THE UNIVERSITY OF CHICAGO

BAYESIAN VARIABLE SELECTION FROM SUMMARY DATA, WITH APPLICATION TO JOINT FINE-MAPPING OF MULTIPLE TRAITS

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF STATISTICS

BY
YUXIN ZOU

CHICAGO, ILLINOIS
AUGUST 2021

In memory of

My Grandfather

Wenbin Wang (1939-2021)

Table of Contents

LIST OF FIGURES vii

LIST OF TABLES xii

ACKNOWLEDGMENTS xiii

ABSTRACT xiv

1 INTRODUCTION 1

2 *SUSIE-SUFF*: FINE-MAPPING USING SUFFICIENT STATISTICS . . . 7
2.1 The *SuSiE* model 7
2.1.1 The Single Effect Regression model 7
2.1.2 The Sum of Single Effects Regression model 10
2.2 *SuSiE-suff*: *SuSiE* using sufficient statistics 12

3 *SUSIE-RSS*: FINE-MAPPING USING SUMMARY STATISTICS 16
3.1 The *SuSiE-RSS* model 17
3.1.1 Likelihood with invertible LD matrix 20
3.1.2 The Single Effect Regression model using Summary Statistics 21
3.1.3 The Sum of Single Effects Regression model using Summary Statistics 22
3.2 Likelihood with non-invertible LD matrix 24
3.3 LD matrix 28
3.4 Detecting allele flips 30
3.5 Algorithm refinement 32
3.6 Numerical Comparisons 36
3.6.1 Posterior inclusion probabilities 37
3.6.2 Credible sets 42
3.7 Discussion 46

4 MULTI-TRAIT FINE-MAPPING 49
4.1 The *mvSuSiE* model 51
4.1.1 The Multivariate Single Effect Regression model 52
4.1.2 The Multivariate Sum of Single Effects Regression model . . . 58
4.2 *mvSuSiE-suff*: *mvSuSiE* using sufficient statistics 58
4.3 The *mvSuSiE-RSS* model 61

4.3.1 The Multivariate Single Effect Regression using Summary Statistics . . . 63
4.3.2 The Multivariate Sum of Single Effects Regression model using Summary Statistics . . . 64
4.3.3 Estimating the residual correlation matrix . . . 65
4.3.4 Generating prior covariance matrices . . . 66
4.4 Posterior inference . . . 66
4.5 Numerical Experiments . . . 68
4.5.1 Posterior Inclusion Probability . . . 77
4.5.2 Credible Sets . . . 81
4.5.3 Runtimes . . . 84
4.6 Fine-mapping on UK Biobank Blood Cell traits . . . 89
4.6.1 UK Biobank Data . . . 89
4.6.2 Multivariate fine-mapping . . . 91
4.6.3 Comparison with single-trait fine-mapping . . . 95
4.6.4 Examples . . . 96
4.7 Discussion . . . 104
4.8 Author Contribution . . . 108

5 ENHANCEMENTS FOR MASH . . . 109
5.1 A review of the `mash` model . . . 110
5.2 Estimating the residual correlation matrix . . . 112
5.2.1 Exact update in M step . . . 114
5.2.2 Ad hoc update in M step . . . 115
5.2.3 Numerical Comparisons . . . 116
5.2.4 Discussion . . . 120
5.3 `mash commonbaseline`: Comparing multiple conditions with the same reference level . . . 120
5.3.1 `mash commonbaseline` with a common control condition . . . 122
5.3.2 `mash commonbaseline` without a common control condition . . . 124
5.3.3 Simple Simulation with a control condition . . . 126
5.3.4 Simple Simulation without a control condition . . . 129
5.3.5 Deviation from Median . . . 133
5.3.6 Application . . . 136
5.3.7 Discussion . . . 142

REFERENCES . . . 143

A SUPPLEMENTARY FOR *SUSIE-SUFF* 155
A.1 *SuSiE* using sufficient statistics 155

B SUPPLEMENTARY FOR *SUSIE-RSS* 157
B.1 Modifying LD matrix with z scores 157

C SUPPLEMENTARY FOR *MVSUSIE-RSS* 160
C.1 Details of posterior computations for the *BMR* model with a mixture prior 160
C.1.1 Bayesian simple multivariate regression with an intercept . . . 160
C.1.2 Bayesian simple multivariate regression with missing data . . . 164
C.1.3 Bayesian simple multivariate regression with intercept and missing data 166
C.2 Details of Multivariate single-effect regression with a mixture prior . . 169

List of Figures

1.1 **Fine-mapping example using *SuSiE*.** Two out of the 1,002 variants have non-zero effects (red points, labeled "SNP 1" and "SNP 2" in the left-hand panel). The strongest marginal association (SMA) is a non-effect SNP (yellow point, labeled "SMA" in the left-hand panel). *SuSiE* finds two 95% CSs, each containing a true effect variant. 5

3.1 **Fine-mapping example with allele mismatch between GWAS and reference panel.** Results are from a simulated data set with $p = 1,002$ SNPs with one SNP having a true effect (red) and one SNP having allele mismatched between study and reference panel (yellow). **Top left:** z scores for each SNP; **Top right:** PIPs computed by *SuSiE-RSS* using the mismatched summary data, with two *CSs* highlighted in blue and green; *SuSiE-RSS* incorrectly identifies a *CS* containing the mismatched (yellow) SNP. **Bottom-left:** Diagnostic plot plotting each observed z score against its expected value. The mismatched SNP stands out as a potential outlier. **Bottom right:** PIPs computed by *SuSiE-RSS* with correct allele alignment, with *CS* highlighted in blue; *SuSiE-RSS* finds one *CS*, which contains the true effect SNP, and the false positive *CS* is avoided. 33

3.2 **Fine-mapping example to illustrate that IBSS algorithm with additional refinement steps can deal with a challenging case.** Results are from a simulated data set with $p = 1,001$ SNPs. Two out of the 1,001 SNPs have non-zero effects (red points, labeled "SNP 1" and "SNP 2" in the left-hand panel). The strongest marginal association (SMA) is a non-effect SNP (yellow point, labeled "SMA" in the left-hand panel). The IBSS algorithm with default settings (middle panel), identifies three *CSs*, two of them false positives containing no true effect SNPs (one contains the SMA). The refinement procedure finds two 95% *CSs*, each containing a true effect SNP. 36

3.3 **Comparison of *SuSiE-suff* with and without the refinement procedure.** The plot summarizes Power versus FDR using the PIPs from *SuSiE-suff*. The open circles in the highlight power versus FDR at PIP threshold of 0.95. 39

3.4 **Comparison of Power vs FDR for each method with in-sample LD matrix.** The plot shows how Power and FDR co-vary as PIP threshold changes. Circles indicate results at PIP threshold 0.95. 40

3.5 **Comparison of Power and FDR for each method with different LD matrix.** Each curve shows how Power and FDR co-vary as PIP threshold changes. We used LD matrices estimated from reference panels of different sizes (500 and 1,000 samples) and with different regularization parameter s. Circles indicate results at PIP threshold 0.95. 43
3.6 **Comparing estimated $\hat{s}$ using different LD matrix.** 44
3.7 **Comparison of 95% credible sets from *SuSiE-suff*, *SuSiE-RSS* and FINEMAP.** Panels show coverage, power, median size and median purity. These statistics are computed by pooling all CSs from all data sets. The error bars in coverage and power plots show 2× standard error. 45

4.1 **Simulated structure for the "artificial mixture" in 20 traits for Scenario 1.** Each heatmap represents a covariance matrix $\boldsymbol{U}_k$, w_k gives the relative frequency of $\boldsymbol{U}_k$. The simulated signal has 20% chance to be shared in block ($\boldsymbol{U}_1$), 15% chance to be specific in trait 1 ($\boldsymbol{U}_2$), 30% chance to be shared in 2 traits ($\boldsymbol{U}_3$ and $\boldsymbol{U}_4$ with equal weights), 25% chance to be shared across traits with different heterogeneity ($\boldsymbol{U}_5$ - $\boldsymbol{U}_9$ with equal weights). 71
4.1 **Simulated structure for the "artificial mixture" in 20 traits for Scenario 1 (cont.).** The simulated signal has 10% chance to be specific in trait other than the first one ($\boldsymbol{U}_{10}$ - $\boldsymbol{U}_{19}$ with equal weights). 72
4.2 **Simulated structure for the "UK Biobank Blood Cells" in 16 traits for Scenario 2.** Each heatmap represents a covariance matrix $\boldsymbol{U}_k$, w_k gives the relative frequency of $\boldsymbol{U}_k$. 73
4.3 **Estimated "artificial mixture" prior for 20 traits via ED in `mashr` package.** Each heatmap represents a covariance matrix $\boldsymbol{U}_k$, w_k gives the relative frequency of $\boldsymbol{U}_k$. 75
4.4 **Estimated "UK Biobank Blood Cell traits" prior via ED in `mashr` package.** Each heatmap represents a covariance matrix $\boldsymbol{U}_k$, w_k gives the relative frequency of $\boldsymbol{U}_k$. 76
4.5 **Power versus FDR in Scenario 3.** *mvSuSiE-RSS* was fitted using default prior and residual correlation matrix estimated from z scores. . . 78
4.6 **Comparison of Power and FDR for different method in Scenario 1 and 2.** 80

4.7 **Calibration of PIP for *mvSuSiE-suff* and *mvSuSiE-RSS* in Scenario 1.** The plots show the proportion of effect SNPs versus the mean PIP for each bin. We expect all points are aligned in the diagonal line for a well-calibrated method. The gray error bars show ±2 standard errors. Points below the diagonal line imply the corresponding PIPs are anti-conservative and points above the diagonal line imply the PIPs are conservative. 82
4.8 **Calibration of PIP for *mvSuSiE-suff* and *mvSuSiE-RSS* in Scenario 2.** 83
4.9 **Compare 95% Credible Sets (*CSs*) from *mvSuSiE-suff* and *mvSuSiE-RSS*.** The coverage, power, size and purity are computed using all *CSs* in all data sets. The error bar is computed as 2 × standard error. The panel (a) and (b) are from simulation Scenario 1. 85
4.9 **Compare 95% Credible Sets (*CSs*) from *mvSuSiE-suff* and *mvSuSiE-RSS* (cont.).** The panel (c) and (d) are from simulation Scenario 2. . . 86
4.10 **Compare 95% trait-specific Credible Sets (*CSs*) from *SuSiE-suff*, *SuSiE-RSS*, *mvSuSiE-suff* and *mvSuSiE-RSS*.** The coverage, power, size and purity are computed using all *CSs* in all traits in all data sets. The error bar is computed as 2 × standard error. The panel (a) and (b) are from simulation Scenario 1. 87
4.10 **Compare 95% trait-specific Credible Sets (*CSs*) from *SuSiE-suff*, *SuSiE-RSS*, *mvSuSiE-suff* and *mvSuSiE-RSS* (cont.).** The panel (c) and (d) are from simulation Scenario 2. 88
4.11 **Estimated UK Biobank 16 Blood Cell traits prior via ED in `mashr` package.** Each heatmap represents a covariance matrix $\boldsymbol{U}_k$, w_k gives the relative frequency of $\boldsymbol{U}_k$. Traits are color-coded by cell types. . 92
4.12 **Summary of patterns identified by ED in `mashr` package.** For each covariance matrix $\boldsymbol{U}_k$, the figure shows the heatmap of $\boldsymbol{U}_k$, and bar plots of the top eigenvectors of $\boldsymbol{U}_k$. Traits are color-coded by cell types. Component (a) reflects effects sharing among red cells. Component (b) captures correlations among compound white cells and platelet. 93
4.12 **Summary of patterns identified by ED in `mashr` package (cont.).** Component (c) captures platelet effects. Component (b) captures compound white cells effects. 94
4.13 **Number of traits share a *CS* from *mvSuSiE-RSS*.** The histogram shows number of traits in which the *CS* are significant in. 96

4.14 **Pairwise sharing of significant *CS* among blood cell traits.** For each pair of blood traits, we consider the *CSs* that are significant ($lfsr < 0.01$) in at least one of the two blood traits, and plot the proportion of these that are shared. 97
4.15 **Comparison of number of significant *CSs* in *mvSuSiE-RSS* vs number of *CSs* in *SuSiE-RSS*.** The traits are colored coded by cell types. The dashed line represents $y = x$. 98
4.16 **Cross-trait Posterior Inclusion Probability from *mvSuSiE-RSS* for AK3 region.** The *CSs* are color coded. The *CS* 1 contains rs12005199 (chr9: 4763491), which is very close to rs409950 (chr9: 4763368) in *CS* 2. The *CS* 3 contains 20 SNPs with minimum pairwise correlation 0.968. . . 101
4.17 **Observed z scores for SNPs in the identified *CSs* for AK3 region.** The locations for variants are color coded by *CSs*. The color of bubble represents effect size and the size of bubble represents $-\log_{10}$(p value). . 102
4.18 **Posterior effects for SNPs in the identified *CSs* for AK3 region.** The locations for variants are color coded by *CSs*. The color of bubble represents posterior effect size and the size of bubble represents $-\log_{10}(CS\ lfsr)$. For each trait, we plot bubbles with significant *CSs*. . . 102
4.19 **Associations for haemoglobin and reticulocyte percentage from original GWAS for AK3 region.** The left panel shows the identified *CSs* using *SuSiE-RSS*. The right panel shows the identified *CSs* using *mvSuSiE-RSS*. 103
4.20 **Cross-trait Posterior Inclusion Probability from *mvSuSiE-RSS* for GLIS3 region.** The *CSs* are color coded. The *CS* 1 contains rs6415788 (chr9: 4118111). The *CS* 2 contains rs7033677 (chr9: 4049942). The *CS* 3 contains 8 SNPs with minimum pairwise correlation 0.903. . . 105
4.21 **Observed z scores for SNPs in the identified *CSs* for GLIS3 region.** 106
4.22 **Posterior effects for SNPs in the identified *CSs* for GLIS3 region.** 106

5.1 **Comparison of log-likelihood ratio using $\hat{\boldsymbol{C}}$ from different methods.** The likelihood ratio compares the model with the one using simple (5.2.1) method. 117
5.2 **Power and Accuracy with different estimated $\boldsymbol{C}$.** The left plot compares RRMSE using different $\boldsymbol{C}$. The right plot gives ROC curves for detecting significant genes. 119

5.3 **Power and Accuracy for different methods.** The deviations are computed over the common control group. The left plot compares RRMSE from different models. The right plot gives ROC curves for detecting significant genes. 129
5.4 **Power and Accuracy for different methods.** The deviations are computed over the mean. The left plot compares RRMSE from different models. The right plot gives ROC curves for detecting significant genes. 133
5.5 **Simulation with Deviation and compare with median.** (a) shows the accuracy of the estimated deviations; (b) shows the ROC curve. . . 136
5.6 **Inferred Patterns of Sharing using `Cormotif`.** The plot shows the five patterns identified in Blischak et al. (2015). 137
5.7 **Inferred Patterns of Sharing.** The plot (a) shows the most common pattern of sharing identified in `mash commonbaseline`. The corresponding first three eigenvectors are in (c), (d) and (e). The plot (b) shows another pattern we identified in `mash commonbaseline` model. 138
5.8 **Plots for similarity of deviations by sign and magnitude.** (a) shows the proportion of significant genes that are "shared in sign"; (b) shows the proportion of significant genes that are "shared in magnitude". 139
5.9 `mash commonbaseline` **examples.** `mash commonbaseline` uses learned patterns of sharing to capture more subtle patterns. For each subfigure, the dots in the left plot are raw deviations for each infection with bars indicating ± 2 se. The dots in the right plot are `mash commonbaseline` posterior estimated deviations with bars indicating ± 2 se. 141

B.1 **Evaluation of posterior inclusion probabilities (PIPs) using LD from reference panel with correction from z scores.** 159

List of Tables

3.1 **Runtimes in seconds** 41

4.1 **Runtimes in seconds with 20 traits.** *mvSuSiE-suff* and *mvSuSiE-RSS* were fitted using default priors. The residual covariance matrix in *mvSuSiE-suff* is the empirical covariance matrix of phenotypes. The residual covariance matrix in *mvSuSiE-RSS* is the empirical correlation matrix of from z scores (4.3.12). 84
4.2 **UK Biobank blood cell traits.** 90
4.3 **Summary of single-trait fine-mapping results from *SuSiE-RSS*.** 99

5.1 **Frobenius norm between estimated $\hat{C}$ and the true value.** 118
5.2 **Runtimes in seconds.** The running time for the exact EM updates and ad hoc EM updates 119

ACKNOWLEDGMENTS

First and foremost, I would like to express my deepest gratitude to my adviser, Professor Matthew Stephens, for the continuous support for my Ph.D. study and research, for his superb guidance and limitless patience. I have benefited greatly from all invaluable insights and constructive feedback that he has shared over the years. I am truly honored to work with such an extraordinary statistician and geneticist. Without his guidance and persistent encouragement, this dissertation would not have been possible.

I also want to express my gratitude to other members of my dissertation committee, Dan Nicolae and Mary Sara McPeek, for their inspiring suggestions and insightful discussions about the projects.

I would like to thank other faculty members in the Departments of Statistics and Human Genetics: Yali Amit, Peter McCullagh, Mihai Anitescu, Xin He, Mei Wang, Fei Liu and others, for their generous help in my Ph.D. study.

I would also like to thank my friends and all the past and current peers in the Department of Statistics, the Stephens Lab, and He Lab: Gao Wang, Peter Carbonetto, Kevin Luo, Fabio Morgante, Yifan Zhou, Alan Selewa, Yusha Liu, Yunqi Yang, Yanyu Liang, Jason Willwerscheid, Joyce Hsiao, Wei Wang, Lei Sun, Zhonglin Li, Ruoqi Yu, Tae Hyun Kim, Micol Tresoldi and many others, for their care and support.

Last but not least, I would like to thank my parents, Jinkui Wang and Jian Zou, for their unconditional love and support throughout my life.

ABSTRACT

Bayesian methods provide attractive approaches to select relevant variables in multiple regression models, particularly in settings with very highly correlated variables. For example, they are popular in genetic fine-mapping problems, aiming to identify the genetic variants that causally affect some phenotypes of interest. However, Bayesian methods are limited by the computational speed and the interpretability of the posterior distribution. Wang et al. (2020) presented a simple and computationally scalable approach to variable selection, the "Sum of Single Effects" (*SuSiE*) model, which provides a Credible Set for each selection, making the results easy to interpret. The *SuSiE* model requires access to individual genotypes and phenotypes.

