BrainGeneBot: A GPT-Engineered, User-Driven Genetic Data Exploration with Polygenic Risk Scores Ranking in Alzheimer's Disease

Zhongming Zhao, PhD

McWilliams School of Biomedical Informatics, The University of Texas Health Science Center at Houston

ABSTRACT

Polygenic risk scores (PRS) are widely used to assess genetic susceptibility in Alzheimer's Disease (AD) research. However, the rapid expansion of PRS studies has led to dataset-specific biases leading to inconsistent variant prioritization and limit generalizability and reproducibility. To address these challenges, we propose a transductive learning framework that integrates multiple PRS datasets for more robust risk variant prioritization, incorporating Genome-Wide Association Study (GWAS) priority scores as biologically informed priors. Additionally, we introduce BrainGeneBot, an AI-driven tool leveraging Generative Pre-trained Transformers (GPT) with Retrieval-Augmented Generation (RAG) technology to streamline genomic analyses in AD, including the STRING for protein interaction analysis, Enrichr for gene set enrichment, ClinVar for genetic variant interpretation, and Biopython for conducting literature searches. We apply our approach to publicly available AD datasets from the PGS Catalog and conduct further analyses to validate its efficacy. In parallel, we perform conventional unsupervised rank aggregation as a baseline. The transductive learning approach not only verifies high-risk variants identified by traditional methods but also reveals unique insights that better correlate with GWAS signals. Our framework streamlines data retrieval and interpretation, effectively prioritizing genetic variants in multiple PRS studies. In summary, the implementation of BrainGeneBot is set to transform genomic research for brain diseases by improving data accessibility, accelerating discovery processes, and refining the precision of genetic insights.

Keywords: GPT-powered informatics, polygenic score, rank aggregation, transductive learning

High-Dimensional Markov-Switching Ordinary Differential Processes

Katherine Tsai[1,3*], Mladen Kolar[2], Sanmi Koyejo[3]

[1]*Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign*

[2]*Marshall School of Business, University of Southern California*

[3]*Department of Computer Science, Stanford University*

ABSTRACT

We investigate the parameter recovery of Markov-switching ordinary differential processes from discrete observations, where the differential equations are nonlinear additive models. This framework has been widely applied in biological systems, control systems, and other domains; however, limited research has been conducted on reconstructing the generating processes from observations. In contrast, many physical systems, such as human brains, cannot be directly experimented upon and rely on observations to infer the underlying systems. To address this gap, this manuscript presents a comprehensive study of the model, encompassing algorithm design, optimization guarantees, and quantification of statistical errors. Specifically, we develop a two-stage algorithm that first recovers the continuous sample path from discrete samples and then estimates the parameters of the processes. We provide novel theoretical insights into the statistical error and linear convergence guarantee when the processes are beta-mixing. Our analysis is based on the truncation of the latent posterior processes and demonstrates that the truncated processes approximate the true processes under mixing conditions. We apply this model to investigate the differences in resting-state brain networks between the ADHD group and normal controls, revealing differences in the transition rate matrices of the two groups.

Keywords: Ordinary Differential Processes; Regime Switchings; Latent Models

Controlling False Discover Rate for High Dimensional Mediator Selection in Non-linear Models

Runqiu Wang[1], **Ran Dai[1]**, Jieqiong Wang[2], Kah Meng Soh[1], Ziyang Xu[2], Mohamed Azzam[2], Hongying Dai[1], Cheng Zheng[1]

[1]*Department of Biostatistics, University of Nebraska Medical Center, 984375 Nebraska Medical Center, 68198, Nebraska, U.S.A.*

[2]*Department of Neurological Sciences, University of Nebraska Medical Center, 984375 Nebraska Medical Center, 68198, Nebraska, U.S.A.*

ABSTRACT

There is a challenge in selecting high-dimensional mediators when the mediators have complex correlation structures and interactions. In this work, we frame the high-dimensional mediator selection problem into a series of hypothesis tests with composite nulls, and develop a method to control the false discovery rate (FDR) which has mild assumptions on the mediation model. We show the theoretical guarantee that the proposed method and algorithm achieve FDR control. We present extensive simulation results to demonstrate the power and finite sample performance compared with existing methods. Lastly, we demonstrate the method for analyzing the Alzheimer's Disease Neuroimaging Initiative (ADNI) data, in which the proposed method selects the volume of the hippocampus and amygdala, as well as some other important MRI-derived measures as mediators for the relationship between gender and dementia progression.

Keywords: False Discovery Rate; high-dimensional mediators; imaging data; knockoff

When Few Labeled Target Data Suffice: A Theory of Semi-Supervised Domain Adaptation via Fine-Tuning from Multiple Adaptive Starts

Wooseok Ha[1], Yuansi Chen[2]

[1]*Department of Mathematical Sciences, KAIST*

[2]*Department of Mathematics, ETH Zu¨rich*

ABSTRACT

Semi-supervised domain adaptation (SSDA) aims to achieve high predictive performance in the target domain with limited labeled target data by exploiting abundant source and unlabeled target data. Despite its significance in numerous applications, theory on the effectiveness of SSDA remains largely unexplored, particularly in scenarios involving various types of source-target distributional shifts. In this talk, I will present a theoretical framework based on structural causal models (SCMs) which allows us to analyze and quantify the performance of SSDA methods when labeled target data is limited. Within this framework, I introduce three SSDA methods, each having a fine-tuning strategy tailored to a distinct assumption about the source and target relationship. Under each assumption, I demonstrate how extending an unsupervised domain adaptation (UDA) method to SSDA can achieve minimax-optimal target performance with limited target labels. Finally, when the relationship between source and target data is only vaguely known—a common practical concern—I will describe the Multi Adaptive-Start FineTuning (MASFT) algorithm, which fine-tunes UDA models from multiple starting points and selects the best-performing one based on a small hold-out target validation dataset. Combined with model selection guarantees, MASFT achieves near-optimal target predictive performance across a broad range of types of distributional shifts while significantly reducing the need for labeled target data.

Keywords: Semi-supervised domain adaptation; distribution shifts; fine-tuning; invariance