Latent Noise Injection for Private and Statistically Aligned Synthetic Data Generation

Rex Shen[1], **Lu Tian[2]**

[1]*Google Inc.z*

[2] *Stanford University*

ABSTRACT

Synthetic data generation has become essential for scalable, privacy-preserving statistical analysis. While standard generative-model–based approaches— such as those using Normalizing Flows— are widely adopted, they often exhibit slow convergence in high-dimensional settings, frequently failing to achieve the canonical root n rate when approximating the true data distribution. To address these limitations, we propose a *Latent Noise Injection* method built on Masked Autoregressive Flows. Rather than sampling directly from the trained model, our approach perturbs each observed data point in the latent space and maps it back to the data domain. This construction preserves a one-to-one correspondence between observed and synthetic samples, enabling synthetic outputs that more faithfully reflect the underlying distribution—particularly in challenging high-dimensional regimes where traditional sampling deteriorates. Our procedure satisfies local (ε,δ)-differential privacy and introduces a single perturbation parameter governing the privacy–utility trade-off. Although estimators based on a single synthetic dataset may converge slowly, we show both theoretically and empirically that aggregating results across multiple studies in a meta-analytic framework restores classical efficiency and yields consistent, reliable inference. With an appropriately calibrated perturbation parameter, Latent Noise Injection achieves strong statistical fidelity to the original data while providing robustness against membership-inference attacks. These results position our method as a compelling alternative to conventional flow-based sampling for synthetic data sharing in decentralized and privacy-sensitive domains, such as biomedical research.

Keywords: Synthetic data, privacy protection, normalizing flows.

Back to Sessions List

MR2G: A Novel Framework for Causal Network Inference Using GWAS Summary Data

Zhaotong Lin[1], Wei Pan[2], **Haoran Xue[3]**

[1]*Department of Statistics, Florida State University*

[2]*Division of Biostatistics & Health Data Science, University of Minnesota*

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

ABSTRACT

Inferring a causal network among multiple traits is essential for unravelling complex biological relationships and informing interventions. Mendelian randomization (MR) has emerged as a powerful tool for causal inference, utilizing genetic variants as instrumental variables (IVs) to estimate causal effects. However, when the directions of causal relationships among traits are unknown, reconstructing the underlying causal network becomes challenging. In particular, the presence of cycles or feedback loops, which are common in biological systems, poses additional challenges for causal network inference, and remains largely under-studied with standard MR approaches and existing IV-based network inference methods. To address these issues, we introduce MR2G, a new statistical framework that enables robust inference of causal networks, including those with cycles, directly from GWAS summary statistics. MR2G is built on a formally defined recursive causal graph model that rigorously links direct causal effects to MR estimands. It recovers a biologically interpretable causal network from pairwise MR effect estimates, while incorporating a network-informed IV screening strategy to reduce pleiotropic bias and improve robustness. Through realistic simulations, MR2G demonstrates superior accuracy and robustness in recovering complex causal structures, including those involving feedback loops. We apply MR2G to GWAS summary statistics for six complex diseases and nine cardiometabolic risk factors. MR2G not only recovers well-established causal pathways but also uncovers multiple feedback relationships, highlighting its utility in disentangling complex and biologically plausible causal networks from large-scale genetic data.

Keywords: Mendelian randomization; cyclic causal network; instrumental variable.

Back to Sessions List

Modeling Non-Uniform Hypergraphs Using Determinantal Point Processes

Ji Zhu

University of Michigan

ABSTRACT

Most statistical models for networks focus on pairwise interactions between nodes. However, many real-world networks involve higher-order interactions among multiple nodes, such as co-authors collaborating on a paper. Hypergraphs provide a natural representation for these networks, with each hyperedge representing a set of nodes. The majority of existing hypergraph models assume uniform hyperedges (i.e., edges of the same size) or rely on diversity among nodes. In this work, we propose a new hypergraph model based on non-symmetric determinantal point processes. The proposed model naturally accommodates non-uniform hyperedges, has tractable probability mass functions, and accounts for both node similarity and diversity in hyperedges. For model estimation, we maximize the likelihood function under constraints using a computationally efficient projected adaptive gradient descent algorithm. We establish the consistency and asymptotic normality of the estimator. Simulation studies confirm the efficacy of the proposed model, and its utility is further demonstrated through edge predictions on several real-world datasets.

Keywords: Hypergraph; Determinantal Point Process; Network.

Back to Sessions List