In this dissertation, we provide a method to fit the *SuSiE* model using summary statistics from univariate regression results. To improve the accuracy and power for variable selection, we further generalize the *SuSiE* framework to select variables jointly for multiple outcomes and account for complicated effect size heterogeneity among outcomes. We provide multivariate variable selection methods using individual-level data, sufficient statistics, and summary statistics. We illustrate the power and flexibility of our method using realistic numerical simulations and real data applications.

CHAPTER 1

INTRODUCTION

Genome-wide association studies (GWAS) have successfully identified many genomic regions containing risk variants associated with complex diseases and traits (e.g. Hakonarson et al., 2007; Zeggini et al., 2007; Donnelly, 2008; Visscher et al., 2012; Köttgen et al., 2013; Fritsche et al., 2016). However, the majority of the GWAS reported associated variants are not causally affecting the trait of interest, but rather correlated to the true causal variants through linkage disequilibrium (LD) (Ott, 1999). To gain more precise biological understanding of these associations, researchers have turned to fine-mapping methods (Spain and Barrett, 2015; Schaid et al., 2018) to identify putative causal variants contributing to these diseases and traits.

Genetic fine-mapping is often framed as a variable selection problem. Suppose $\boldsymbol{y}$ is a centered quantitative phenotype vector for N individuals, $\boldsymbol{X} = [\boldsymbol{x}_1 \cdots \boldsymbol{x}_J]$ is an $N \times J$ column centered genotype matrix for N individuals at J genetic variants (typically Single Nucleotide Polymorphisms, or SNPs) in a genomic contiguous region. Consider the multiple linear regression model:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{e}, \tag{1.0.1}$$

where $\boldsymbol{b}$ is the $J \times 1$ vector of multiple regression coefficients, $\boldsymbol{e}$ is an N-vector of error terms with distribution $N_N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_N)$, $\sigma^2 > 0$ is the residual variance, $\boldsymbol{I}_N$ is the $N \times N$ identity matrix, and $N_r(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the r-variate normal distribution

with mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$. We assume the non-zero effect variants are included in the genotype matrix (either directly typed or well imputed). Assuming the effects $\boldsymbol{b}$ are sparse and the absence of unmeasured confounders, performing variable selection on $\boldsymbol{b}$ in regression model (1.0.1) identifies variants that causally affect the phenotype $\boldsymbol{y}$. In genetic fine-mapping, the variants can be very highly correlated and it is challenging to identify the true causal variants from many other highly correlated nearby variants.

A simple variable selection approach is stepwise conditional analysis (Friedman et al., 2001, Section 3.3.3; Yang et al. 2012). It starts by identifying the variant most correlated with the response. It then iterates, at each step, finding the next most correlated variant conditional on the selected variants in previous steps. This is continued until no significant variant remains in the associated region at a nominal or Bonferroni-corrected significance level. This simple stepwise conditional approach is widely used in fine-mapping applications (e.g. Allen et al., 2010; Trynka et al., 2011; Flister et al., 2013; Astle et al., 2016). It provides a list of variants that could be causal. However, when two variants are perfectly correlated and one of the variants is the causal variant, it is impossible to distinguish which one should be selected since both are correlated with the phenotype. Stepwise conditional analysis arbitrarily selects one of the two, which loses information and leads to an incomplete list of possible causal variants. Moreover, stepwise conditional analysis does not provide any assessment of confidence in selected variants. There are many variable selection methods suffering similar issues, for example, methods using penalized likelihood (Tibshirani, 1996; Zou and Hastie, 2005; Tibshirani, 2011) and selective inference

(Taylor and Tibshirani, 2015; Lee et al., 2016; Berk et al., 2013). To overcome these issues, the majority of current approaches to fine-mapping are based on Bayesian variable selection methods (Mitchell and Beauchamp, 1988; George and McCulloch, 1997).

Bayesian variable selection in regression (BVSR) has been widely applied to genetic fine-mapping and related applications (e.g. Meuwissen et al., 2001; Sillanpaa and Bhattacharjee, 2005; Servin and Stephens, 2007; Hoggart et al., 2008; Stephens and Balding, 2009; Logsdon et al., 2010; Guan and Stephens, 2011; Bottolo et al., 2011; Carbonetto et al., 2012; Zhou et al., 2013; Kichaev et al., 2014; Hormozdiari et al., 2014; Chen et al., 2015; Wallace et al., 2015; Moser et al., 2015; Benner et al., 2016; Newcombe et al., 2016; Wen et al., 2016b; Lee et al., 2018; Wang et al., 2020). BVSR quantifies uncertainty in the causal variants by taking into account patterns of correlations among variants. It assigns a prior distribution to $\boldsymbol{b}$ that induces sparsity, and computes the posterior probability that variant j has non-zero effect in the model (this is called posterior inclusion probability, or PIP).

Although the Bayesian approach is attractive to variable selection problem with highly correlated variables, computing the posterior distribution is computationally challenging. Some approaches exhaustive search all possible models involving a small number of causal variants (Kichaev et al., 2014; Hormozdiari et al., 2014; Chen et al., 2015), but this quickly becomes computationally infeasible as the number of possible causal variants increases in the model. Other approaches such as sophisticated Markov Chain Monte Carlo or stochastic search algorithms can help (Guan and Stephens, 2011; Wallace et al., 2015; Benner et al., 2016; Wen et al., 2016b; New-

combe et al., 2016), but remain computationally intensive. Another challenge for BVSR is to summarize and interpret the complex posterior distribution, especially with highly correlated variants and many causal variants.

Motivated by the shortcomings above, Wang et al. (2020) introduced a novel formulation of BVSR, Sum of Single Effects (*SuSiE*) model. The novel structure of *SuSiE* suggests a simple and fast variational algorithm, Iterative Bayesian Stepwise Selection (IBSS), whose computation scales linearly with the number of possible causal variants. The *SuSiE* structure also naturally yields independent credible sets. Each Credible Set is designed to capture one non-zero effect with high probability, making the posterior results easy to interpret and ideal for guiding follow-up studies. The definition of "Credible Set" is as follows,

Definition 1. *A level-ρ Credible Set (CS) is defined to be a subset of variants that has probability $\geqslant \rho$ of containing at least one effect variant (i.e. a variant with non-zero effect).*

The *SuSiE* model reports as many *CSs* as the data support, each *CS* contains as few variants as possible. In the case where two variants (say, $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$) are completely correlated and one of them is causal, *SuSiE* will report a *CS* containing both variants. This indicates that there is (at least) one causal variant, which is either $\boldsymbol{x}_1$ or $\boldsymbol{x}_2$ but we cannot say which.

In Figure 1.1, we illustrate *SuSiE* output in a more realistic simulated fine-mapping example from Wang et al. (2020). There are two variants with non-zero effects. Because of the correlation structure between variants, the strongest marginal association (SMA) is not one of the causal variants. Stepwise conditional analysis

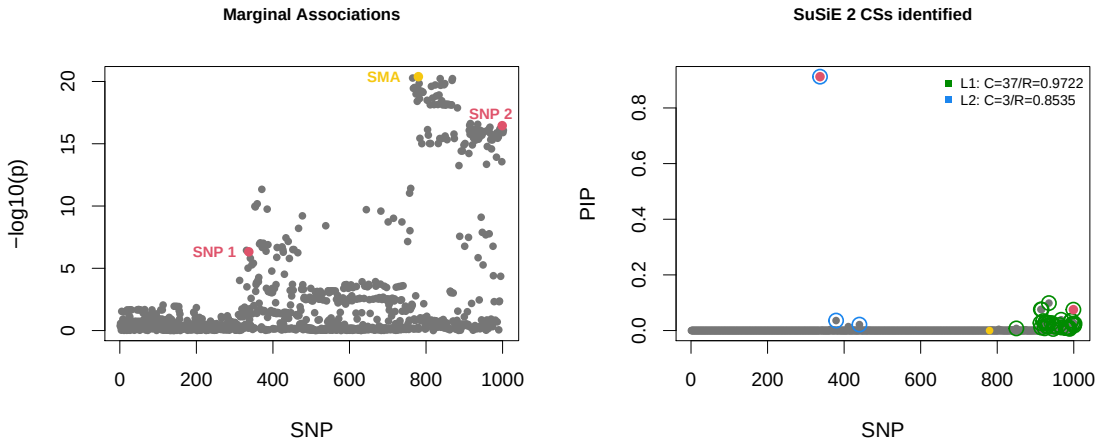


Figure 1.1: **Fine-mapping example using *SuSiE*.** Two out of the 1,002 variants have non-zero effects (red points, labeled "SNP 1" and "SNP 2" in the left-hand panel). The strongest marginal association (SMA) is a non-effect SNP (yellow point, labeled "SMA" in the left-hand panel). *SuSiE* finds two 95% CSs, each containing a true effect variant.

would select the wrong variant (SMA) at the first step. *SuSiE* reports two high-purity 95% *CSs*, indicating there are (at least) two causal variants. Each *CS* contains a true effect variant, and neither contains the SMA. The blue *CS* contains 3 variants and one of them is the actual causal variant. The other green *CS* contains 37 variants with minimum correlation 0.97. Because these 37 variants are strongly correlated, we cannot say which one has non-zero effect.

SuSiE requires individual-level genotype and phenotype data, $\{\boldsymbol{y}, \boldsymbol{X}\}$. In this thesis, we generalize the *SuSiE* framework to do fine-mapping using summary statistics and multi-trait fine-mapping. The dissertation consists of the following chapters. From Chapter 2 to Chapter 4, we extend the *SuSiE* framework in several aspects. In Chapter 2, we extend *SuSiE* to use sufficient statistics, which gives the same re-

sult as *SuSiE* but with different computation complexity. In Chapter 3, we modify *SuSiE* model to use summary statistics from GWAS and an accurate correlation matrix (also known as LD matrix). Some caveats are further discussed. In Chapter 4, we generalize *SuSiE* framework to do joint fine-mapping of multiple phenotypes using summary statistics. Joint analysis of multiple phenotypes improves power and precision to identify relevant variants (e.g. Shriner, 2012; Stephens, 2013).

In addition to the work on fine-mapping, in Chapter 5, we introduce two enhancements for `mash` (Multivariate Adaptive Shrinkage), a flexible empirical Bayes multivariate association testing method (Urbut et al., 2019). The two enhancements are all about properly including the error correlations among measurements in different conditions in the `mash` model.

CHAPTER 2

SUSIE-SUFF: FINE-MAPPING USING SUFFICIENT STATISTICS

In this chapter, we summarize the "Sum of Single Effects" (*SuSiE*) model introduced by Wang et al. (2020), and the algorithm they introduced to fit this model, which they called Iterative Bayesian Stepwise Selection (IBSS). We also describe a new, equivalent algorithm for fitting the *SuSiE* model using sufficient statistics. Compared with the original algorithm, our new algorithm can be computationally advantageous when the sample size is large (e.g. for biobank-scale data, UK Biobank (Sudlow et al., 2015)), because after initial computation of the sufficient statistics, the computations at each iteration do not depend on sample size. The ideas in this chapter also form the building blocks for subsequent extensions to use (non-sufficient) summary statistics in Chapter 3.

2.1 The *SuSiE* model

The *SuSiE* model introduced in Wang et al. (2020) is based on an even simpler model, the "single effect regression" (SER) model, so we begin by describing the SER model before describing the *SuSiE* model.

2.1.1 The Single Effect Regression model

The SER is a multiple linear regression (1.0.1) in which *exactly one* of the regression coefficients is non-zero. This simple idea of assuming exactly one effect dates back

at least to Servin and Stephens (2007), and, despite its simplicity, variations on this model have been used in many genomic applications related to fine-mapping (e.g Veyrieras et al., 2008; Pickrell, 2014; Maller et al., 2012).

Specifically, Wang et al. (2020) considered the following SER:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{e} \tag{2.1.1}$$

$$\boldsymbol{e} \sim N(0, \sigma^2 \boldsymbol{I}_N) \tag{2.1.2}$$

$$\boldsymbol{b} = \boldsymbol{\gamma} b \tag{2.1.3}$$

$$\boldsymbol{\gamma} \sim \text{Mult}(1, \boldsymbol{\pi}) \tag{2.1.4}$$

$$b \sim N(0, \sigma_0^2). \tag{2.1.5}$$

Here $\text{Mult}(m, \boldsymbol{\pi})$ denotes the multinomial distribution obtained when m samples are drawn with category probabilities given by $\boldsymbol{\pi}$. Thus $\boldsymbol{\gamma} \in \{0, 1\}^J$ is a J-vector with exactly one non-zero element, and so $\boldsymbol{b}$ is also a vector with one non-zero element. (We use the term "single effect vector" to refer to any vector with one non-zero element, so both $\boldsymbol{\gamma}$ and $\boldsymbol{b}$ are single effect vectors.) The scalar b represents the value of the one non-zero element in $\boldsymbol{b}$. The prior inclusion probabilities, $\boldsymbol{\pi} = (\pi_1, \cdots, \pi_J)$, which we assume fixed and known, give the prior probability for each variant j being "the" non-zero effect. The prior variance of the single effect, σ_0^2, and the residual variance, σ^2, are hyper-parameters that can be pre-specified or (more commonly) estimated.

The likelihood for σ_0^2 and σ^2 under the *SER* model is

$$L_{SER}(\sigma_0^2, \sigma^2; \boldsymbol{y}) \coloneqq p(\boldsymbol{y}|\boldsymbol{X}, \sigma_0^2, \sigma^2) \tag{2.1.6}$$

$$= p(\boldsymbol{y}|\boldsymbol{X}, \sigma^2, \boldsymbol{b} = 0) \sum_j \pi_j \mathrm{BF}(\boldsymbol{y}, \boldsymbol{x}_j; \sigma^2, \sigma_0^2), \tag{2.1.7}$$

where $\mathrm{BF}(\boldsymbol{y}, \boldsymbol{x}_j; \sigma^2, \sigma_0^2)$ denotes the Bayes Factor for variant j being associated with $\boldsymbol{y}$ (Servin and Stephens, 2007), which is given in (2.1.13) below. Maximum likelihood estimates of the hyper-parameters can be obtained by maximizing this likelihood using numerical optimization methods.

Given the hyper-parameters, the posterior distribution for $\boldsymbol{b}$ is easy to compute analytically (Servin and Stephens, 2007; Wang et al., 2020), and is summarized in Proposition 1.

Proposition 1. *Consider the SER model with known σ_0^2 and σ^2. The posterior distribution on $\boldsymbol{b}$ can be written in terms of univariate least-squares estimate of b_j, $\hat{b}_j \coloneqq \boldsymbol{x}_j^\mathsf{T}\boldsymbol{y}/\boldsymbol{x}_j^\mathsf{T}\boldsymbol{x}_j$, and its variance $s_j^2 \coloneqq \sigma^2/\boldsymbol{x}_j^\mathsf{T}\boldsymbol{x}_j$. Specifically, the posterior distribution on $\boldsymbol{b} = \boldsymbol{\gamma} b$ is*

$$\boldsymbol{\gamma}|\boldsymbol{y}, \boldsymbol{X}, \sigma^2, \sigma_0^2 \sim \mathit{Mult}(1, \boldsymbol{\alpha}) \tag{2.1.8}$$

$$b|\boldsymbol{y}, \boldsymbol{X}, \sigma^2, \sigma_0^2, \gamma_j = 1 \sim N(\mu_{1j}, \sigma_{1j}^2), \tag{2.1.9}$$

where

$$\sigma_{1j}^2 \coloneqq \frac{1}{1/s_j^2 + 1/\sigma_0^2} \tag{2.1.10}$$

$$\mu_{1j} \coloneqq \frac{\sigma_{1j}^2}{s_j^2}\hat{b}_j \tag{2.1.11}$$

$$\alpha_j = \frac{\pi_j BF(\boldsymbol{y}, \boldsymbol{x}_j; \sigma^2, \sigma_0^2)}{\sum_{j'=1}^{J} \pi_{j'} BF(\boldsymbol{y}, \boldsymbol{x}_{j'}; \sigma^2, \sigma_0^2)} \tag{2.1.12}$$

$$BF(\boldsymbol{y}, \boldsymbol{x}_j; \sigma^2, \sigma_0^2) = \sqrt{\frac{s_j^2}{s_j^2 + \sigma_0^2}} \exp\left(\frac{1}{2}\frac{\hat{b}_j^2}{s_j^2}\frac{\sigma_0^2}{\sigma_0^2 + s_j^2}\right). \tag{2.1.13}$$

The vector $\boldsymbol{\alpha}$ gives posterior inclusion probabilities for the variants. Given $\boldsymbol{\alpha}$, it is also straightforward to compute a level-ρ *CS*, which is a set of variants that has probability at least ρ of containing the causal variant (Maller et al., 2012). Specifically, *CS* is obtained by first sorting variants in decreasing order of α_j, and then including variants in the CS until their cumulative probability exceeds ρ. To formalize the *CS*, let $r = (r_1, \cdots, r_J)$ denote the indices of the variants ranked in order of decreasing α_j, so that $\alpha_{r_1} \geqslant \cdots \geqslant \alpha_{r_J}$, and let S_k denote the cumulative sum of the k largest PIPs. The level-ρ *CS* is

$$CS(\boldsymbol{\alpha}; \rho) \coloneqq \{r_1, \ldots, r_{k_0}\}, \quad k_0 = \min\{k : S_k \geqslant \rho\}. \tag{2.1.14}$$

2.1.2 The Sum of Single Effects Regression model

While the SER model is attractive in its simplicity, the assumption of a single non-zero effect is rather limiting. To address this, while preserving some of the compu-

tational benefits of the SER model, Wang et al. (2020) introduced the *SuSiE* model, which parameterizes the sparse vector $\boldsymbol{b}$ as a sum of single effect vectors:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{e} \tag{2.1.15}$$

$$\boldsymbol{e} \sim N(0, \sigma^2 \boldsymbol{I}_N) \tag{2.1.16}$$

$$\boldsymbol{b} = \sum_{l=1}^{L} \boldsymbol{b}_l \tag{2.1.17}$$

$$\boldsymbol{b}_l = \boldsymbol{\gamma}_l b_l \tag{2.1.18}$$

$$\boldsymbol{\gamma}_l \sim \text{Mult}(1, \boldsymbol{\pi}) \tag{2.1.19}$$

$$b_l \sim N(0, \sigma_{0l}^2). \tag{2.1.20}$$

In this formulation L, which must be specified, is an upper bound on the number of non-zero entries in $\boldsymbol{b}$. Note that the model allows each single effect to have its own variance, σ_{0l}^2; if $\sigma_{0l}^2 = 0$ then that effect disappears. In the special case $L = 1$, the *SuSiE* model becomes the SER model.

