; Data integration; Orthogonality constraint

Back to Sessions List