Bayesian Causal Discovery with Cycles and Latent Confounders

Wei Jin[1], Lang Lang[1], Leah H. Rubin[2], **Yanxun Xu[1]**

[1]*Department of Applied Mathematics and Statistics, Johns Hopkins University*

[2]*Department of Neurology, School of Medicine, Johns Hopkins University*

ABSTRACT

Learning causality from observational data has received increasing interest across various scientific fields. However, most existing methods assume the absence of latent confounders and restrict the underlying causal graph to be acyclic, assumptions that are often violated in many real-world applications. In this paper, we address these challenges by proposing a novel framework for causal discovery that accommodates both cycles and latent confounders. By leveraging the identifiability results from noisy independent component analysis and recent advances in factor analysis, we establish the unique causal identifiability of the proposed method under mild conditions. We further develop a fully Bayesian approach for causal structure learning and evaluate its identifiability, utility, and superior performance against state-of-the-art alternatives through extensive simulation studies. Application to a dataset from the Women's Interagency HIV Study yields interpretable and clinically meaningful insights. To facilitate broader applications, we have implemented the proposed Bayesian causal discovery method in an R package, \pkg{BayCausal}, which is the first publicly available software capable of achieving unique causal identification in the presence of both cycles and latent confounders.

Keywords: Bayesian structural learning; Causal identification; Directed cyclic graph; Latent confounding; Observational data

Bayesian Automated Learning of Sparsity in Risk Prediction with Application to Whole-brain Functional Connectivity Analysis

Taiwo Fagbohungbe, Eric Odoom, **Xia Wang**, Xuan Cao

Division of Statistics and Data Science, Department of Mathematical Sciences, University of Cincinnati, USA

ABSTRACT

Challenges in disease risk models lie in identifying key biomarkers and estimating their associated coefficients. Many existing approaches rely on prespecified tuning parameters that implicitly control the number of relevant variables, the so-called sparsity level. We propose a fully Bayesian framework based on spike-and-slab priors for logistic risk prediction that automatically learns sparsity from the data. Specifically, we place hierarchical priors on the variable inclusion probability and on the slab variance (shrinkage) parameter. Learning both parameters enables automated adaptation to the true model sparsity, yielding flexible yet accurate coefficient estimates and predictions. Under mild conditions, the method attains model selection consistency with computational complexity comparable to existing approaches. We analyze an Autism dataset for risk stratification using whole-brain functional connectivity features, training on the NYU site and evaluating on an independent UCLA site. The proposed model achieves 81% classification accuracy, outperforming competing methods. We further evaluate its performance on a mouse gene expression dataset, attaining 75% accuracy and surpassing strong baselines.

Keywords: fMRI, Hierarchical model, Logistic regression, Risk prediction, Spike-and-slab

Bayesian Design and Analysis Methods for Decentralized Clinical Trials

Ruitao Lin

The University of Texas MD Anderson Cancer Center

ABSTRACT

Decentralized clinical trials (DCTs) extend trial activities beyond traditional sites, improving access, convenience, efficiency, and result generalizability. They are particularly promising for chronic conditions like diabetes and obesity, which require longer study durations to assess drug effects. However, decentralized data collection raises concerns about increased variability and potential biases. In this talk, I will present several novel Bayesian methodologies developed at MD Anderson Cancer Center in collaboration with researchers from major pharmaceutical companies to address the design and analysis of decentralized clinical trials with longitudinal data. In particular, I will discuss a novel Bayesian integrated learning procedure that combines centralized and decentralized data collection for analyzing longitudinal data in DCTs. Through simulations and sensitivity analyses, we demonstrate that the proposed Bayesian integrated learning method performs well across various scenarios. Notably, it matches the efficiency of traditional trials when decentralized data collection introduces no additional variability or error. Even when such issues arise, it remains less biased and more efficient than naïve methods that rely solely on centralized data or indiscriminately pool data from both sources.

Keywords: Bayesian methods, clinical trials, decentralization, measurement errors.

Bayesian Dose-Response Meta-Analysis for Predictive Biomarkers Using Aggregate and Individual Participant Data with Data Augmentation for Precision Medicine

J. Jack Lee and Wayne Yi-Hung Wu

Department of Biostatistics, The University of Texas MD Anderson Cancer Center, USA

ABSTRACT

A predictive biomarker is a biological indicator that identifies individuals who are more likely than those without the biomarker to experience either a positive or negative effect from a medical product. It plays a pivotal role in precision medicine. For example, high PD-L1 expression is associated with high efficacy for immune checkpoint inhibitors. In reported studies, some may have individual participant data (IPD) and many others have aggregate data (AD) but may have varying biomarker cutoff values from studies to studies.

To synthesize predictive biomarker evidence from studies with different data types and cutoff values, we construct a Bayesian hierarchical dose–response meta-analysis for time-to-event outcomes that integrates both IPD and AD through a four-parameter log-logistic model. For AD-only settings, a data-augmentation method using quasi–Monte Carlo integration is proposed to average over the biomarker distribution within each interval. Simulation studies demonstrate that incorporating IPD and prognostic factor adjustment substantially reduced bias and mean-squared error, while the augmented AD model improved slope estimation, especially under steep nonlinear relationships. In addition, we introduce a covariate-adjusted meta-analytic combined (cMAC) framework for augmenting the control arm of a current randomized trial by borrowing from historical controls available as IPD, AD, or both. The model links aggregate-level likelihoods to individual-level models via quasi–Monte Carlo integration, enabling coherent borrowing while accounting for covariate imbalance. Simulation studies across varying heterogeneity and covariate-shift scenarios show that hybrid IPD + AD borrowing improves precision and power relative to no borrowing, while automatically downweighting incompatible data to control bias.

Keywords: Bayesian hierarchical model, log-logistic dose–response model, meta-analysis, data augmentation, covariate-adjusted meta-analytic combined framework