Wang et al. (2020) developed a simple algorithm for fitting the *SuSiE* model, which they called Iterative Bayesian Stepwise Selection (IBSS). The idea behind IBSS is that, given $\boldsymbol{b}_1, \cdots, \boldsymbol{b}_{L-1}$, fitting $\boldsymbol{b}_L$ corresponds to fitting an *SER* model, which is straightforward as described above. Thus the IBSS algorithm simply iterates this procedure: it repeatedly computes the posterior distribution for $\boldsymbol{b}_l$ under the *SER* model given current estimates of other $\boldsymbol{b}_{l'}$, $l' \neq l$. The theoretical background for the IBSS algorithm is established in details in Appendix B of Wang et al. (2020). In short, the IBSS algorithm optimizes a variational approximation (VA) to the posterior dis-

tribution for $\boldsymbol{b}_1, \cdots, \boldsymbol{b}_L$ under the *SuSiE* model. The algorithm finds an approximation $q(\boldsymbol{b}_1, \cdots, \boldsymbol{b}_L)$ to the posterior distribution $p_{\text{post}} = p(\boldsymbol{b}_1, \cdots, \boldsymbol{b}_L | \boldsymbol{X}, \boldsymbol{y}, \sigma^2, \boldsymbol{\sigma}_0^{\boldsymbol{2}})$ by minimizing the Kullback-Leibler (KL) divergence from q to p_{post}. In *SuSiE*, the approximation q is chosen to factorize into L independent components, each corresponding to a *SER* model:

$$q(\boldsymbol{b}_1, \ldots, \boldsymbol{b}_L) = \prod_{l=1}^{L} q_l(\boldsymbol{b}_l). \tag{2.1.21}$$

The L single effects are assumed independent a posteriori. This allows q_l to capture the strong dependencies among elements of $\boldsymbol{b}_l$ that are induced by the assumption that exactly one element of $\boldsymbol{b}_l$ is non-zero.

The hyper-parameters $\boldsymbol{\sigma}_0^{\boldsymbol{2}} = (\sigma_{01}^2, \cdots, \sigma_{0L}^2)$ and σ^2 are estimated using an empirical Bayes approach. The prior variance for the l-th single effect is estimated by maximum-likelihood before computing the posterior distribution for $\boldsymbol{b}_l$. The residual variance σ^2 is estimated by maximizing the expected likelihood under VA.

SuSiE summarizes the approximated posterior distribution $q(\boldsymbol{b}_1, \cdots, \boldsymbol{b}_L)$ using posterior inclusion probabilities (PIPs) and independent *CSs*.

2.2 *SuSiE-suff*: *SuSiE* using sufficient statistics

The IBSS algorithm in Wang et al. (2020) takes as input the data $\boldsymbol{y}$ and $\boldsymbol{X}$. In this section we describe a new, but equivalent, IBSS algorithm that instead takes as input the sufficient statistics, which are given in the following Lemma.

Lemma 1. *Given residual variance σ^2, the sufficient statistics of the multiple linear*

regression (1.0.1) *are* $\boldsymbol{X}^\intercal\boldsymbol{X}$ *and* $\boldsymbol{X}^\intercal\boldsymbol{y}$. *To estimate the residual variance* σ^2, *the sufficient statistics are* $\boldsymbol{X}^\intercal\boldsymbol{X}, \boldsymbol{X}^\intercal\boldsymbol{y}, \boldsymbol{y}^\intercal\boldsymbol{y}$ *and sample size* N.

Proof. The likelihood for $\boldsymbol{b}$ and σ^2 is

$$L(\boldsymbol{b}, \sigma^2; \boldsymbol{y}, \boldsymbol{X}) \coloneqq p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{b}, \sigma^2) \tag{2.2.1}$$

$$= (2\pi\sigma^2)^{-N/2} \exp\left(-\frac{1}{2\sigma^2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\|^2\right) \tag{2.2.2}$$

$$= (2\pi\sigma^2)^{-N/2} \exp\left(-\frac{1}{2\sigma^2}\left(\boldsymbol{y}^\intercal\boldsymbol{y} - 2\boldsymbol{b}^\intercal\boldsymbol{X}^\intercal\boldsymbol{y} + \boldsymbol{b}^\intercal\boldsymbol{X}^\intercal\boldsymbol{X}\boldsymbol{b}\right)\right). \tag{2.2.3}$$

Given the residual variance σ^2, it is clear the sufficient statistics for parameters $\boldsymbol{b}$ are $\boldsymbol{X}^\intercal\boldsymbol{X}$ and $\boldsymbol{X}^\intercal\boldsymbol{y}$. The sufficient statistics for parameters $\boldsymbol{b}$ and σ^2 are $\boldsymbol{X}^\intercal\boldsymbol{X}, \boldsymbol{X}^\intercal\boldsymbol{y}, \boldsymbol{y}^\intercal\boldsymbol{y}$ and sample size N. □

To develop an IBSS algorithm that uses these sufficient statistics, we compute the posterior distribution for $\boldsymbol{b}$ under the *SER* model using sufficient statistics. We define a function, *SER-suff*, that takes input as sufficient statistics and returns the posterior distribution for $\boldsymbol{b} = \boldsymbol{\gamma}b$ under the *SER* model with fixed hyper-parameters σ^2, σ_0^2. We write it as

$$\textit{SER-suff}(\boldsymbol{X}^\intercal\boldsymbol{X}, \boldsymbol{X}^\intercal\boldsymbol{y}; \sigma^2, \sigma_0^2) \coloneqq (\boldsymbol{\alpha}, \boldsymbol{\mu}_1, \boldsymbol{\sigma}_1^{\boldsymbol{2}}), \tag{2.2.4}$$

where $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_J)$ is the vector of PIPs under *SER*, with $\alpha_j \coloneqq \Pr(\gamma_j = 1 \,|\boldsymbol{X}^\intercal\boldsymbol{X}, \boldsymbol{X}^\intercal\boldsymbol{y}, \sigma^2, \sigma_0^2)$, and μ_{1j}, σ_{1j}^2 are the posterior mean and variance of b given $\gamma_j = 1$. For later convenience, we use $L(\boldsymbol{b}, \sigma^2; \boldsymbol{X}^\intercal\boldsymbol{X}, \boldsymbol{X}^\intercal\boldsymbol{y}, \boldsymbol{y}^\intercal\boldsymbol{y}, N)$ and $L_{SER}(\sigma_0^2, \sigma^2; \boldsymbol{X}^\intercal\boldsymbol{y})$ to denote the likelihood for $\boldsymbol{b}$ and σ^2 (2.2.3) and the *SER* likeli-

hood for σ_0^2 and σ^2 (2.1.7) using sufficient statistics.

The IBSS algorithm using sufficient statistics is outlined in Algorithm 1, and we call the *SuSiE* model fitting with the IBSS-suff algorithm as *SuSiE-suff*. When the sufficient statistics are correctly computed using the column centered individual-level data $\{\boldsymbol{y}, \boldsymbol{X}\}$, the result from *SuSiE-suff* is same as applying the original IBSS algorithm in Wang et al. (2020) to the original $\{\boldsymbol{y}, \boldsymbol{X}\}$, which for brevity we call *SuSiE*. However, the computational complexity of the *SuSiE-suff* and *SuSiE* algorithms differ: the computational complexity of *SuSiE* is $O(NJL)$ per iteration, whereas *SuSiE-suff* is $O(J^2L)$ per iteration (and the number of iterations will be the same). If $N > J$, which can be the case in fine-mapping applications, then *SuSiE-suff* will be faster. However, computing the matrix $\boldsymbol{X}^\intercal\boldsymbol{X}$ — which is required for *SuSiE-suff* but not for the original algorithm — is expensive, requiring $O(NJ^2)$ operations. Thus in practice *SuSiE-suff* will only be preferred when $N \gg J$ or when $\boldsymbol{X}^\intercal\boldsymbol{X}$ has been pre-computed.

Note that one can compute the sufficient statistics using the following commonly-computed summary statistics from J simple linear regressions: $\hat{b}_j = \boldsymbol{x}_j^\intercal\boldsymbol{y}/\boldsymbol{x}_j^\intercal\boldsymbol{x}_j$ with standard error, $\hat{s}_j$, the sample LD matrix, the variance of $\boldsymbol{y}$ and sample size N. See Appendix A.1 for details.

We have implemented Algorithm 1 in the R package `susieR` available at `https://github.com/stephenslab/susieR`.

Algorithm 1 IBSS algorithm using sufficient statistics

Require: Sufficient Statistics $\boldsymbol{X}^\intercal\boldsymbol{X}, \boldsymbol{X}^\intercal\boldsymbol{y}, \boldsymbol{y}^\intercal\boldsymbol{y}$ and sample size N.
Require: Number of effects, L.
Require: A function *SER-suff*$(\boldsymbol{X}^\intercal\boldsymbol{X}, \boldsymbol{X}^\intercal\boldsymbol{y}; \sigma^2, \sigma_0^2) \to (\boldsymbol{\alpha}, \boldsymbol{\mu}_1, \boldsymbol{\sigma}_1^2)$ that computes the posterior distribution for $\boldsymbol{b}$ under the *SER* model.

Initial settings of σ^2, $\boldsymbol{\sigma_0^2}$ and posterior means $\bar{\boldsymbol{b}}_l = 0$, for $l = 1, \ldots, L$.
repeat
  for l in $1, \ldots, L$ **do**
    $\boldsymbol{u} \leftarrow \boldsymbol{X}^\intercal\boldsymbol{y} - \boldsymbol{X}^\intercal\boldsymbol{X}\sum_{l' \neq l}\bar{\boldsymbol{b}}_{l'}$
    $\sigma_{0l}^2 \leftarrow$ argmax $L_{SER}(\sigma_{0l}^2, \sigma^2; \boldsymbol{u})$ ▷ Update σ_{0l}^2 (optional).
    $(\boldsymbol{\alpha}_l, \boldsymbol{\mu}_{1l}, \boldsymbol{\sigma}_{1l}) \leftarrow$ *SER-suff*$(\boldsymbol{X}^\intercal\boldsymbol{X}, \boldsymbol{u}; \sigma^2, \sigma_{0l}^2)$
    $\bar{\boldsymbol{b}}_l \leftarrow \boldsymbol{\alpha}_l \circ \boldsymbol{\mu}_{1l}$ ▷ "$\circ$" denotes element-wise multiplication.
  $\sigma^2 \leftarrow$ argmax $\mathbb{E}_q\left[\log L(\boldsymbol{b}, \sigma^2; \boldsymbol{X}^\intercal\boldsymbol{X}, \boldsymbol{X}^\intercal\boldsymbol{y}, \boldsymbol{y}^\intercal\boldsymbol{y}, N)\right]$ ▷ Update σ^2 (optional).
until convergence criterion satisfied
return $\boldsymbol{\alpha}_1, \boldsymbol{\mu}_{11}, \boldsymbol{\sigma}_{11}^2, \ldots, \boldsymbol{\alpha}_L, \boldsymbol{\mu}_{1L}, \boldsymbol{\sigma}_{1L}^2$.

CHAPTER 3

SUSIE-RSS: FINE-MAPPING USING SUMMARY STATISTICS

Multiple regression models, such as the linear model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{e}$, are widely used in genetic studies to relate a phenotype $\boldsymbol{y}$ to genotypes $\boldsymbol{X}$, and many methods have been developed to fit such models (e.g. Meuwissen et al., 2001; Bottolo et al., 2010, 2011; Guan and Stephens, 2011; Wen et al., 2016b; Wang et al., 2020). These methods naturally require access to the individual-level genotype and phenotype data. However, in many genetic applications individual-level data can be difficult to obtain, both for practical reasons (e.g. the need to obtain many data sets collected by many different groups) and for reasons to do with consent and privacy. In contrast, summary data, such as z scores from single SNP analysis, are much easier to obtain, and many publications share such summary data through the Internet (e.g. Pasaniuc and Price, 2017, Table 1). In addition, information on LD among SNPs is available from public reference genotype panels such as Consortium et al. (2015). The ease of access to such "summary data" has motivated the development of methods to fit the multiple regression model using summary data (e.g. Hormozdiari et al., 2014; Kichaev et al., 2014; Chen et al., 2015; Vilhjálmsson et al., 2015; Benner et al., 2016; Mak et al., 2017; Lloyd-Jones et al., 2019). In this chapter we develop method to fit the *SuSiE* model to such summary data. Because the method combines the "regression with summary statistics" (RSS) likelihood from Zhu and Stephens (2017) with the *SuSiE* model of Wang et al. (2020), we call it *SuSiE-RSS*.

Our basic approach is similar to the approaches taken by previous fine-mapping methods that use summary data (e.g. Hormozdiari et al., 2014; Kichaev et al., 2014; Chen et al., 2015; Benner et al., 2016; Newcombe et al., 2016; Lee et al., 2018). The main advantage of our method over these previous methods is that the *SuSiE* model leads to fast and accurate fine-mapping. However, our analysis also highlights some subtle issues that arise in dealing with non-invertible LD matrices, which have not been addressed in previous work (Section 3.2).

We describe the *SuSiE-RSS* approach in Section 3.1. Section 3.3 shows the importance of having an accurate LD matrix. In Section 3.4, we describe a simple approach to detect variants with flipped allele between summary statistics and reference panel. In Section 3.5, we describe a refinement procedure for the IBSS algorithm. Section 3.6 shows the performance of *SuSiE-RSS* using numerical experiments. We discuss future work in Section 3.7.

3.1 The *SuSiE-RSS* model

Following previous methods for fine-mapping from summary data (e.g. Hormozdiari et al., 2014; Kichaev et al., 2014), we assume that we have access to the following summary data: i) z scores from marginal tests of association between each SNP and the phenotype in a GWAS sample; ii) an estimate $\hat{\boldsymbol{R}}$ of the LD (correlation) matrix among variants. For example, $\hat{\boldsymbol{R}}$ could be the sample correlation matrix computed from the GWAS samples if available, or computed from a suitable reference panel that is similar to the GWAS study population.

Also following previous methods (e.g. Kichaev et al., 2014; Hormozdiari et al.,

2014; Benner et al., 2016), we assume the model

$$\hat{\boldsymbol{z}}|\boldsymbol{z}, \hat{\boldsymbol{R}} \sim N_J(\hat{\boldsymbol{R}}\boldsymbol{z}, \hat{\boldsymbol{R}}), \tag{3.1.1}$$

where $\boldsymbol{z}$ represents an unobserved $J \times 1$ vector of standardized true effects (also known as the noncentrality parameter, or NCP). Intuitively this model captures an important property of marginal z scores: because of LD, the observed marginal z score for variant j, $\hat{z}_j$, does not represent the actual standardized true effect, but a combined effect of all variants in LD with variant j, it is an "LD convolved" effect,

$$\mathbb{E}(\hat{z}_j|\hat{\boldsymbol{R}}) = \sum_{i=1}^{J} r_{ij} z_i, \tag{3.1.2}$$

where r_{ij} is the (i, j)-th entry of $\hat{\boldsymbol{R}}$.

Various, more-or-less formal, justifications have been give for assuming that (3.1.1) will hold approximately, both for z scores from quantitative phenotypes and binary (case-control) data. For quantitative phenotypes, Hormozdiari et al. (2014) argue for (3.1.1) using the usual marginal z scores

$$\hat{z}_j := \hat{b}_j / \hat{s}_j, \tag{3.1.3}$$

where

$$\hat{b}_j := (\boldsymbol{x}_j^\intercal \boldsymbol{x}_j)^{-1} \boldsymbol{x}_j^\intercal \boldsymbol{y}, \tag{3.1.4}$$

$$\hat{s}_j^2 := (N \boldsymbol{x}_j^\intercal \boldsymbol{x}_j)^{-1} (\boldsymbol{y} - \boldsymbol{x}_j \hat{b}_j)^\intercal (\boldsymbol{y} - \boldsymbol{x}_j \hat{b}_j). \tag{3.1.5}$$

(These expressions assume the phenotypes $\boldsymbol{y}$ and genotypes $\boldsymbol{x}_j$ are all centered to have mean 0.) In this case the NCPs are $z_j = \frac{b_j\sqrt{\boldsymbol{x}_j^\intercal \boldsymbol{x}_j}}{\sigma}$, where b_j and σ are defined in (1.0.1).

For case-control data, Han et al. (2009) argue for (3.1.1) when the z scores are computed using a two proportions test using balanced case-control samples. That is,

$$\hat{z}_j := \frac{\sqrt{N}(f_j^+ - f_j^-)}{\sqrt{2f_j(1-f_j)}}, \tag{3.1.6}$$

where f_j^+ and f_j^- are the frequency of the variant j in case and control samples; $f_j = \frac{f_j^+ + f_j^-}{2}$ is the frequency of the variant j in the sample. In this case the NCPs are $z_j = \frac{\sqrt{N}(p_j^+ - p_j^-)}{\sqrt{2p_j(1-p_j)}}$, where p_j^+ and p_j^- are the frequency of the variant j in case and control populations, $p_j = \frac{p_j^+ + p_j^-}{2}$. The two proportions test statistics (3.1.6) is a scaled version of $\hat{\boldsymbol{b}}_j$ (3.1.4), where $\boldsymbol{y}$ is a binary vector with 1, -1 indicating the case and control samples.

In brief, the assumptions underlying the use of (3.1.1) are i) the correlation of response, $\boldsymbol{y}$, with any single variant $\boldsymbol{x}_j$ is small; ii) the LD matrix $\hat{\boldsymbol{R}}$ accurately reflects the correlation of variants in the study samples; iii) the same samples are used to compute all marginal z scores; and iv) genotypes used to compute summary statistics are accurate, since we ignore the imputation error for imputed genotypes in the model. See Zhu and Stephens (2017) for more precise derivations and discussion.

3.1.1 Likelihood with invertible LD matrix

For model (3.1.1) to have a valid density, the LD matrix $\hat{\boldsymbol{R}}$ must be invertible. We assume the LD matrix is invertible in Section 3.1. The non-invertible case is discussed in Section 3.2.

Assuming the LD matrix is invertible, the density for $\hat{\boldsymbol{z}}$ is

$$p(\hat{\boldsymbol{z}}|\boldsymbol{z}, \hat{\boldsymbol{R}}) = (2\pi)^{-\frac{J}{2}} |\hat{\boldsymbol{R}}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\hat{\boldsymbol{z}} - \hat{\boldsymbol{R}}\boldsymbol{z})^\intercal \hat{\boldsymbol{R}}^{-1}(\hat{\boldsymbol{z}} - \hat{\boldsymbol{R}}\boldsymbol{z})\right), \tag{3.1.7}$$

where $|\hat{\boldsymbol{R}}|$ represents the matrix determinant. The density leads to the "Regression with Summary Statistics" (RSS) likelihood for $\boldsymbol{z}$,

$$L_{RSS}(\boldsymbol{z}; \hat{\boldsymbol{z}}, \hat{\boldsymbol{R}}) \coloneqq \exp\left(-\frac{1}{2}\boldsymbol{z}^\intercal \hat{\boldsymbol{R}}\boldsymbol{z} + \boldsymbol{z}^\intercal \hat{\boldsymbol{z}}\right). \tag{3.1.8}$$

Note that the RSS likelihood (3.1.8) is a special case of the likelihood (2.2.3) with $\boldsymbol{X}^\intercal\boldsymbol{X} = \hat{\boldsymbol{R}}$, $\boldsymbol{X}^\intercal\boldsymbol{y} = \hat{\boldsymbol{z}}$, $\sigma^2 = 1$ (for any values of $\boldsymbol{y}^\intercal\boldsymbol{y}$ and N).

