Statistica Sinica **36** (2026), 1-23
doi:https://doi.org/10.5705/ss.202023.0201

AN EMPIRICAL BAYES REGRESSION FOR MULTI-TISSUE GENE EXPRESSION PREDICTION

Fei Xue and Hongzhe Li*

Purdue University and University of Pennsylvania

Abstract: The Genotype-Tissue Expression (GTEx) project collects samples from multiple human tissues to study the relationship between genetic variation or single nucleotide polymorphisms (SNPs) and gene expression in each tissue. However, most existing eQTL analyses only focus on single tissue information. In this paper, we develop a multi-tissue method that improves prediction of gene expression based on cis-SNPs by borrowing information across tissues. Specifically, we propose an empirical Bayes regression model for SNP-expression association using data from multiple tissues. To allow the effects of SNPs to vary greatly among tissues, we use a mixture distribution as the prior, which is a mixture of a multivariate Gaussian distribution and a Dirac mass at zero. We show that the proposed estimator of the cis-SNP effects on gene expression asymptotically achieves the minimum Bayes risk among all estimators. Analyses of the GTEx data show that our proposed method is superior to existing methods in terms of prediction accuracy for gene expression using cis-SNPs in testing sets.

Key words and phrases: Bayes risk, data integration, missing data, mixture model.

1. Introduction

Genome-wide association studies (GWAS) have successfully associated single nucleotide polymorphisms (SNPs) with complex human traits (Uffelmann et al., 2021). However, there are still problems in statistical power and interpretation of GWAS results due to complexity of linkage disequilibrium (LD) and gene regulation (Boyle, Li and Pritchard, 2017). To alleviate these problems, a popular approach is the transcriptome-wide association study (TWAS) that integrates the SNP-trait association with SNP-based prediction of gene expression (Wainberg et al., 2019). Specifically, TWAS first predicts expression levels using SNPs, and then tests whether the predicted values are associated with human traits. In this paper, we focus on the first part in TWAS and aim to improve the SNP-based prediction of gene expression.

Many large data sets have been generated for such genetics of gene expression studies for various tissues, which have provided important insights into gene regulations. Among these studies, the Genotype-Tissue Expression (GTEx) project aims to characterize variation in gene expression levels across individuals

*Corresponding author. E-mail: hongzhe@upenn.edu