Statistica Sinica **32** (2022), 1-22
doi:https://doi.org/10.5705/ss.202019.0445

OPTIMAL ESTIMATION OF SIMULTANEOUS SIGNALS USING ABSOLUTE INNER PRODUCT WITH APPLICATIONS TO INTEGRATIVE GENOMICS

Rong Ma, T. Tony Cai and Hongzhe Li

University of Pennsylvania

Abstract: Integrating the summary statistics from a genome-wide association study and expression quantitative trait loci data provides a powerful way of identifying genes with expression levels that are potentially associated with complex diseases. We introduce a parameter called T-score that quantifies the genetic overlap between a gene and the disease phenotype based on the summary statistics, based on the mean values of two Gaussian sequences. Specifically, given two independent samples $\mathbf{x}_n \sim N(\theta, \mathbf{\Sigma}_1)$ and $\mathbf{y}_n \sim N(\mu, \mathbf{\Sigma}_2)$, the T-score is defined as $\sum_{i=1}^n |\theta_i \mu_i|$, a nonsmooth functional, that characterizes the number of shared signals between two absolute normal mean vectors $|\theta|$ and $|\mu|$. Using approximation theory, estimators are constructed and shown to be minimax rate-optimal and adaptive over various parameter spaces. Simulation studies demonstrate the superiority of the proposed estimators over existing methods. Lastly, the method is applied to an integrative analysis of heart failure genomics data sets and we identify several genes and biological pathways that are potentially causal to human heart failure.

Key words and phrases: Approximation theory, eQTL, GWAS, minimax lower bound, non-smooth functional.

1. Introduction

1.1. Integrating summary data from genome-wide association studies and expression quantitative trait loci studies

Integrative genomics aims to integrate various biological data sets for the systematic discovery of a genetic basis that underlies and modifies a human disease (Giallourakis et al. (2005)). To realize its full potential in genomic research, methods are needed that exhibit both computational efficiency and a theoretical guarantee for such integrative analyses. This study proposes a method that combines data sets from genome-wide association studies (GWASs) and expression quantitative trait loci (eQTL) studies in order to identify genetically regulated disease genes. Furthermore, we provide an integrative view of the underlying bi-

Corresponding author: Hongzhe Li, Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, PA 19104, USA. E-mail: hongzhe@upenn.edu.