One nice property of distribution (3.1.1) shown by Benner et al. (2016) is that the support for the model that a particular set of variants is causal depends only on the marginal z scores of those causal variants, and the LD matrix between the causal variants. We state it in the following proposition and the derivation is in Benner et al. (2016).

Proposition 2. *Let $\mathcal{C}$ and $\mathcal{N}$ be the set of causal and non-causal variants, respectively, that is $\boldsymbol{z}_\mathcal{C} \neq \boldsymbol{0}$, $\boldsymbol{z}_\mathcal{N} = \boldsymbol{0}$. We partition the marginal z-scores $\hat{\boldsymbol{z}}$ into $\hat{\boldsymbol{z}}_\mathcal{C}$ and $\hat{\boldsymbol{z}}_\mathcal{N}$. The conditional distribution of $\hat{\boldsymbol{z}}_\mathcal{N}$ given $\hat{\boldsymbol{z}}_\mathcal{C}$, $\boldsymbol{z}_\mathcal{N} = 0$ does not depend on $\boldsymbol{z}_\mathcal{C}$,*

i.e.

$$p(\hat{\boldsymbol{z}}_{\mathcal{N}}|\hat{\boldsymbol{z}}_{\mathcal{C}}, \boldsymbol{z}_{\mathcal{N}} = \boldsymbol{0}, \boldsymbol{z}_{\mathcal{C}}, \hat{\boldsymbol{R}}) = p(\hat{\boldsymbol{z}}_{\mathcal{N}}|\hat{\boldsymbol{z}}_{\mathcal{C}}, \boldsymbol{z}_{\mathcal{N}} = \boldsymbol{0}, \hat{\boldsymbol{R}}). \tag{3.1.9}$$

The Bayes Factor for assessing the support for the model with causal variants in $\mathcal{C}$ *against the null model is*

$$BF := \frac{p(\hat{\boldsymbol{z}}|\boldsymbol{z}_{\mathcal{N}} = \boldsymbol{0}, \boldsymbol{z}_{\mathcal{C}} \neq \boldsymbol{0}, \hat{\boldsymbol{R}})}{p(\hat{\boldsymbol{z}}|\boldsymbol{z} = \boldsymbol{0}, \hat{\boldsymbol{R}})} \tag{3.1.10}$$

$$= \frac{p(\hat{\boldsymbol{z}}_{\mathcal{C}}|\boldsymbol{z}_{\mathcal{C}} \neq \boldsymbol{0}, \hat{\boldsymbol{R}}_{\mathcal{C}\mathcal{C}})}{p(\hat{\boldsymbol{z}}_{\mathcal{C}}|\boldsymbol{z}_{\mathcal{C}} = \boldsymbol{0}, \hat{\boldsymbol{R}}_{\mathcal{C}\mathcal{C}})}, \tag{3.1.11}$$

where $\hat{\boldsymbol{R}}_{\mathcal{C}\mathcal{C}}$ *is the LD matrix of the causal variants in* $\mathcal{C}$*. The Bayes Factor uses only the information about causal variants in* $\mathcal{C}$*.*

3.1.2 The Single Effect Regression model using Summary Statistics

As with *SuSiE* model using individual-level data (Wang et al., 2020), *SuSiE-RSS* model is based on a simpler model, the "single effect regression" model using summary statistics (*SER-RSS*). The *SER-RSS* model assumes exactly one of the J variants has non-zero effect, that is,

$$\hat{\boldsymbol{z}} \sim N_J(\hat{\boldsymbol{R}}\boldsymbol{z}, \hat{\boldsymbol{R}}) \tag{3.1.12}$$

$$\boldsymbol{z} = \boldsymbol{\gamma} z \tag{3.1.13}$$

$$\boldsymbol{\gamma} \sim \text{Mult}(1, \boldsymbol{\pi}) \tag{3.1.14}$$

$$z \sim N(0, \omega^2). \tag{3.1.15}$$

The single effect vector $\boldsymbol{z}$ has exactly one non-zero element (equals to z). The position of the non-zero effect is determined by the binary single effect vector $\boldsymbol{\gamma}$. The prior variance of the single effect, ω^2, can be pre-specified or estimated by maximizing likelihood.

Under the *SER-RSS* model, the following corollary is an immediate consequence of Proposition 2 with only one variant in the causal set $\mathcal{C}$.

Corollary 1. *Under the SER-RSS model, the posterior inference is independent of the LD matrix* $\hat{\boldsymbol{R}}$.

Given the prior variance of the single effect, the posterior computations to the *SER-RSS* model follow straightforwardly using Proposition 1, in which $\hat{b}_j = \hat{z}_j$, $s_j = 1$. A level-ρ CS, $CS(\boldsymbol{\alpha}; \rho)$, is then computed using the posterior distribution on $\boldsymbol{z}$.

3.1.3 The Sum of Single Effects Regression model using Summary Statistics

Conventional methods for sparse regression give a sparse prior on $\boldsymbol{z}$ that allows for multiple non-zero entries (e.g. Hormozdiari et al., 2014; Kichaev et al., 2014; Benner et al., 2016). However, the posterior distribution becomes intractable and the computation becomes expensive. Here we use the "Sum of Single Effects" prior (Wang et al., 2020) on NCP $\boldsymbol{z}$, which parameterizes the sparse vector $\boldsymbol{z}$ as a sum of

single effect vectors. The *SuSiE-RSS* model is

$$\hat{\boldsymbol{z}} \sim N_J(\hat{\boldsymbol{R}}\boldsymbol{z}, \hat{\boldsymbol{R}}) \tag{3.1.16}$$

$$\boldsymbol{z} = \sum_{l=1}^{L} \boldsymbol{z}_l \tag{3.1.17}$$

$$\boldsymbol{z}_l = \boldsymbol{\gamma}_l z_l \tag{3.1.18}$$

$$\boldsymbol{\gamma}_l \sim \text{Mult}(1, \boldsymbol{\pi}) \tag{3.1.19}$$

$$z_l \sim N(0, \omega_l^2). \tag{3.1.20}$$

There are at most L non-zero effects. Each single effect z_l has its own variance, ω_l^2.

Because the *SuSiE-RSS* model uses the sum of single effects prior and the RSS likelihood (3.1.8) is a special case of the likelihood (2.2.3), we can easily fit the model using the IBSS algorithm outlined in Algorithm 1 with inputs $\boldsymbol{X}^\intercal\boldsymbol{X} = \hat{\boldsymbol{R}}$, $\boldsymbol{X}^\intercal\boldsymbol{y} = \hat{\boldsymbol{z}}$ and fixed residual variance ($\sigma^2 = 1$). Of course this will not give the same result as *SuSiE-suff* applied to the actual sufficient statistics; it is simply a convenient way to re-use the sufficient statistic algorithm to fit the *SuSiE-RSS* model. The IBSS algorithm finds an approximation $q(\boldsymbol{z}_1, \cdots, \boldsymbol{z}_L) = \prod_{l=1}^{L} q(\boldsymbol{z}_l)$ to the posterior distribution $p_{\text{post}} = p(\boldsymbol{z}_1, \cdots, \boldsymbol{z}_L | \hat{\boldsymbol{z}}, \hat{\boldsymbol{R}}, \boldsymbol{\omega^2})$ by minimizing the Kullback-Leibler (KL) divergence from q to p_{post}. We summarizes the approximated posterior distribution $q(\boldsymbol{z}_1, \cdots, \boldsymbol{z}_L)$ using posterior inclusion probabilities (PIPs) and independent CSs.

The *SuSiE-RSS* model is implemented in the R package `susieR` available at `https://github.com/stephenslab/susieR`.

3.2 Likelihood with non-invertible LD matrix

The LD matrix $\hat{\boldsymbol{R}}$ is assumed to be invertible in Section 3.1. However, the LD matrix $\hat{\boldsymbol{R}}$ could be non-invertible because of the possible collinearity between variants. As a result, the density for $\hat{\boldsymbol{z}}$ and the likelihood are not well defined. There are several approaches to deal with non-invertible LD matrix.

Approach 1 The LD matrix is modified to be invertible, e.g. adding a small diagonal term to $\hat{\boldsymbol{R}}$, $\hat{\boldsymbol{R}}_\lambda = \hat{\boldsymbol{R}} + \lambda \boldsymbol{I}$, λ is a small positive number (Hormozdiari et al., 2014; Kichaev et al., 2014). The likelihood for $\boldsymbol{z}$ is

$$L_1(\boldsymbol{z}; \hat{\boldsymbol{z}}, \hat{\boldsymbol{R}}_\lambda) \coloneqq \exp\left(-\frac{1}{2}\boldsymbol{z}^\intercal \hat{\boldsymbol{R}}_\lambda \boldsymbol{z} + \boldsymbol{z}^\intercal \hat{\boldsymbol{z}}\right) \tag{3.2.1}$$

$$= L_{RSS}(\boldsymbol{z}; \hat{\boldsymbol{z}}, \hat{\boldsymbol{R}}_\lambda). \tag{3.2.2}$$

Approach 2 For model (3.1.1) to have a valid density, the variance-covariance matrix needs to be invertible. We use the modified invertible LD matrix $\hat{\boldsymbol{R}}_\lambda$ only in the variance part of (3.1.1), and keep the mean part unchanged. The likelihood for $\boldsymbol{z}$ is

$$L_2(\boldsymbol{z}; \hat{\boldsymbol{z}}, \hat{\boldsymbol{R}}, \lambda) \coloneqq \exp\left(-\frac{1}{2}\boldsymbol{z}^\intercal \hat{\boldsymbol{R}} \hat{\boldsymbol{R}}_\lambda^{-1} \hat{\boldsymbol{R}} \boldsymbol{z} + \boldsymbol{z}^\intercal \hat{\boldsymbol{R}} \hat{\boldsymbol{R}}_\lambda^{-1} \hat{\boldsymbol{z}}\right). \tag{3.2.3}$$

Approach 3 The variable is transformed into a lower dimension space, so the transformed variable has a valid density (Lozano et al., 2017; Park et al., 2017). Suppose

the LD matrix $\hat{\boldsymbol{R}}$ is positive semi-definite, the eigen-decomposition of $\hat{\boldsymbol{R}}$ is

$$\hat{\boldsymbol{R}} = \boldsymbol{Q}\boldsymbol{D}\boldsymbol{Q}^{\intercal}, \tag{3.2.4}$$

where $\boldsymbol{D}$ is a diagonal matrix with eigenvalues $d_1 \geqslant d_2 \cdots \geqslant d_r > 0$, r is the rank of $\hat{\boldsymbol{R}}$; $\boldsymbol{Q}$ is a $J \times r$ matrix of eigenvectors corresponding to non-zero eigenvalues, $\boldsymbol{Q}^{\intercal}\boldsymbol{Q} = \boldsymbol{I}_r$. We transform the model (3.1.1) as follows,

$$\boldsymbol{D}^{-1/2}\boldsymbol{Q}^{\intercal}\hat{\boldsymbol{z}} \sim N_r(\boldsymbol{D}^{1/2}\boldsymbol{Q}^{\intercal}\boldsymbol{z}, \boldsymbol{I}_r). \tag{3.2.5}$$

The transformed data, $\boldsymbol{D}^{-1/2}\boldsymbol{Q}^{\intercal}\hat{\boldsymbol{z}}$, become independent. The likelihood for $\boldsymbol{z}$ using the transformed model is

$$L_3(\boldsymbol{z}; \hat{\boldsymbol{z}}, \hat{\boldsymbol{R}}) := \exp\left(-\frac{1}{2}\boldsymbol{z}^{\intercal}\hat{\boldsymbol{R}}\boldsymbol{z} + \boldsymbol{z}^{\intercal}\boldsymbol{Q}\boldsymbol{Q}^{\intercal}\hat{\boldsymbol{z}}\right). \tag{3.2.6}$$

The likelihood (3.2.6) is equivalent to the likelihood one would obtain by using density (3.1.7), but replacing $\hat{\boldsymbol{R}}^{-1}$ with the Moore–Penrose pseudo-inverse of $\hat{\boldsymbol{R}}$, $\boldsymbol{Q}\boldsymbol{D}^{-1}\boldsymbol{Q}^{\intercal}$.

Approach 4 We use likelihood (3.1.8) even $\hat{\boldsymbol{R}}$ is non-invertible. Note that this function exists, and is easily computed, for non-invertible $\hat{\boldsymbol{R}}$.

The four methods above are not equivalent in general, but they are connected. Proposition 3 summarizes the connection between likelihoods.

Proposition 3. *Let* $\hat{\boldsymbol{R}}_\lambda = \hat{\boldsymbol{R}} + \lambda \boldsymbol{I}$,

$$\lim_{\lambda \to 0} L_1(\boldsymbol{z}; \hat{\boldsymbol{z}}, \hat{\boldsymbol{R}}_\lambda) = L_{RSS}(\boldsymbol{z}; \hat{\boldsymbol{z}}, \hat{\boldsymbol{R}}); \tag{3.2.7}$$

$$\lim_{\lambda \to 0} L_2(\boldsymbol{z}; \hat{\boldsymbol{z}}, \hat{\boldsymbol{R}}, \lambda) = L_3(\boldsymbol{z}; \hat{\boldsymbol{z}}, \hat{\boldsymbol{R}}); \tag{3.2.8}$$

$$\lim_{\lambda \to 0} L_3(\boldsymbol{z}; \hat{\boldsymbol{z}}, \hat{\boldsymbol{R}}_\lambda) = L_{RSS}(\boldsymbol{z}; \hat{\boldsymbol{z}}, \hat{\boldsymbol{R}}). \tag{3.2.9}$$

Proof. As $\lambda \to 0$, $\hat{\boldsymbol{R}}_\lambda \to \hat{\boldsymbol{R}}$, (3.2.7) is trivial.

Next, we show (3.2.8). Let $\boldsymbol{P} = \boldsymbol{Q}\boldsymbol{D}^{1/2}$ with Moore-Penrose inverse $\boldsymbol{P}^\dagger = \boldsymbol{D}^{-1/2}\boldsymbol{Q}^\intercal$, we can write $L_2(\boldsymbol{z}; \hat{\boldsymbol{z}}, \hat{\boldsymbol{R}}, \lambda)$ as follows,

$$\begin{aligned} &L_2(\boldsymbol{z}; \hat{\boldsymbol{z}}, \hat{\boldsymbol{R}}, \lambda) \\ &= \exp\left(-\frac{1}{2}\boldsymbol{z}^\intercal \boldsymbol{P}\boldsymbol{P}^\intercal (\boldsymbol{P}\boldsymbol{P}^\intercal + \lambda \boldsymbol{I})^{-1} \boldsymbol{P}\boldsymbol{P}^\intercal \boldsymbol{z} + \boldsymbol{z}^\intercal \boldsymbol{P}\boldsymbol{P}^\intercal (\boldsymbol{P}\boldsymbol{P}^\intercal + \lambda \boldsymbol{I})^{-1} \hat{\boldsymbol{z}}\right). \end{aligned} \tag{3.2.10}$$

In the limit $\lambda \to 0$, $\boldsymbol{P}^\intercal(\boldsymbol{P}\boldsymbol{P}^\intercal + \lambda \boldsymbol{I})^{-1} \to \boldsymbol{P}^\dagger$ (Theorem 3.4 in Albert (1972)). Therefore,

$$\lim_{\lambda \to 0} L_2(\boldsymbol{z}; \hat{\boldsymbol{z}}, \hat{\boldsymbol{R}}, \lambda) = \exp\left(-\frac{1}{2}\boldsymbol{z}^\intercal \boldsymbol{P}\boldsymbol{P}^\dagger \boldsymbol{P}\boldsymbol{P}^\intercal \boldsymbol{z} + \boldsymbol{z}^\intercal \boldsymbol{P}\boldsymbol{P}^\dagger \hat{\boldsymbol{z}}\right) \tag{3.2.11}$$

$$= \exp\left(-\frac{1}{2}\boldsymbol{z}^\intercal \hat{\boldsymbol{R}} \boldsymbol{z} + \boldsymbol{z}^\intercal \boldsymbol{Q}\boldsymbol{Q}^\intercal \hat{\boldsymbol{z}}\right) \tag{3.2.12}$$

$$= L_3(\boldsymbol{z}; \hat{\boldsymbol{z}}, \hat{\boldsymbol{R}}). \tag{3.2.13}$$

Finally, we show (3.2.9). Since $\hat{\boldsymbol{R}}_\lambda$ is full rank, the matrix of eigenvectors $\boldsymbol{Q}_\lambda \in$

$\mathbb{R}^{J\times J}$ satisfies $\boldsymbol{Q}_\lambda\boldsymbol{Q}_\lambda^\intercal = \boldsymbol{I}_J$. Therefore, the likelihood for $\boldsymbol{z}$ is

$$L_3(\boldsymbol{z};\hat{\boldsymbol{z}},\hat{\boldsymbol{R}}_\lambda) \coloneqq \exp\left(-\frac{1}{2}\boldsymbol{z}^\intercal\hat{\boldsymbol{R}}_\lambda\boldsymbol{z} + \boldsymbol{z}^\intercal\hat{\boldsymbol{z}}\right). \tag{3.2.14}$$

As λ goes to 0, $L_3(\boldsymbol{z};\hat{\boldsymbol{z}},\hat{\boldsymbol{R}}_\lambda)$ converges to $L_{RSS}(\boldsymbol{z};\hat{\boldsymbol{z}},\hat{\boldsymbol{R}})$. □

As the small diagonal term λ goes to 0 in Approach 1 and 2, Approach 2 is equivalent to Approach 3, Approach 1 is equivalent to Approach 4. Moreover, if we use $\hat{\boldsymbol{R}}_\lambda$ in Approach 3 to do the transformation and take λ goes to 0, the likelihood converges to likelihood (3.1.8), not the likelihood in Approach 3 (3.2.6).

