A Perturbation Subsampling for Large Scale Data

Yujing Yao[1] and **Zhezhen Jin[2]**

[1]*Department of Neurology*

[2]*Department of Biostatistics, Columbia University, New York, NY*

ABSTRACT

When analyzing large-scale data, subsampling methods and divide-and-conquer procedures are appealing, because they ease the computational burden, while preserving the validity of inferences. Here, sampling may occur with or without replacement. In this paper, we propose a perturbation subsampling approach based on independent and identically distributed stochastic weights for analyzing large-scale data. We justify the method based on optimizing convex objective functions by establishing the asymptotic consistency and normality of the resulting estimators. This method simultaneously provides consistent point and variance estimators. We demonstrate the finite-sample performance of the proposed method using simulation studies and two real-data analyses.

Keywords: Convex objective function, distributed computing, optimization, perturbation, subsampling.

Subgroup Mixture Challenges in Bridging Studies for Predictive Biomarkers

Szu-Yu Tang

Pfizer, Inc.

ABSTRACT

Predictive biomarkers play a critical role in precision medicine by identifying patient subgroups most likely to benefit from specific treatments. In scenarios where a companion diagnostic (CDx) is unavailable, patients are enrolled using a clinical trial assay (CTA) and subsequently retested with the CDx. This requires a bridging study to evaluate the clinical utility of the CDx, particularly its efficacy.

Recent insights from the Oncology Working Group (Liu, 2023) highlight potential logical inconsistencies arising from improper mixing of biomarker-positive and -negative subgroups in different efficacy endpoints. This research investigates the impact of subgroup mixture in the bridging study context. Through illustrative examples and simulation studies of different predictive and prognostic biomarker scenarios, this presentation will:

1. Examine the logical pitfalls associated with subgroup mixture in bridging studies.
2. Apply the Subgroup Mixture Estimation (SME) framework (Ding, 2016) to derive logically consistent estimates.
3. Examine the influence of assay concordance on subgroup mixture challenge.

These findings aim to enhance the robustness of CDx evaluation and support more reliable decision-making in precision oncology.

Keywords: bridging study, clinical trial assay (CTA), companion diagnostic test (CDx), subgroup mixture estimation (SME)

Scalable Joint Modeling of Multiple Longitudinal Biomarkers and Competing Risks Time-to-Event Data: with Applications to Mega-Scale Health Research

Shanpeng Li*[1,3], Emily Ouyang*[2], Jin Zhou[3], **Xinping Cui[2]**, Gang Li[3]

[1]*Department of Computational and Quantitative Medicine, City of Hope, Duarte, CA, USA.*

[2]*Department of Biostatistics, University of California at Riverside, Riverside, CA, USA.*

[3]*Department of Biostatistics, University of California at Los Angeles, Los Angeles, CA, USA.*

ABSTRACT

Joint modeling has become increasingly popular for characterizing the association between one or more longitudinal biomarkers and competing risks time-to-event outcomes. However, semiparametric multivariate joint modeling for large-scale data encounter substantial statistical and computational challenges, primarily due to the high dimensionality of random effects and the complexity of estimating nonparametric baseline hazards. These challenges often lead to prolonged computation time and excessive memory usage, limiting the utility of joint modeling for biobank-scale datasets. In this work, we introduce an efficient implementation of a semiparametric multivariate joint model, supported by a normal approximation and customized linear scan algorithms within an expectation-maximization (EM) framework. Our method significantly reduces computation time and memory consumption, enabling the analysis of data from thousands of subjects. The scalability and estimation accuracy of our approach are demonstrated through simulation studies. We also present an application to UK Biobank primary care study as an illustrative example. A user-friendly R package, FastJM, has been developed for the shared random effects joint model with efficient implementation. The package is publicly available on the Comprehensive R Archive Network https://CRAN.R-project.org/package=FastJM.

Keywords: longitudinal data, competing risks, normal approximation, linear scan algorithms, large-scale biobank data

Metaheuristics as a General-Purpose Optimization Tool for Statistical Research

Weng Kee Wong

Department of Biostatistics, Fielding School of Public Health

University of California at Los Angeles

ABSTRACT

Nature-metaheuristics have been widely used in engineering and computer science to address various types of optimization problems for decades and are now increasingly used across disciplines. They are increasingly popular in industry and academia for tackling all kinds of complex and high-dimensional optimization problems. Interestingly, metaheuristics seems to be still relatively underused in the statistical research community.

I present an overview of nature-inspired metaheuristics and some of their applications in statistics. The main appealing features of these algorithms are their speed, flexibility, availability of codes in different platforms, and ease of implementation and usage. Above all, they are virtually assumptions-free, which allows us to apply them to solve a huge range of optimization tasks. I will enumerate the advantages of nature-inspired metaheuristic algorithms over existing optimization algorithms and illustrate their diverse applications in biostatistics, and beyond. If time permits, I will demonstrate how nature-inspired algorithms can find more flexible and computationally challenging designs for early-phase clinical trials.

Keywords: Design Efficiency, Early Phase Trials, Optimal Experiment Designs, Particle Swarm Optimization. mixture regression models, multiple constraints, swarm intelligence