Bayesian large-scale regression with GWAS summary statistics

Xiang Zhu\(^1\) and Matthew Stephens\(^1,2\)

\(^1\)Department of Statistics, \(^2\)Department of Human Genetics

How can summary statistics be used in multiple-SNP analysis?

- Recent work has revealed potential merits of multiple-SNP analysis.
- Existing methods are often complicated by access to full data.
- Summary statistics from single-SNP analysis are widely available.

A novel statistical problem

Consider the multiple linear regression,

\[ y = X\beta + \epsilon \]

where \( y \) is an \( n \times 1 \) vector, \( X \) is an \( n \times p \) matrix, \( \beta \) is the \( p \times 1 \) regression coefficient, and \( \epsilon \) is the error term. In regression analysis, we observe the individual-level data \( \{X, y\} \) and use them to infer the parameter \( \beta \). Here we assume that the full data \( \{X, y\} \) are not available, and only summary statistics of simple linear regression are provided:

\[ \hat{\beta}_j := (X_j^T X_j)^{-1} X_j^T y_j, \quad \hat{\epsilon}_j := (X_j^T X_j)^{-1}(y_j - X_j\hat{\beta}_j) \]

where \( X_j \) is the \( j \)th column of \( X \). \( j = 1, \ldots, p \).

How do we infer \( \beta \) using \( \{\hat{\beta}_j, \hat{\epsilon}_j\} \)?

Examples of tools for multiple-SNP analysis

A growing number of GWAS summary statistics-based methods have recently been published.

- QCTA-1002 [1]: approximate the standard multiple linear regression
- MVAN [2]: model 2-scores at a locus as multivariate normal
- LDSC [3]: regresses genome-wide \( \chi^2 \) statistics on "LD scores"

Shortcomings of existing methods

- Their connections with methods using full data are not clear.
- They cannot be easily applied to various multiple-SNP problems.

These concerns can be addressed if \( \beta \) has an explicit likelihood based on summary-level data.

References


Acknowledgments

We thank Xian He, Yongtao Guan, Peter Carbonetto for helpful discussions. We thank Raman Shah and John Zekos for expert technical support.

This study makes use of data generated by the Wellcome Trust Case Control Consortium. A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk.

We acknowledge the University of Chicago Research Computing Center for support of this work.

Regression with Summary Statistics (RSS) provides a solution.

Likelihood

We derive the following regression model for GWAS summary statistics:

\[ y_j = X_j \beta + \epsilon_j, \quad j = 1, \ldots, p \]

where \( \hat{\beta}_j \) is the single-SNP effect size estimate of SNP \( j \); \( S = diag(\hat{s}_j) \), \( s = (\hat{s}_1, \ldots, \hat{s}_p) \), where \( \hat{s}_j \) is the standard error of \( \hat{\beta}_j \); \( R \) is the population linkage disequilibrium (LD) matrix.

We term the model Regression with Summary Statistics.

Features of RSS model

- It produces an explicit likelihood of multiple-SNP effect \( \beta \).
- It is mathematically justified by asymptotic theory [4].
- It is computationally tractable for genome-wide analysis.
- It answers multiple questions within a single framework.

Dual role of population LD

- \( \hat{\beta}_j \) includes the effects of all SNPs that SNP \( j \) tags.
- \( \hat{\beta}_j \) and \( \hat{\beta}_k \) are correlated if SNP \( j \) and SNP \( k \) are in LD.

We estimate \( R \) using a shrinking model based on population genetic principles [6].

Prior

Four types of prior on \( \beta \) are considered:

- Linear mixed model (LMM) prior:
  \[ \hat{\beta}_j \sim N(0, \sigma_j^2) \]
- Bayesian variable selection regression (BVSR) prior:
  \[ \hat{\beta}_j \sim pN(0, \sigma_j^2) + (1 - p)N(0, \tilde{\sigma}_j^2) \]
- Bayesian sparse linear mixed model (BLSMM) prior:
  \[ \hat{\beta}_j \sim N(0, \sigma_j^2 + \tilde{\sigma}_j^2) + (1 - \tilde{p})N(0, \tilde{\sigma}_j^2) \]
- Adaptive shrinkage (ASH) prior:
  \[ \hat{\beta}_j \sim \pi_j N(0, \sigma_j^2) + \ldots + \pi_p N(0, \sigma_p^2) \]

They depict three genetic architectures. 

- infinitesimal (LMM), sparse (BVSR), hybrid (BLSMM & ASH)

Posterior

We provide efficient MCMC schemes to simulate posterior distributions of \( \beta \). Multiple tasks can be performed simultaneously using the same posterior samples.

Extension

One important extension is to integrate additional genomic information with the RSS model [6, 7, 8, 9]. For example, together with the prior from [6],

\[ \hat{\beta}_j \sim (1 - \pi_j) N(0, \sigma_j^2) + \pi_j N(0, \tilde{\sigma}_j^2) \]

logit(\( \pi_j \)) = \( \beta \cdot S \) (SNP \( j \) is in the gene set)

RSS is able to infer gene set enrichment. Details will be presented at [10].

RSS on height GWAS supports a polygenic architecture of human stature.

We applied the RSS model on GWAS summary statistics of 1.06 million SNPs for adult human height from 253,288 individuals of European (EUR) ancestry [11].

The population LD matrix \( R \) was estimated from the 1000 Genomes [12] EUR samples.

RSS yields results comparable to methods that require full data.

Estimating SNP heritability

Phenotypic variation explained (PVE) by available genotypes:

\[ \text{PVE}(j) = \frac{R_{jj}}{R_{jj} + S_{jj}} \]

Full-data counterpart: GEMMA-BVSR and GEMMA-BLSMM [14, 15, 16]

Testing SNP set association

Multiple-SNP Bayes factor (BF) of SNP set \( C \) under LMM prior:

\[ \text{BF}(C) = p(R_{ij} | R_{ij}, \sigma = 0) / p(R_{ij} | R_{ij}, \sigma \neq 0) \]

Full-data counterpart: BIMAR [17, 18]

Detecting genome-wide association

Posterior inclusion probability (PIP) of SNP \( j \) under BVSR prior:

\[ \text{PIP}(j) = P(\hat{\beta}_j \neq 0 | \beta, R) \]

Full-data counterpart: GEMMA-BVSR [14, 15]

Software

Software of fitting the RSS model is freely available from https://github.com/etatselab/rss.

More information at xiangzhutuchicago.edu