The Approach 3 and 4 are equivalent when the observed z scores $\hat{\boldsymbol{z}}$ lies in the subspace spanned by the eigenvectors of the LD matrix $\hat{\boldsymbol{R}}$ (i.e. $\boldsymbol{Q}\boldsymbol{Q}^\intercal\hat{\boldsymbol{z}} = \hat{\boldsymbol{z}}$). When the observed z scores does not lie in the subspace spanned by the eigenvectors of $\hat{\boldsymbol{R}}$, which could happen when $\hat{\boldsymbol{R}}$ is estimated from a reference panel, the Approach 3 and 4 are not equivalent. The Approach 3 transforms the z scores into the subspace spanned by the eigenvectors of $\hat{\boldsymbol{R}}$, whereas Approach 4 uses the observed z scores even it does not lie in the subspace spanned by the eigenvectors of $\hat{\boldsymbol{R}}$.

For Approach 1 and 4, the posterior inference is independent of the LD matrix $\hat{\boldsymbol{R}}_\lambda$ or $\hat{\boldsymbol{R}}$ under the *SER-RSS* model (Corollary 1). However, this is not true for Approach 2 and 3. The Approach 2 uses $\hat{\boldsymbol{R}}$ in the mean, but $\hat{\boldsymbol{R}}_\lambda$ in the variance. Consequently, the inference depends on all observed z scores, LD matrix $\hat{\boldsymbol{R}}$ and the extra diagonal term λ. The inference from Approach 3 depends on the LD matrix as well, because of the $\boldsymbol{Q}\boldsymbol{Q}^\intercal\hat{\boldsymbol{z}}$ in the likelihood (3.2.6).

We use Approach 4 in *SuSiE-RSS*, because the inference does not depend on

the LD matrix under the *SER-RSS* model, and the likelihood is equivalent to the likelihood in Approach 1 as λ goes to 0.

3.3 LD matrix

A key assumption in *SuSiE-RSS* model is that the LD matrix $\hat{\boldsymbol{R}}$ is an accurate estimate of the correlation among SNPs in the original GWAS samples. Computing the sample correlation matrix using the original genotype matrix, $\boldsymbol{X}$, provides an accurate correlation among SNPs in the sample, and it is the best LD we can obtain. However, the original genotype data are not always publicly available for privacy reasons, and researchers are not sharing the LD matrix from the GWAS analysis. Typically, $\hat{\boldsymbol{R}}$ is estimated from some suitable public reference genotype panels, and we hope that LD information from these reference panels could accurately represent the correlations of SNPs in the study population. Benner et al. (2017) did a thorough assessment about the influence of the reference panel, and concluded that inaccurate LD information leads to an inflation of false positives. They showed the size of the reference panel needs to scale with the GWAS sample size, and suggested a reference panel of 1,000 individuals from the target population is adequate for a GWAS cohort of up to 10,000 individuals.

To make some correction for the LD matrix obtained from the reference panel, we use a simple method to modify the LD matrix. We use regularized LD matrix $\tilde{\boldsymbol{R}}_s = (1-s)\hat{\boldsymbol{R}} + s\boldsymbol{I}$, for some $0 < s < 1$. The similar regularization for the LD matrix is also used in covariance estimation (Ledoit and Wolf, 2004) and methods related to summary data (e.g. Pasaniuc et al., 2014; Kichaev et al., 2014; Mak et al.,

2017). In previous works, the parameter s is usually fixed at an arbitrary small value, or chosen using cross validation. We estimate the parameter s by maximizing likelihood under the null ($\boldsymbol{z} = \boldsymbol{0}$),

$$\hat{s} = \arg \max_{s \in (0,1)} N(\hat{\boldsymbol{z}}; \boldsymbol{0}, (1-s)\hat{\boldsymbol{R}} + s\boldsymbol{I}). \tag{3.3.1}$$

When the LD matrix $\hat{\boldsymbol{R}}$ is computed using the original genotypes that were used to compute the marginal z scores, the vector of observed marginal z scores lies in the subspace spanned by the eigenvectors of $\hat{\boldsymbol{R}}$ approximately. The part of z scores that lies outside the subspace spanned by the eigenvectors of $\hat{\boldsymbol{R}}$ is very small, thus the estimated $\hat{s}$ is small. If the LD matrix is obtained from the reference panel, its accuracy is limited by the quality of the panel. The marginal z scores does not lie in the subspace spanned by the eigenvectors of $\hat{\boldsymbol{R}}$, and the part of z scores that is outside the subspace spanned by the eigenvectors of $\hat{\boldsymbol{R}}$ is large, thus the estimated $\hat{s}$ is larger. The estimated $\hat{s}$ gives information about the consistency between the z scores and LD matrix. To solve (3.3.1), we used numerical optimization (the Brent method (Brent, 2002) implemented in the optim function in R). The main computational expense is an initial eigen-decomposition of $\hat{\boldsymbol{R}}$, which is $O(J^3)$ and thus non-trivial if J is large. Because we found regularization typically provided small benefits in *SuSiE* methods (see Section 3.6), our software sets $s = 0$ as default and avoid this computational expense.

In our earlier works, we explored modification for the LD matrix from reference panels to make it more consistent with the GWAS z scores. As we refined our

method, we found the modification could make things worse rather than better. We include the details in Appendix B.1.

3.4 Detecting allele flips

A common mistake in fine-mapping with summary statistics is "allele flips", where different coding of the two alleles at a SNP are used in the study sample (used to compute $\hat{\boldsymbol{z}}$) and the reference panel (used to compute $\hat{\boldsymbol{R}}$). Reference and alternative alleles may mismatch between summary statistics from GWAS and the reference panel. Suppose two SNPs are perfectly positively correlated in the reference panel (i.e. pairwise correlation equals 1), the corresponding z scores for the two SNPs should be same with the same sign. But the reference and alternative allele is flipped for one SNP in GWAS. Consequently, the z scores for the two SNPs have opposite sign in GWAS and the z scores are not consistent with the LD matrix from the reference panel.

To detect the variants with flipped allele, we use the likelihood ratio test. Based on the model for $\hat{\boldsymbol{z}}$ with regularized LD matrix $\tilde{\boldsymbol{R}}_{\hat{s}} = (1-\hat{s})\hat{\boldsymbol{R}} + \hat{s}\boldsymbol{I}$,

$$\hat{\boldsymbol{z}}|\tilde{\boldsymbol{R}}_{\hat{s}}, \boldsymbol{z} \sim N(\tilde{\boldsymbol{R}}_{\hat{s}}\boldsymbol{z}, \tilde{\boldsymbol{R}}_{\hat{s}}), \tag{3.4.1}$$

the distribution of marginal z score for SNP j conditional on all other observed z scores is

$$\hat{z}_j|\hat{\boldsymbol{z}}_{-j}, \tilde{\boldsymbol{R}}_{\hat{s}}, \boldsymbol{z} \sim N(-\frac{1}{\boldsymbol{\Omega}_{jj}}\boldsymbol{\Omega}_{j,-j}\hat{\boldsymbol{z}}_{-j} + \frac{1}{\boldsymbol{\Omega}_{jj}}z_j, \frac{1}{\boldsymbol{\Omega}_{jj}}), \tag{3.4.2}$$

in which $\boldsymbol{\Omega}$ is the precision matrix, $\boldsymbol{\Omega} = \tilde{\boldsymbol{R}}_{\hat{s}}^{-1}$, $\boldsymbol{\Omega}_{j,-j}$ is the j-th row of $\boldsymbol{\Omega}$ omitting

the j-th element, $\hat{\boldsymbol{z}}_{-j}$ is the observed z scores without the j-th one. If SNP j is either in strong LD with other SNPs (which implies $1/\boldsymbol{\Omega}_{jj} \approx 0$) or has small effect ($z_j \approx 0$), then this yields the approximation

$$\hat{z}_j|\hat{\boldsymbol{z}}_{-j}, \tilde{\boldsymbol{R}}_{\hat{s}}, z_j = 0 \sim N(-\frac{1}{\boldsymbol{\Omega}_{jj}}\boldsymbol{\Omega}_{j,-j}\hat{\boldsymbol{z}}_{-j}, \frac{1}{\boldsymbol{\Omega}_{jj}}). \tag{3.4.3}$$

When there is no flipped allele and the LD matrix is estimated from the reference panel, the distribution of the standardized differences between the observed and predicted z score, $t_j := \boldsymbol{\Omega}_{jj}(\hat{z}_j + \frac{1}{\boldsymbol{\Omega}_{jj}}\boldsymbol{\Omega}_{j,-j}\hat{\boldsymbol{z}}_{-j})$, has longer tail than standard normal $N(0,1)$. Therefore, we use a mixture of normals to model the conditional distribution empirically,

$$\hat{z}_j|\hat{\boldsymbol{z}}_{-j}, \tilde{\boldsymbol{R}}_{\hat{s}}, z_j = 0 \sim \sum_{k=1}^{K} p_k N(-\frac{1}{\boldsymbol{\Omega}_{jj}}\boldsymbol{\Omega}_{j,-j}\hat{\boldsymbol{z}}_{-j}, \frac{\sigma_k^2}{\boldsymbol{\Omega}_{jj}}). \tag{3.4.4}$$

We take $\sigma_1, \cdots, \sigma_K$ to be a grid of fixed positive numbers with minimum value as 0.8, maximum value as $2\sqrt{\max(t_j^2)}$. We take $\sigma_k = 1.05\sigma_{k-1}$. The mixture proportions $\boldsymbol{p} = (p_1, \cdots, p_K)$ are non-negative and sum to one. We estimate $\boldsymbol{p}$ by maximum likelihood, which is a convex optimization problem and can be solved effectively using `mixsqp` (Kim et al., 2020). The distribution (3.4.3) is a special case of (3.4.4) with the grid of $\sigma_1 = 1$, $K = 1$. The distribution (3.4.3) with $s = 0$ is also used in z-score imputation (Lee et al., 2013) and GWAS quality control (Chen et al., 2020).

We test the null hypothesis H_0 : the sign of $\hat{z}_j$ is correct v.s. the alternative

hypothesis H_1 : the sign of $\hat{z}_j$ is flipped using likelihood ratio test,

$$\text{LR} := \frac{\sum_{k=1}^{K} \hat{p}_k N(\hat{z}_j; \frac{1}{\boldsymbol{\Omega}_{jj}} \boldsymbol{\Omega}_{j,-j} \hat{\boldsymbol{z}}_{-j}, \frac{\sigma_k^2}{\boldsymbol{\Omega}_{jj}})}{\sum_{k=1}^{K} \hat{p}_k N(\hat{z}_j; -\frac{1}{\boldsymbol{\Omega}_{jj}} \boldsymbol{\Omega}_{j,-j} \hat{\boldsymbol{z}}_{-j}, \frac{\sigma_k^2}{\boldsymbol{\Omega}_{jj}})}. \tag{3.4.5}$$

We apply the test for variants with z scores greater than 2 in magnitude.

We show one example in Figure 3.1. The summary statistics are obtained using 10,000 individuals from UK Biobank and simulated phenotypes. The LD matrix is estimated from another 1,000 random individuals in UK Biobank. However, we deliberately mismatched ("flipped") the reference and alternative allele for one SNP (yellow point in Figure 3.1). Using the resulting z scores and mismatched LD matrix, *SuSiE-RSS* identifies two *CSs*, one containing the true signal, and another containing the SNP with flipped allele. With the correct allele alignment, this false discovery is avoided. Our diagnostic plot comparing each z score against its expected value shows the yellow SNP as a potential outlier, and our logLR statistic identifies this SNP as a likely allele flip (log LR = 9.01; no other SNP here has log LR exceeding 2).

3.5 Algorithm refinement

All *SuSiE* related models (i.e. *SuSiE*, *SuSiE-suff* , *SuSiE-RSS*) use the IBSS algorithm to approximate the posterior distribution. The IBSS algorithm is optimizing the variational objective function, known as the "evidence lower bound" (ELBO) (see Wang et al. (2020) for details). Wang et al. (2020) pointed out although the algorithm

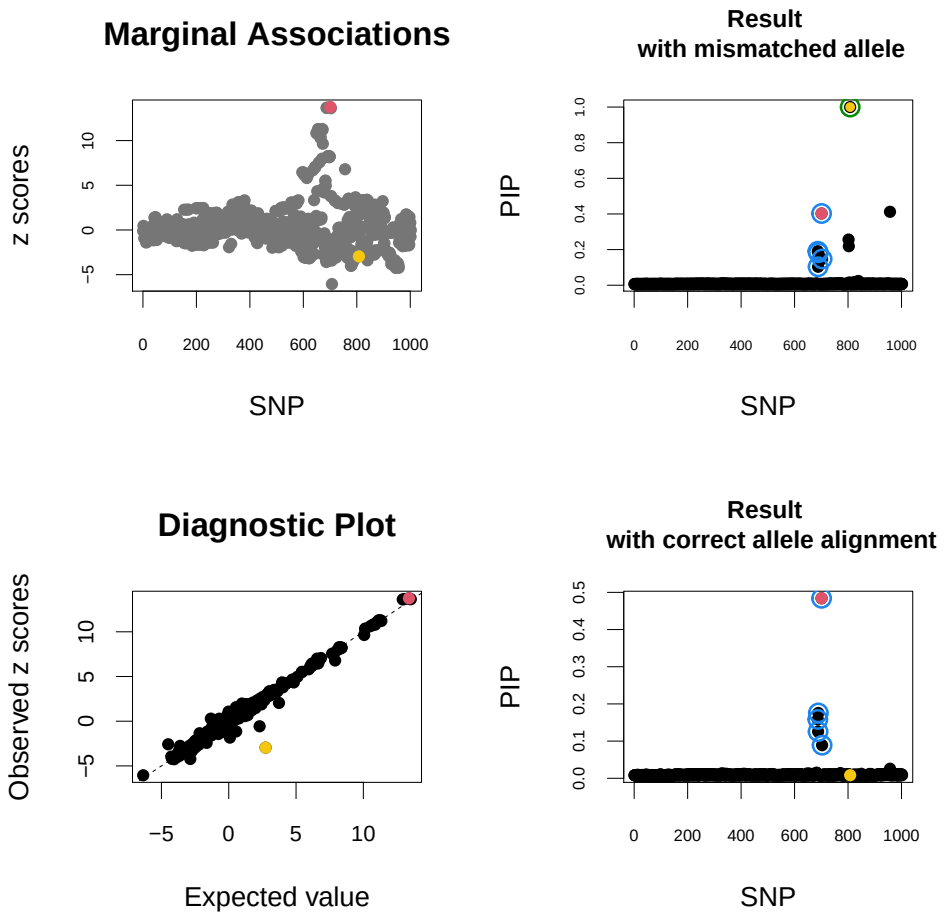


Figure 3.1: **Fine-mapping example with allele mismatch between GWAS and reference panel.** Results are from a simulated data set with $p = 1,002$ SNPs with one SNP having a true effect (red) and one SNP having allele mismatched between study and reference panel (yellow). **Top left:** z scores for each SNP; **Top right:** PIPs computed by *SuSiE-RSS* using the mismatched summary data, with two *CSs* highlighted in blue and green; *SuSiE-RSS* incorrectly identifies a *CS* containing the mismatched (yellow) SNP. **Bottom-left:** Diagnostic plot plotting each observed z score against its expected value. The mismatched SNP stands out as a potential outlier. **Bottom right:** PIPs computed by *SuSiE-RSS* with correct allele alignment, with *CS* highlighted in blue; *SuSiE-RSS* finds one *CS*, which contains the true effect SNP, and the false positive *CS* is avoided.

usually performs well, sometimes it can produce a poor fit, due to getting stuck in a local optimum. Although this is rare, it can provide misleading results when it occurs. In particular we found that the IBSS algorithm occasionally converges to a solution with one or more false positive *CSs* (i.e. *CSs* containing all null SNPs) and misses alternative explanations that avoid these false positives and have higher objective value.

Motivated by this we developed a simple refinement procedure for finding and comparing these alternative explanations. It is equally applicable for all methods using the IBSS algorithm. We explain the details using *SuSiE* model. Based on a previous *SuSiE* fit, we loop through all identified *CSs*. For each identified Credible Set, CS_k, we remove all SNPs in CS_k and refit *SuSiE* model, t_k. This forces *SuSiE* to consider alternative explanations other than SNPs in CS_k. Then we fit *SuSiE* model with all SNPs but initializing at t_k, which yields a new *SuSiE* model s_k. Comparing achieved ELBO for s_k, $k = 1, \cdots, K$ and s, we take the model with the highest ELBO and check all identified *CSs* again. This process is repeated until no improvements in the objective occur. By construction, the refinement procedure produces a solution whose objective is at least as big as the original IBSS solution. See Algorithm 2 for details. Since this refinement procedure involves re-fitting model using the IBSS algorithm with different parameters, it incurs additional computational expense.

Algorithm 2 Refinement procedure

Require: a *SuSiE* object, s, with K *CSs*

while TRUE **do**
  for $k \leftarrow 1$ to number of *CSs* in s **do**
    Remove all SNPs included in the kth *CS* (set SNPs in CS_k to have prior weight 0) and fit *SuSiE* model, t_k
    Fit *SuSiE* model, s_k, with initialization at t_k
  if $\max_k \text{ELBO}(s_k) > \text{ELBO}(s)$ **then**
    $s \leftarrow s_{k'}$, where $k' = \arg_k \max \text{ELBO}(s_k)$
  else return s

We use one example from our simulations to illustrate that the refinement procedure (Algorithm 2) works well in a challenging fine-mapping setting. The simulation details are in Section 3.6. The example is summarized in Figure 3.2. In this example, the causal SNP 1 has moderate correlation with both SMA and causal SNP 2; the correlation between SMA and SNP 2 is weak. The causal SNP 1 and SNP 2 have opposite effects, therefore the SNP 2 cancels out some of the marginal effect of SNP 1, and the SNP with the strongest marginal association (SMA) is not one of the actual effect SNPs. The default IBSS algorithm yields three *CSs*, two of them do not contain any true effect SNP (Figure 3.2, middle panel). Using the refinement steps, it finds two *CSs*, neither containing the SMA, and each containing one of the effect SNPs. The achieved ELBO from the refinement procedure becomes higher. The default IBSS algorithm converges to a local optimum and the refinement procedure helps it escape from the local optimum.

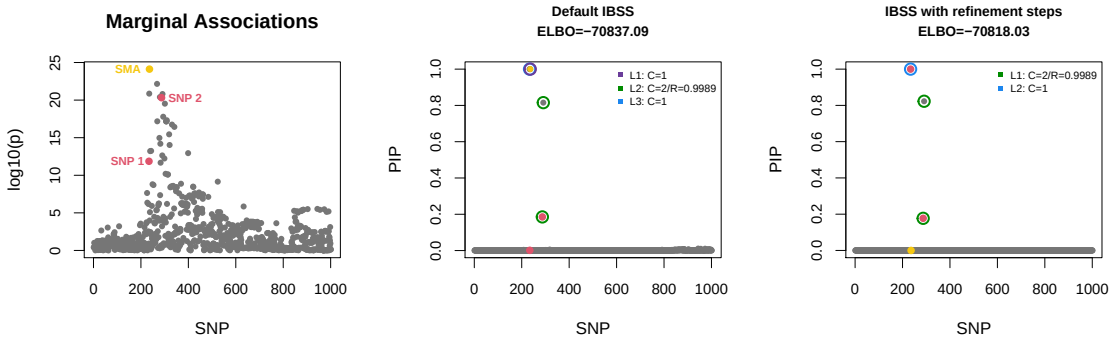


Figure 3.2: **Fine-mapping example to illustrate that IBSS algorithm with additional refinement steps can deal with a challenging case.** Results are from a simulated data set with $p = 1,001$ SNPs. Two out of the 1,001 SNPs have non-zero effects (red points, labeled "SNP 1" and "SNP 2" in the left-hand panel). The strongest marginal association (SMA) is a non-effect SNP (yellow point, labeled "SMA" in the left-hand panel). The IBSS algorithm with default settings (middle panel), identifies three *CSs*, two of them false positives containing no true effect SNPs (one contains the SMA). The refinement procedure finds two 95% *CSs*, each containing a true effect SNP.

3.6 Numerical Comparisons

We performed numerical experiments using real genotype data from UK Biobank (Sudlow et al., 2015; Bycroft et al., 2018) to mimic the real fine-mapping application. There are 274,549 unrelated White British individuals after removing outlier individuals and individuals with different self-reported and genetic sex. We randomly selected 200 non-overlapping regions, each region contains around 1,000 SNPs. We included SNPs with minor allele frequency greater than 0.01. We randomly sampled 50,000 individuals to simulate phenotypes and computed in-sample LD matrix. To investigate the impact of misspecification of LD matrix, we randomly sampled another 500 and 1,000 individuals as two reference panels. We assess the performance

of *SuSiE-RSS* using in-sample LD matrix and reference LD matrices.

The simulation setting is similar to *SuSiE* simulations in Wang et al. (2020) and we briefly describe it here. We column-standardized the genotype matrix $\boldsymbol{X}$, so rare SNPs have larger effects than the common SNPs in the original genotype scale. We generated $\boldsymbol{y}$ under the multiple regression model (1.0.1). The number of non-zero effects is specified by S. We randomly sampled S effect variants from $\{1, \ldots, J\}$ and generated S effects independently from $N(0, 1)$. We drew $\boldsymbol{y} \sim N(\boldsymbol{Xb}, \sigma^2 \boldsymbol{I}_N)$, where σ^2 achieves the specified PVE, PVE $= \frac{\text{Var}(\boldsymbol{Xb})}{\sigma^2 + \text{Var}(\boldsymbol{Xb})}$, Var(.) denotes sample variance. For each SNP j, we fitted simple linear regression with $\boldsymbol{y}$ and $\boldsymbol{x}_j$ and obtained the estimated z score, $\hat{z}_j$. We generated data using $S \in \{1, 2, 3\}$ and PVE = 0.005. In total, we generated $200 \times 3 = 600$ data sets.

We used Dynamic Statistical Comparisons (DSC) to simulate data and compare different methods. More details of DSC are in `https://stephenslab.github.io/dsc-wiki/overview`. The simulation code and results are available at `https://github.com/zouyuxin/dsc_susierss`.

3.6.1 Posterior inclusion probabilities

Posterior inclusion probability (PIP) is a standard output for most fine-mapping methods. Here we compare our method, *SuSiE-RSS*, with three other software that are commonly used in fine-mapping: *SuSiE* (Wang et al., 2020), CAVIAR (Hormozdiari et al., 2014, version 2.2) and FINEMAP (Benner et al., 2016, version 1.4). Since the sample size is large, we use *SuSiE* with sufficient statistics, which gives the same result as individual-level data (see details are in Chapter 2). *SuSiE-suff* and *SuSiE-*

RSS are implemented in R, CAVIAR and FINEMAP are all implemented in C++. CAVIAR and FINEMAP implement similar Bayesian variable selection models with different algorithms and priors on the effect sizes. CAVIAR exhaustively searches all possible combinations of up to L non-zero effects among the J variants. FINEMAP uses shotgun stochastic search to explore different combinations.

In CAVIAR, we set all prior inclusion probabilities to $1/J$ to match the default settings used in other methods. We set the maximum number of causal SNPs to the true value S in CAVIAR and FINEMAP. For FINEMAP, we also try to increase the maximum number of causal SNPs to 4. For *SuSiE-suff* and *SuSiE-RSS*, we set $L = 10$, which is allowing maximum 10 causal SNPs.

We evaluate power versus False Discovery Rate (FDR). The FDR and power are calculated as FDR $:= \frac{\text{FP}}{\text{TP+FP}}$, power $:= \frac{\text{TP}}{\text{TP+FN}}$, where FP, TP and FN denote the number of False Positives, True Positives and True Negatives at a given PIP threshold.

The *SuSiE-suff* method with the refinement procedure has lower FDR at a given PIP threshold (Figure 3.3). In particular, among the SNPs assigned PIP ≈ 1, where the FDR after refinement is ≈ 0 whereas the FDR without refinement is ≈ 0.04. All subsequent results in this chapter therefore use the refinement procedure.

In-sample LD matrix

Figure 3.4 shows Power vs FDR from all methods as PIP threshold varies, with all methods run using the in-sample LD matrix. *SuSiE-RSS* performs similarly to *SuSiE-suff* and FINEMAP, all of which outperform CAVIAR in these simulations.

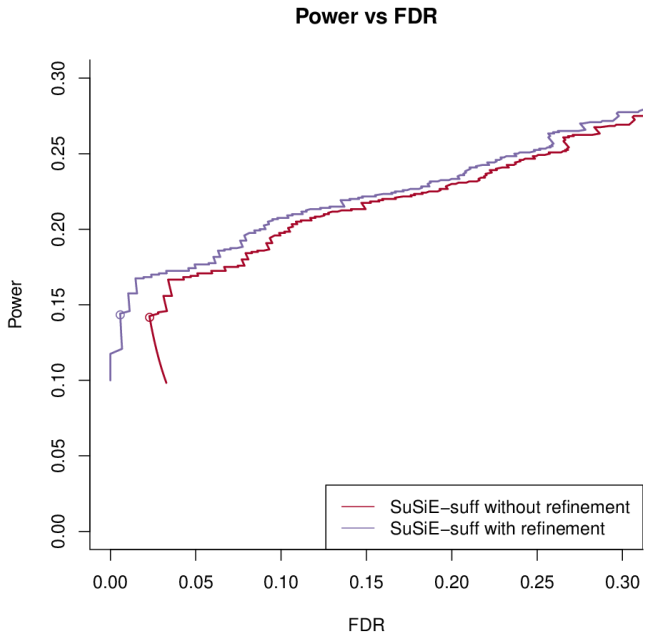


Figure 3.3: **Comparison of *SuSiE-suff* with and without the refinement procedure.** The plot summarizes Power versus FDR using the PIPs from *SuSiE-suff*. The open circles in the highlight power versus FDR at PIP threshold of 0.95.

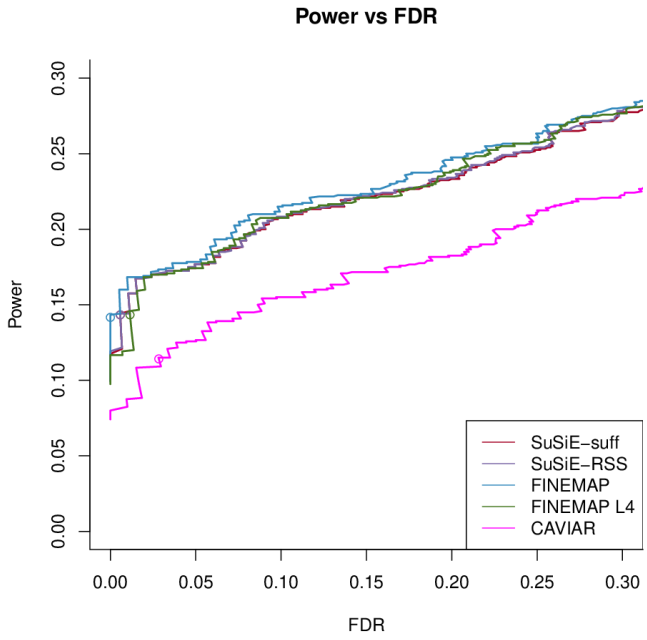


Figure 3.4: **Comparison of Power vs FDR for each method with in-sample LD matrix.** The plot shows how Power and FDR co-vary as PIP threshold changes. Circles indicate results at PIP threshold 0.95.

The close performance of *SuSiE-RSS* and *SuSiE-suff* is expected due to the use of the in-sample LD and the low PVE in these simulations. With the in-sample LD matrix, *SuSiE-suff* and *SuSiE-RSS* use the same $\boldsymbol{X}^\intercal\boldsymbol{X}$ (differ by a factor of sample size N). *SuSiE-RSS* uses marginal z scores $\hat{\boldsymbol{z}}$ as the $\boldsymbol{X}^\intercal\boldsymbol{y}$ in *SuSiE-suff*, which is a good approximation when each variant has small correlation with the quantitative phenotype $\boldsymbol{y}$. *SuSiE-suff* estimates the residual variance σ^2, which is approximately variance of $\boldsymbol{y}$ when the sample size is large and the total PVE is small. Since our simulation has large sample size and small total PVE, the information lost in summary statistics is negligible.

Comparing the running time for different methods, *SuSiE-RSS* is faster than FINEMAP and CAVIAR (Table 3.1), even with the additional burden of running the refinement procedure.

Table 3.1: **Runtimes in seconds**

method	mean	min.	max.
SuSiE-suff without refinement	1.40	0.40	18.61
SuSiE-suff with refinement	4.81	1.44	62.34
SuSiE-RSS without refinement	1.31	0.39	20.42
SuSiE-RSS with refinement	4.64	1.43	74.15
FINEMAP	12.92	1.00	42.93
FINEMAP L4	16.11	1.67	39.27
CAVIAR	1516.91	3.54	4831.95

LD matrix from reference panel

Next, we assess how methods performance change with use of a reference panel. For each method, we tried using reference panels of 500 samples and 1,000 samples, with no regularization, small regularization ($s = 0.001$), or estimated regularization

(Section 3.3). Results (Power vs FDR) are shown in Figure 3.5. The different methods responded differently to changes in the reference panel size and regularization. FINEMAP and *SuSiE-RSS* both show notably better performance with the larger reference panel, whereas CAVIAR performed about the same for both sizes of panel. As a result, FINEMAP and *SuSiE-RSS* outperform CAVIAR with large reference panel, but with small reference panel CAVIAR is competitive. Using reference panel with 500 samples, FINEMAP with maximum 4 causal SNPs has much higher FDR than the *SuSiE-RSS* model with maximum 10 causal SNPs (12% vs. 25% at PIP threshold of 0.95). For FINEMAP, increasing the maximum number of causal SNPs from oracle to 4, the FDR increases dramatically (from 6% to more than 25% at PIP threshold of 0.95).

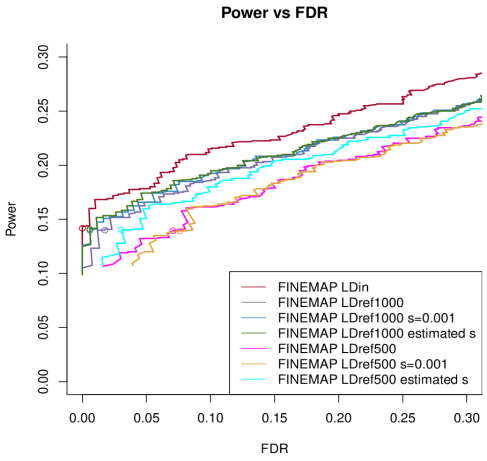
Regularization of the LD matrix improved the performance of FINEMAP, particularly FINEMAP L4 with smaller panel, but did not greatly impact *SuSiE-RSS* or CAVIAR. The reasons for these differences is unclear to us.

The estimated $\hat{s}$ is small when the LD matrix is estimated from the GWAS samples. As the quality of the reference panel decreases (the size of the reference panel decreases in the simulation), the estimated $\hat{s}$ becomes larger (Figure 3.6).

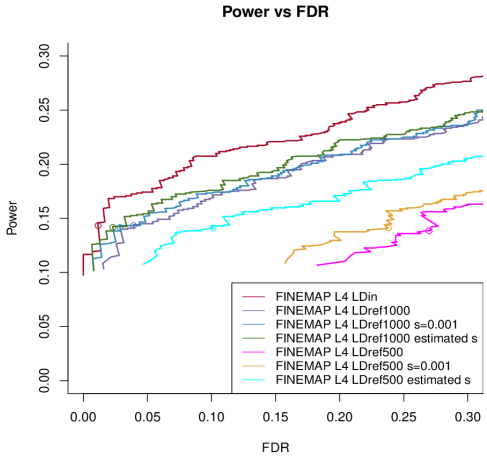
3.6.2 Credible sets

SuSiE-suff and *SuSiE-RSS* produce multiple Credible Sets (*CSs*) for each region, each aimed at capturing one effect SNP. The *CSs* with "purity" less than 0.5 are discarded, where purity is defined as the smallest absolute correlation among all pairs of SNPs within the *CS*. FINEMAP produces similar *CSs* in v1.4 and we discard the *CSs*

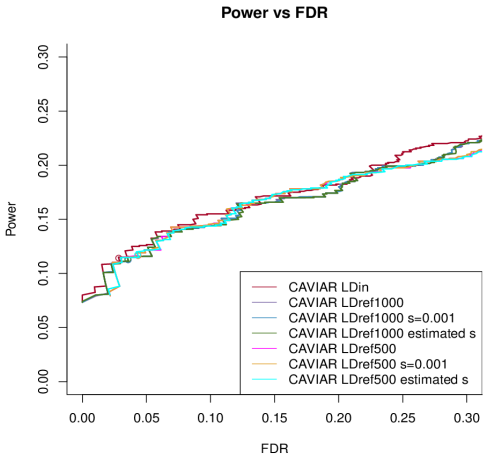
(a) FINEMAP with oracle maximum number of effects



(b) FINEMAP with 4 maximum effects



(c) CAVIAR with oracle maximum number of effects



(d) *SuSiE-RSS* with 10 maximum effects

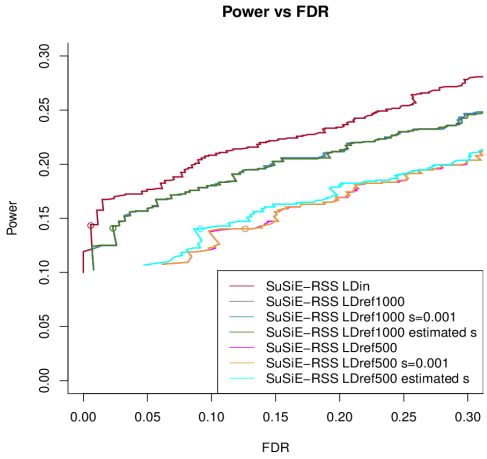


Figure 3.5: **Comparison of Power and FDR for each method with different LD matrix.** Each curve shows how Power and FDR co-vary as PIP threshold changes. We used LD matrices estimated from reference panels of different sizes (500 and 1,000 samples) and with different regularization parameter s. Circles indicate results at PIP threshold 0.95.

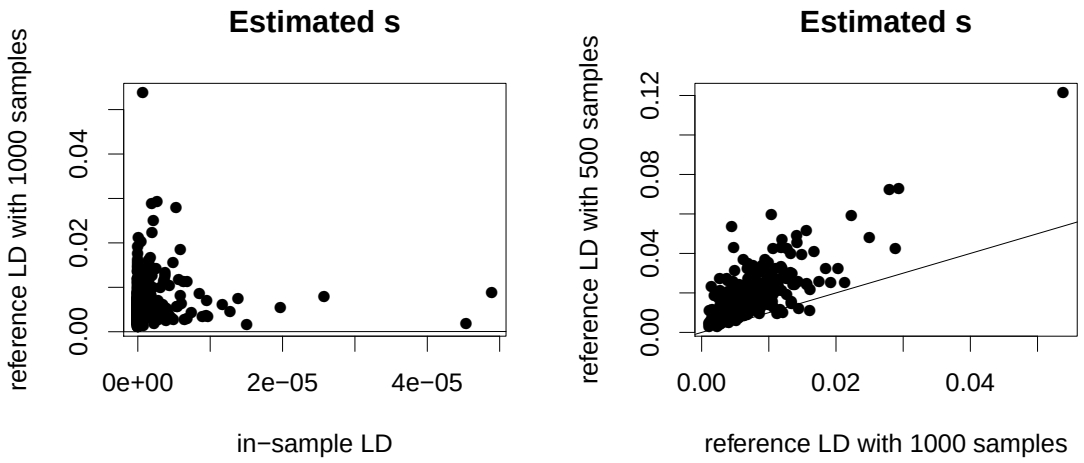


Figure 3.6: **Comparing estimated $\hat{s}$ using different LD matrix.**

with "purity" less than 0.5 as well. In contrast, CAVIAR produces one credible set per region aiming to cover all effect SNPs, which is very different from our credible set definition. Therefore, we compare *CSs* from *SuSiE-RSS* with *SuSiE-suff* and FINEMAP with maximum 4 causal SNPs.

We assess the 95% *CSs* using several criteria: 1. the empirical coverage levels for *CSs*, which is the proportion of *CSs* that contain an effect SNP; 2. power, which is the proportion of true effect SNPs included in a *CS*; 3. median number of SNPs in each *CS*; 4. median purity (Figure 3.7). By all metrics, the *CSs* from *SuSiE-RSS* using in-sample LD performs similarly to *SuSiE-suff* and FINEMAP. Using the LD matrix from reference panels, coverage and power decrease for all methods. For the smaller reference panels (500) the coverage of all methods drops below 0.9, emphasising the importance of using sufficiently large reference panels.

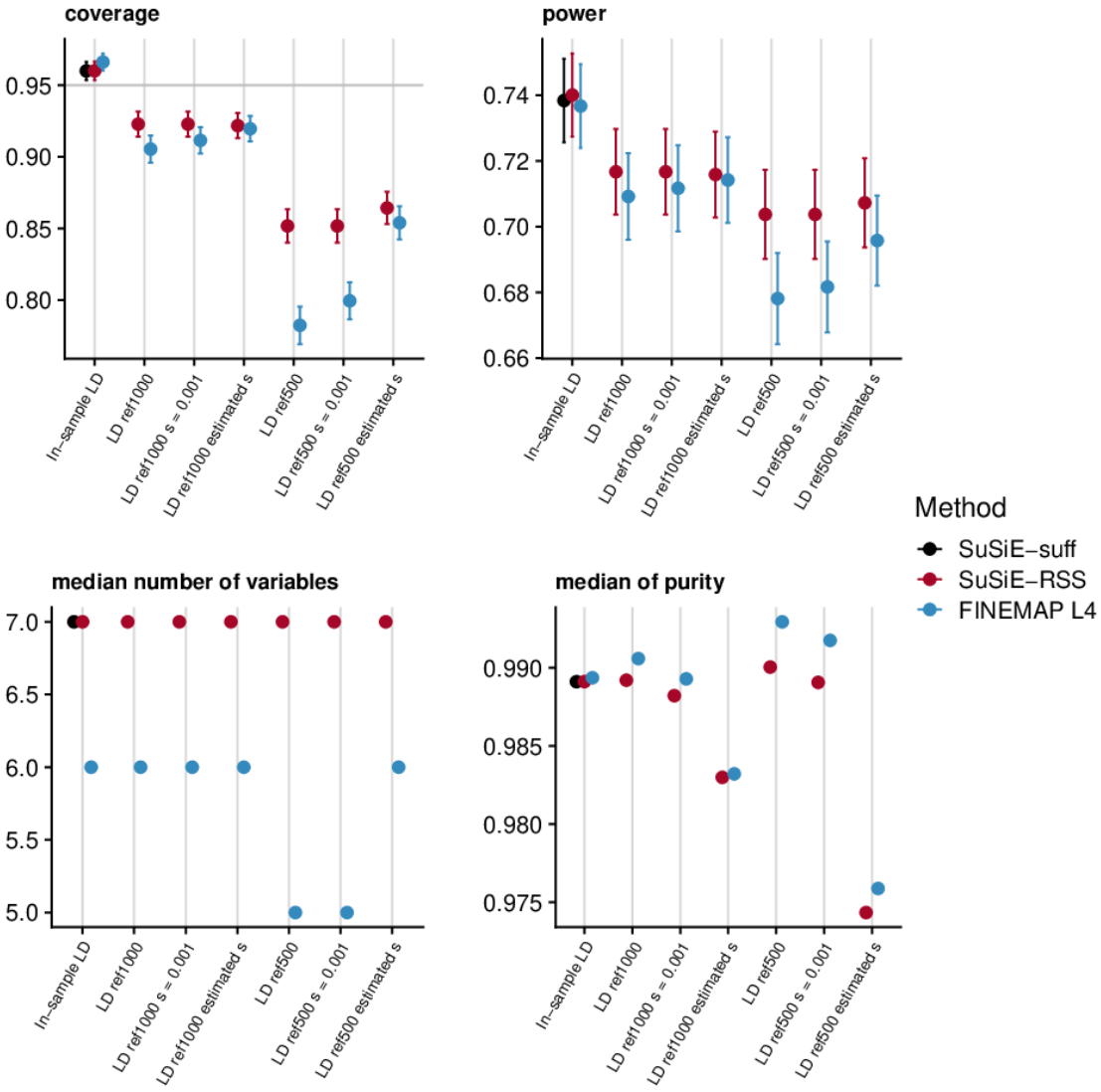


Figure 3.7: **Comparison of 95% credible sets from *SuSiE-suff*, *SuSiE-RSS* and FINEMAP.** Panels show coverage, power, median size and median purity. These statistics are computed by pooling all CSs from all data sets. The error bars in coverage and power plots show 2× standard error.

3.7 Discussion

In this chapter, we have presented a Bayesian approach to variable selection in multiple linear regression using summary statistics. Using simulated data, the results from our method are comparable with existing methods. Our method is more computationally efficient than existing methods (e.g. CAVIAR, FINEMAP). The method provides a list of independent *CSs*, which is useful for follow-up studies and understanding disease biology.

Our method uses the IBSS algorithm. The IBSS works well in most of our fine-mapping experiments, but we have observed it can get stuck in a local optimum in some difficult settings. We have introduced a refinement procedure for the IBSS algorithm to get out of the poor local optima, when there are significant findings using the default algorithm. This refinement procedure can help avoid false positives. However, it does not help with false negatives due to local optima. When there is no significant finding using the default IBSS algorithm (e.g. Wang et al., 2020, change point detection example), our refinement procedure does not change the result. In this case, a better initialization may help, for example, using solution from lasso or other variable selection methods. One could also develop better algorithm to optimize the variational objective function.

Our simulation results highlight the importance of having an accurate LD matrix. For all methods using summary statistics, the reliability of the fine-mapping results depends heavily on the LD information. The false discovery rate increases when the LD matrix is obtained from a small reference panel. Some methods are more robust to misspecification than others. The regularization for the reference LD matrix

$((1-s)\hat{\boldsymbol{R}}+s\boldsymbol{I})$ has slight improvement on the fine-mapping results. To better explore the public available GWAS summary statistics, it is crucial to estimate LD matrix from a large reference panel with similar ancestry to the GWAS sample, and it is better to share the GWAS LD information.

We assume fixed and known prior inclusion probabilities in *SuSiE-RSS*, but *SuSiE-RSS* can be easily extended to incorporate functional genomic annotation data. Previous studies have shown that integrating functional data improves fine-mapping performance (Kichaev et al., 2014; Wen et al., 2016a). We can allow the prior inclusion probabilities depend on functional data, for example,

$$\log\left(\frac{\Pr(\gamma_j = 1|\boldsymbol{\eta}, \boldsymbol{A})}{\Pr(\gamma_j = 0|\boldsymbol{\eta}, \boldsymbol{A})}\right) = \eta_0 + \sum_{m=1}^{M} \eta_m A_{jm}, \tag{3.7.1}$$

where $A_{jm} = 1$ when variant j is part of annotation m. The hyper-parameter η_0 captures the baseline prior log odds for causality of any variant, η_m characterizes the enrichment level of each genomic feature.

For binary phenotype, the justification for model (3.1.1) is based on z scores from two proportions tests with equal case-control samples (Han et al., 2009). Nonetheless, the marginal z scores for case-control study are usually from logistic regression (Chen et al., 2016). We are not aware a formal derivation of the model for z scores from logistic regression, although people do use univariate logistic regression z scores in their fine-mapping analysis. One direction for future work is to check the *SuSiE-RSS* performance using univariate logistic regression z scores. More ambitiously, one could derive the RSS likelihood based on generalized linear models (McCullagh and Nelder,

2019), using asymptotic theories of maximum likelihood estimator. The asymptotic normality of summary statistics holds in a reasonably balanced case-control study. However, because of the low prevalence of many diseases, case-control ratios are often unbalanced (case:control $<$ 1:10) in GWAS. The unbalanced case-control ratio violates asymptotic normal assumption for logistic regression and inflates Type I error (Zhou et al., 2018). One solution to address this is using test statistics with saddle point approximation (Kuonen, 1999), or Firth correction (Firth, 1993; Heinze and Schemper, 2002). Both methods provide good control of Type I error for rare binary traits (Zhou et al., 2018; Mbatchou et al., 2021).

To detect variants with modest genetic effects, meta-analysis of several GWASs has become a common method to increase the sample size and power (e.g. Willer et al., 2010; Lee et al., 2017). Meta-analysis combines individual-level data or summary statistics from different studies/populations. Using summary statistics from meta-analysis, one strategy to do fine-mapping would be applying *SuSiE-RSS* with sample size weighted correlation matrix. This strategy is simple, although we are not sure about the impact of deviations from our model using meta-analysis results. We have not investigated this in much detail. A more principled approach would be to do joint fine-mapping of multiple studies/ethnics/traits (Kichaev and Pasaniuc, 2015). We describe an approach to do multi-trait fine-mapping in Chapter 4.

CHAPTER 4

MULTI-TRAIT FINE-MAPPING [1]

Genome-wide association analyses have been performed for thousands of phenotypes and identified many genomic regions associated with complex traits (e.g. Canela-Xandri et al., 2018; Kanai et al., 2018). Many statistical fine-mapping methods have been developed to prioritize putative causal variants for a single phenotype (e.g. Guan and Stephens, 2011; Kichaev et al., 2014; Hormozdiari et al., 2014; Chen et al., 2015; Benner et al., 2016; Wen et al., 2016b; Newcombe et al., 2016; Lee et al., 2018; Wang et al., 2020). A simple strategy to do fine-mapping with multiple phenotypes is analyzing each phenotype separately, and then examining the overlap of results among phenotypes. However, this trait-by-trait analysis fails to leverage information across phenotypes to improve the power to detect the causal variants. To address deficiencies of trait-by-trait analysis, it is desirable to perform multi-phenotype fine-mapping. Multivariate analysis improves power when signals are shared among multiple phenotypes; it increases power even when signals are not shared, but the phenotypes are correlated (Stephens, 2013).

To date, there are few tools available for multi-trait fine-mapping because of the computational challenge in performing inference in the model. To our knowledge, the existing multi-trait fine-mapping methods using individual-level genotype and multiple phenotype data are *MT-HESS* (Lewin et al., 2016) and *atlasqtl* (Ruffieux et al., 2020), and the only existing multi-trait fine-mapping method using GWAS

1. THIS CHAPTER CONTAINS JOINT WORK WITH G. WANG, P. CARBONETTO AND M. STEPHENS. SEE SECTION 4.8.

summary data is *PAINTOR* (Kichaev et al., 2017). All these methods are based on Bayesian multivariate multiple regression. Nonetheless, the existing approaches are limited in practice for at least two reasons. First, the existing methods make restrictive assumptions on non-zero effects. They assume the effect variants have non-zero effects in all traits and the non-zero effects are uncorrelated among traits. However, the effects could be specific to a subset of traits, and some traits may be more correlated than others; for example, in our blood cell traits application (see Section 4.6), there are variants with non-zero effects specific to red blood cells and the effects in red blood cell traits are correlated. Second, the existing methods are computationally intensive. *MT-HESS* uses MCMC sampling for Bayesian inference. *PAINTOR* uses exhaustive search or Importance Sampling. Thus, *MT-HESS* and *PAINTOR* are computationally slow or even impossible for more than 6 phenotypes. *atlasqtl* uses variational approximation for posterior distribution, which makes it possible to handle a large number of phenotypes. However, *atlasqtl* uses "fully factorized" variational approximation, which is not suitable for highly correlated variables (see Carbonetto et al., 2012; Wang et al., 2020, for discussions).

In this chapter, we describe a generalization of *SuSiE* to do multivariate variable selection using individual-level genotype and phenotypes data, Multivariate Sum of Single Effects model (*mvSuSiE*). We also describe *mvSuSiE-suff* that fit the *mvSuSiE* model with sufficient statistics, which gives exactly the results obtained by applying *mvSuSiE* to individual-level data. We further develop method to fit the *mvSuSiE* model using summary statistics, *mvSuSiE-RSS*, which requires GWAS z scores for each phenotype and an estimated LD matrix. Our model efficiently per-

forms joint fine-mapping for multiple phenotypes while accounting for potentially complicated patterns of heterogeneous effect size across traits. It uses flexible priors allowing for arbitrary correlations in effect sizes among phenotypes (Urbut et al., 2019). It performs efficient posterior inference via variational approximation, but the parameterization is different from *atlasqtl*.

We describe the *mvSuSiE* model in Section 4.1. The fitting algorithm using sufficient statistics (*mvSuSiE-suff*) is described in Section 4.2. Section 4.3 describes the *mvSuSiE-RSS* model. Section 4.4 summarizes the posterior inference. Section 4.5 shows the performance of *mvSuSiE-RSS* using simulations. Section 4.6 illustrates the application of *mvSuSiE-RSS* to jointly fine-map 16 blood cell traits from UK Biobank. Section 4.7 concludes and discusses future works.

4.1 The *mvSuSiE* model

Suppose R quantitative phenotypes are observed for N individuals, the standard multivariate multiple regression model is

$$\boldsymbol{Y} \sim \mathrm{MN}_{N\times R}(\boldsymbol{X}\boldsymbol{B}, \boldsymbol{I}_N, \boldsymbol{V}), \tag{4.1.1}$$

where $\boldsymbol{Y} \in \mathbb{R}^{N\times R}$ denotes a matrix of observed responses for N samples across R phenotypes, $\boldsymbol{X} \in \mathbb{R}^{N\times J}$ denotes a matrix of genotype at J genetic variants observed in the N samples, $\boldsymbol{B} \in \mathbb{R}^{J\times R}$ denotes a matrix of regression coefficients for the J variants across R phenotypes, $\boldsymbol{I}_N$ is the $N \times N$ identity matrix, and $\mathrm{MN}_{N\times R}(\boldsymbol{M}, \boldsymbol{\Sigma}, \boldsymbol{V})$ denotes the matrix normal distribution with mean $\boldsymbol{M} \in \mathbb{R}^{N\times R}$, row covariances

$\boldsymbol{\Sigma} \in \mathbb{R}^{N \times N}$ and column covariances $\boldsymbol{V} \in \mathbb{R}^{R \times R}$. We assume column covariance matrix $\boldsymbol{V}$ is invertible. An intercept in the multivariate regression is accounted for by requiring that all the columns of $\boldsymbol{Y}$ and $\boldsymbol{X}$ are centered so that their means are zero.

As in *SuSiE* model, the *mvSuSiE* model is based on a simpler model, the "multivariate single effect regression" (*MSER*) model. So we describe the *MSER* model first.

4.1.1 The Multivariate Single Effect Regression model

The *MSER* model generalizes the "single effect regression" (*SER*) model to multivariate context. We write the coefficients $\boldsymbol{B}$ as

$$\boldsymbol{B} = \boldsymbol{\gamma} \otimes \boldsymbol{b}. \tag{4.1.2}$$

Here, $\boldsymbol{x} \otimes \boldsymbol{y}$ denotes Kronecker product of vectors $\boldsymbol{x}$ and $\boldsymbol{y}$, $\boldsymbol{\gamma} \in \{0, 1\}^J$ is a vector of indicator variables in which exactly one of the J elements is one and the remaining are zero, and $\boldsymbol{b} \in \mathbb{R}^R$ is the vector of regression coefficients for the R phenotypes. By this definition, the coefficient matrix $\boldsymbol{B} \in \mathbb{R}^{J \times R}$ has a single row containing non-zero values, and these non-zero values are determined by $\boldsymbol{b}$. We refer to $\boldsymbol{B}$ as a "single effect matrix" because it captures the effects of a single variant.

As in *SER*, the priors for the indicator variables $\boldsymbol{\gamma}$ and regression coefficients $\boldsymbol{b}$

are

$$\boldsymbol{\gamma} \sim \text{Mult}(1, \boldsymbol{\pi}) \tag{4.1.3}$$

$$\boldsymbol{b} \sim g, \tag{4.1.4}$$

where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_J)$ is a vector of prior inclusion probabilities ($\pi_j \geqslant 0$, $\sum_{j=1}^{J} \pi_j = 1$).

In multivariate analysis, $\boldsymbol{b}$ represents effects for R phenotypes, and we want to allow both shared and trait-specific effects. A natural method would be introducing some indicators for which traits have non-zero effect, but this also introduce computational complexity. To flexibly capture the heterogeneity and correlations among effects for different traits, we take the prior distribution of the coefficients, g, to be a mixture of multivariate normals,

$$g(\boldsymbol{b}) = \sum_{k=1}^{K} \omega_k N_R(\boldsymbol{b}; \mathbf{0}, \sigma_0^2 \boldsymbol{U}_k), \tag{4.1.5}$$

in which each $\boldsymbol{U}_k$ is an $R \times R$ covariance matrix that captures one pattern of effects, and $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_K)$ are mixture proportions ($\omega_k \geqslant 0$, $\sum_{k=1}^{K} \omega_k = 1$). The hyperparameter σ_0^2 controls the prior scale of the single effect, and it can be estimated using an empirical Bayes procedure. We assess the significance of the signal for each trait using the posterior on the effect sizes (see Section 4.4).

In summary, the *MSER* model is

$$\boldsymbol{Y} \sim \mathrm{MN}_{N\times R}(\boldsymbol{XB}, \boldsymbol{I}_N, \boldsymbol{V}) \tag{4.1.6}$$

$$\boldsymbol{B} = \boldsymbol{\gamma} \otimes \boldsymbol{b}, \tag{4.1.7}$$

$$\boldsymbol{\gamma} \sim \mathrm{Mult}(1, \boldsymbol{\pi}), \tag{4.1.8}$$

$$\boldsymbol{b} \sim \sum_{k=1}^{K} \omega_k N_R(\boldsymbol{0}, \sigma_0^2 \boldsymbol{U}_k) \tag{4.1.9}$$

We assume $\boldsymbol{V}$, $\boldsymbol{\pi}$, $\boldsymbol{\omega}$, $\mathcal{U} = \{\boldsymbol{U}_1, \cdots, \boldsymbol{U}_K\}$ are known, or have been estimated previously. For instance, the prior inclusion probabilities $\boldsymbol{\pi}$ can be fixed as uniform among variants, $\boldsymbol{\pi} = (1/J, \cdots, 1/J)$; the residual variance $\boldsymbol{V}$ can be estimated as the empirical covariance matrix of $\boldsymbol{Y}$, which is a "conservative" estimates under the null ($\boldsymbol{B} = \boldsymbol{0}$); the parameters $\boldsymbol{\omega}$ and $\mathcal{U}$ in the prior mixture can be fixed to the so-called *canonical mixture* (Flutre et al., 2013), or they can be estimated using statistical procedures (Section 4.3.4 and Urbut et al. (2019)).

Posterior under the *MSER* model

To derive posterior computations for the *MSER* model, it helps to start with Bayesian simple multivariate regression (*BMR*) model,

$$\boldsymbol{Y} \sim \mathrm{MN}_{N\times R}(\boldsymbol{x}\boldsymbol{b}^\intercal, \boldsymbol{I}_N, \boldsymbol{V}), \tag{4.1.10}$$

$$\boldsymbol{b} \sim \sum_{k=1}^{K} \omega_k N_R(\boldsymbol{0}, \sigma_0^2 \boldsymbol{U}_k). \tag{4.1.11}$$

Here, $\boldsymbol{x} \in \mathbb{R}^N$ is a vector of genotypes for one variant, $\boldsymbol{b} \in \mathbb{R}^R$ is the (unknown) vector of regression for the R phenotypes. The posterior distribution on $\boldsymbol{b}$ can be written using least-squares estimate of $\boldsymbol{b}$, denoted $\hat{\boldsymbol{b}}$, and its variance-covariance matrix, $\boldsymbol{S}$,

$$\hat{\boldsymbol{b}} = (\boldsymbol{x}^\intercal \boldsymbol{x})^{-1} \boldsymbol{Y}^\intercal \boldsymbol{x} \tag{4.1.12}$$

$$\boldsymbol{S} = (\boldsymbol{x}^\intercal \boldsymbol{x})^{-1} \boldsymbol{V}. \tag{4.1.13}$$

The posterior distribution for $\boldsymbol{b}$ is summarized in Proposition 4.

Under the *MSER* model, the posterior distribution on $\boldsymbol{\gamma}$ and $\boldsymbol{B}$ is summarized in Proposition 5.

Proposition 4. *Consider the BMR model with known $\boldsymbol{V}$, $\boldsymbol{\omega}$, $\mathcal{U}$ and σ_0^2. The posterior distribution for $\boldsymbol{b}$ is*

$$p(\boldsymbol{b}|\boldsymbol{Y}, \boldsymbol{x}, \boldsymbol{V}, \boldsymbol{\omega}, \mathcal{U}, \sigma_0^2) = \sum_{k=1}^{K} \omega_{1k} N_R(\boldsymbol{b}; \boldsymbol{\mu}_{1k}, \boldsymbol{\Sigma}_{1k}), \tag{4.1.14}$$

where

$$\boldsymbol{\Sigma}_{1k}(\boldsymbol{x}) := \sigma_0^2 \boldsymbol{U}_k (\boldsymbol{I} + \sigma_0^2 \boldsymbol{S}^{-1} \boldsymbol{U}_k)^{-1} \tag{4.1.15}$$

$$\boldsymbol{\mu}_{1k}(\boldsymbol{x}) := \boldsymbol{\Sigma}_{1k}(\boldsymbol{x}) \boldsymbol{S}^{-1} \hat{\boldsymbol{b}} \tag{4.1.16}$$

$$\omega_{1k} := \frac{\omega_k BF(\boldsymbol{Y}, \boldsymbol{x}; \boldsymbol{V}, \boldsymbol{U}_k, \sigma_0^2)}{\sum_{k=1}^{K} \omega_k BF(\boldsymbol{Y}, \boldsymbol{x}; \boldsymbol{V}, \boldsymbol{U}_k, \sigma_0^2)} \tag{4.1.17}$$

$$BF(\boldsymbol{Y}, \boldsymbol{x}; \boldsymbol{V}, \boldsymbol{U}_k, \sigma_0^2) := \frac{N_R(\hat{\boldsymbol{b}}; \boldsymbol{0}, \boldsymbol{S} + \sigma_0^2 \boldsymbol{U}_k)}{N_R(\hat{\boldsymbol{b}}; \boldsymbol{0}, \boldsymbol{S})}. \tag{4.1.18}$$

The $BF(\boldsymbol{Y}, \boldsymbol{x}; \boldsymbol{V}, \boldsymbol{U}_k, \sigma_0^2)$ (4.1.18) is the Bayes Factor for variant with non-zero ef-

fects from $N_R(\mathbf{0}, \sigma_0^2 \boldsymbol{U}_k)$.

The posterior first and second moments of $\boldsymbol{b}$ *are*

$$\boldsymbol{\mu}_1^{\mathsf{mix}}(\boldsymbol{Y}, \boldsymbol{x}; \boldsymbol{V}, \boldsymbol{\omega}, \mathcal{U}, \sigma_0^2) := \sum_{k=1}^{K} \omega_{1k} \boldsymbol{\mu}_{1k}(\boldsymbol{x}), \tag{4.1.19}$$

$$\boldsymbol{\mu}_2^{\mathsf{mix}}(\boldsymbol{Y}, \boldsymbol{x}; \boldsymbol{V}, \boldsymbol{\omega}, \mathcal{U}, \sigma_0^2) := \sum_{k=1}^{K} \omega_{1k} \left[\boldsymbol{\mu}_{1k}(\boldsymbol{x}) \boldsymbol{\mu}_{1k}(\boldsymbol{x})^{\mathsf{T}} + \boldsymbol{\Sigma}_{1k}(\boldsymbol{x}) \right]. \tag{4.1.20}$$

The Bayes Factor for comparing the model with the null model ($\boldsymbol{b} = \mathbf{0}$*) is given by*

$$BF^{\mathsf{mix}}(\boldsymbol{Y}, \boldsymbol{x}; \boldsymbol{V}, \boldsymbol{\omega}, \mathcal{U}, \sigma_0^2) = \sum_{k=1}^{K} \omega_k BF(\boldsymbol{Y}, \boldsymbol{x}; \boldsymbol{V}, \boldsymbol{U}_k, \sigma_0^2). \tag{4.1.21}$$

Proposition 5. *Under the MSER model with known* $\boldsymbol{V}$*,* $\boldsymbol{\omega}$*,* $\mathcal{U}$ *and* σ_0^2*, the posterior distribution for* $\boldsymbol{\gamma}$ *is*

$$\boldsymbol{\gamma} | \boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{V}, \boldsymbol{\omega}, \mathcal{U}, \sigma_0^2 \sim \mathit{Mult}(1, \boldsymbol{\alpha}), \tag{4.1.22}$$

where

$$\alpha_j = \frac{\pi_j BF^{\mathsf{mix}}(\boldsymbol{Y}, \boldsymbol{x}_j; \boldsymbol{V}, \boldsymbol{\omega}, \mathcal{U}, \sigma_0^2)}{\sum_{j=1}^{J} \pi_j BF^{\mathsf{mix}}(\boldsymbol{Y}, \boldsymbol{x}_j; \boldsymbol{V}, \boldsymbol{\omega}, \mathcal{U}, \sigma_0^2)}, \tag{4.1.23}$$

with BF^{mix} *given by* (4.1.21)*. The posterior first and second moments of* $\boldsymbol{b}_j$ *(the* j*th row of* $\boldsymbol{B}$*) are*

$$\mathbb{E}[\boldsymbol{b}_j] = \alpha_j \boldsymbol{\mu}_1^{\mathsf{mix}}(\boldsymbol{Y}, \boldsymbol{x}_j; \boldsymbol{V}, \boldsymbol{\omega}, \mathcal{U}, \sigma_0^2) \tag{4.1.24}$$

$$\mathbb{E}[\boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{T}}] = \alpha_j \boldsymbol{\mu}_2^{\mathsf{mix}}(\boldsymbol{Y}, \boldsymbol{x}_j; \boldsymbol{V}, \boldsymbol{\omega}, \mathcal{U}, \sigma_0^2), \tag{4.1.25}$$

where $\boldsymbol{\mu}_1^{\mathsf{mix}}$ *and* $\boldsymbol{\mu}_2^{\mathsf{mix}}$ *are the posterior first and second moments from the BMR model, given at* (4.1.19) *and* (4.1.20).

Estimating Prior Scalar

We estimate the prior scalar σ_0^2 using an expectation-maximization approach (Dempster et al., 1977). The update for σ_0^2 is

$$\sigma_0^2 = \sum_{k=1}^{K} \phi_k \frac{1}{\sum_{k'=1}^{K} \phi_{k'} \mathrm{rank}(\boldsymbol{U}_{k'})} \sum_{j=1}^{J} \alpha_j \mathrm{tr}(\boldsymbol{U}_k^{\dagger} \boldsymbol{\mu}_{jk}^2), \tag{4.1.26}$$

where $\boldsymbol{U}_k^{\dagger}$ is the Moore–Penrose inverse of $\boldsymbol{U}_k$, $\boldsymbol{\mu}_{jk}^2$ is the posterior second moment for variant j with $\boldsymbol{U}_k$ as the prior (i.e. $\boldsymbol{\mu}_{jk}^2 = \boldsymbol{\mu}_{1k}(\boldsymbol{x}_j)\boldsymbol{\mu}_{1k}(\boldsymbol{x}_j)^{\mathsf{T}} + \boldsymbol{\Sigma}_{1k}(\boldsymbol{x}_j)$), ϕ_k is the posterior weight of component k based on single effect regression,

$$\phi_k := \frac{\omega_k \sum_{j=1}^{J} \mathrm{BF}(\boldsymbol{Y}, \boldsymbol{x}_j; \boldsymbol{V}, \boldsymbol{U}_k, \sigma_0^2)}{\sum_{k'=1}^{K} \omega_{k'} \sum_{j=1}^{J} \mathrm{BF}(\boldsymbol{Y}, \boldsymbol{x}_j; \boldsymbol{V}, \boldsymbol{U}_{k'}, \sigma_0^2)}. \tag{4.1.27}$$

4.1.2 The Multivariate Sum of Single Effects Regression model

To allow multiple effect variants, we parameterize coefficients $\boldsymbol{B}$ as the sum of L "single effect matrices". The *mvSuSiE* model is

$$\boldsymbol{Y} \sim \mathrm{MN}_{N\times R}(\boldsymbol{XB}, \boldsymbol{I}_N, \boldsymbol{V}) \tag{4.1.28}$$

$$\boldsymbol{B} = \sum_{l=1}^{L} \boldsymbol{B}^l \tag{4.1.29}$$

$$\boldsymbol{B}^l = \boldsymbol{\gamma}^l \otimes \boldsymbol{b}^l \tag{4.1.30}$$

$$\boldsymbol{\gamma}^l \sim \mathrm{Mult}(1, \boldsymbol{\pi}) \tag{4.1.31}$$

$$\boldsymbol{b}^l \sim \sum_{k=1}^{K} \omega_k N_R(\boldsymbol{b}; \boldsymbol{0}, \sigma_{0l}^2 \boldsymbol{U}_k). \tag{4.1.32}$$

The coefficients matrix $\boldsymbol{B}$ has at most L rows containing non-zero elements. Each σ_{0l}^2 controls the prior scale of the l-th single effect, thus the single effects could on different scales.

We extend the *mvSuSiE* model (4.1.28) - (4.1.32) to allow for missing values in $\boldsymbol{Y}$, see Appendix C.1 and C.2 for details. For simplicity, we consider $\boldsymbol{Y}$ without missing values in this chapter.

4.2 *mvSuSiE-suff*: *mvSuSiE* using sufficient statistics

Wang et al. (2020), Appendix B, described a general variational approach to fit additive effects models. The *mvSuSiE* model (4.1.28) – (4.1.32) is an additive effects model, with each additive effect being a multivariate single-effect regression (Section

4.1.1). Therefore applying the results from Wang et al. (2020) immediately produces an IBSS algorithm for fitting the *mvSuSiE* model from the individual-level data $\{\boldsymbol{Y}, \boldsymbol{X}\}$. Analogous to the result for univariate *SuSiE*, this IBSS algorithm finds an approximation $q(\boldsymbol{B}^1, \cdots, \boldsymbol{B}^L) = \prod_{l=1}^{L} q(\boldsymbol{B}^l)$ to the posterior distribution $p_{\text{post}} = p(\boldsymbol{B}^1, \cdots, \boldsymbol{B}^L | \boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\omega}, \mathcal{U}, \boldsymbol{\sigma}_0^{\mathbf{2}})$ by minimizing the Kullback-Leibler (KL) divergence from q to p_{post}.

As with univariate *SuSiE*, one can also fit the *mvSuSiE* model using sufficient statistics, as we now describe. The sufficient statistics for $\boldsymbol{B}$ of the multivariate multiple regression model (4.1.1) are $\boldsymbol{X}^\intercal\boldsymbol{X}$, $\boldsymbol{X}^\intercal\boldsymbol{Y}$, which is clear from the likelihood for $\boldsymbol{B}$,

$$\begin{aligned} &L(\boldsymbol{B}; \boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{V}) \coloneqq p(\boldsymbol{Y} | \boldsymbol{X}, \boldsymbol{B}, \boldsymbol{V}) \\ &= |2\pi\boldsymbol{V}|^{-N/2} \exp\left(-\frac{1}{2}\text{tr}\left[\boldsymbol{V}^{-1}(\boldsymbol{Y}^\intercal\boldsymbol{Y} - 2\boldsymbol{B}^\intercal\boldsymbol{X}^\intercal\boldsymbol{Y} + \boldsymbol{B}^\intercal\boldsymbol{X}^\intercal\boldsymbol{X}\boldsymbol{B})\right]\right). \end{aligned} \tag{4.2.1}$$

Algorithm 3 outlines the IBSS algorithm for *mvSuSiE* using sufficient statistics. The main building block for this algorithm is a function, *MSER-suff*, that takes input as sufficient statistics and returns the posterior distribution for $\boldsymbol{B} = \boldsymbol{\gamma} \otimes \boldsymbol{b}$ under the *MSER* model. That is,

$$\textit{MSER-suff}(\boldsymbol{X}^\intercal\boldsymbol{X}, \boldsymbol{X}^\intercal\boldsymbol{Y}; \sigma_{0l}^2, \mathcal{U}, \boldsymbol{V}, \boldsymbol{\omega}) \coloneqq (\boldsymbol{\alpha}, [\boldsymbol{\mu}], [\boldsymbol{\mu}^2]), \tag{4.2.2}$$

where the vector $\boldsymbol{\alpha}$ gives the posterior inclusion probabilities under *MSER* (4.1.23); the matrix $[\boldsymbol{\mu}]$ is the posterior mean of $\boldsymbol{B}$ (4.1.24); the array $[\boldsymbol{\mu}^2]$ contains the posterior second moment of $\boldsymbol{B}$ (4.1.25).

In summary, just as with univariate *SuSiE*, we have two IBSS algorithms for fitting the multivariate model, one (*mvSuSiE*) based on the individual-level data $\{\boldsymbol{Y}, \boldsymbol{X}\}$ and one (*mvSuSiE-suff*) based on the sufficient statistics $\boldsymbol{X}^\intercal\boldsymbol{X}$, $\boldsymbol{X}^\intercal\boldsymbol{Y}$. The two algorithms will give the same result when the sufficient statistics are correctly computed using the column-centered individual-level data $\{\boldsymbol{Y}, \boldsymbol{X}\}$. However, the computational complexity of the two algorithms differs. In our current implementation, when estimating the prior scalar parameters $\boldsymbol{\sigma_0^2}$, the computational complexity is $O(L(NJR + KJR^3))$ per iteration using individual-level data, and $O(L(J^2R + KJR^3))$ per iteration using sufficient statistics. With fixed prior scalar parameters $\boldsymbol{\sigma_0^2}$, we can precompute the matrix inversions involved in the posterior distribution (4.1.15), which reduces the computational complexity to $O(L(NJR + KJR^2))$ per iteration using individual-level data, and $O(L(J^2R + KJR^2))$ per iteration using sufficient statistics. If $\boldsymbol{X}$ is column-standardized, we can remove the J multiplier before R^2 or R^3 in the computational complexity, because the $\boldsymbol{S}$ (4.1.13) is same for all variants.

Algorithm 3 IBSS algorithm for *mvSuSiE* using sufficient statistics

Require: Sufficient Statistics $\boldsymbol{X}^\intercal\boldsymbol{X}, \boldsymbol{X}^\intercal\boldsymbol{Y}$.

Require: Number of effects, L; priors, $\boldsymbol{\omega}$, $\mathcal{U}$; residual covariance, $\boldsymbol{V}$ and initial estimates of $\boldsymbol{\sigma_0^2}$.

Require: A function *MSER-suff*$(\boldsymbol{X}^\intercal\boldsymbol{X}, \boldsymbol{X}^\intercal\boldsymbol{Y}; \sigma_{0l}^2, \mathcal{U}, \boldsymbol{V}, \boldsymbol{\omega}) \rightarrow (\boldsymbol{\alpha}, [\boldsymbol{\mu}], [\boldsymbol{\mu}^2])$ that computes the posterior distribution for $\boldsymbol{B}^l$ under the *MSER* model.

Initialize posterior means $\bar{\boldsymbol{B}}^l = 0$, for $l = 1, \ldots, L$.

repeat

  for l in $1, \ldots, L$ **do**

    $\boldsymbol{U} \leftarrow \boldsymbol{X}^\intercal\boldsymbol{Y} - \boldsymbol{X}^\intercal\boldsymbol{X} \sum_{l' \neq l} \bar{\boldsymbol{B}}^{l'}$

    $(\boldsymbol{\alpha}^l, [\boldsymbol{\mu}]^l, [\boldsymbol{\mu}^2]^l) \leftarrow$ *MSER-suff*$(\boldsymbol{X}^\intercal\boldsymbol{X}, \boldsymbol{U}; \sigma_{0l}^2, \mathcal{U}, \boldsymbol{V}, \boldsymbol{\omega})$

    $\bar{\boldsymbol{B}}^l \leftarrow \boldsymbol{\alpha}^l \circ [\boldsymbol{\mu}]^l$ ▷ compute posterior mean by multiplying elements of $\boldsymbol{\alpha}^l$ to rows of $[\boldsymbol{\mu}]^l$

    $\sigma_{0l}^2 \leftarrow$ (4.1.26) ▷ Update σ_{0l}^2 (optional).

until convergence criterion satisfied

return $\boldsymbol{\alpha}^1, [\boldsymbol{\mu}]^1, [\boldsymbol{\mu}^2]^1, \ldots, \boldsymbol{\alpha}^L, [\boldsymbol{\mu}]^L, [\boldsymbol{\mu}^2]^L$.

4.3 The *mvSuSiE-RSS* model

As with univariate *SuSiE*, we develop a method to fit the *mvSuSiE* model to summary data. We assume all phenotype measurements have been performed in a single sample, so the marginal z scores are computed using the same samples for all phenotypes, and the correlations among variants are same for all phenotypes. We assume that we have an estimate $\hat{\boldsymbol{R}}$ of the LD matrix between the J variants. In addition, we assume

we have access to the marginal z scores from linear models between each variant and each phenotype. Let $\hat{\boldsymbol{Z}}$ denotes the $J \times R$ matrix of observed marginal z scores, each element $\hat{z}_{jr}$ is defined as (3.1.3). To simplify notation, we assume the genotypes for each variant are standardized, that is, $\boldsymbol{x}_j^\intercal \boldsymbol{x}_j = N$. Assuming the correlation between phenotype and any single variant is small, the observed marginal z scores are approximately $\hat{\boldsymbol{Z}} \approx \frac{1}{\sqrt{N}} \boldsymbol{X}^\intercal \boldsymbol{Y} \boldsymbol{S}^{-1}$, where $\boldsymbol{S}^2 \coloneqq \text{diag}(\text{Var}(\boldsymbol{y}_1), \cdots, \text{Var}(\boldsymbol{y}_R))$, and Var(.) denotes the sample variance.

Consider the multivariate multiple linear regression model (4.1.1) with $\boldsymbol{V} = \text{Var}(\boldsymbol{Y})$. We obtain the model for $\hat{\boldsymbol{Z}}$ by multiplying (4.1.1) by $\frac{1}{\sqrt{N}} \boldsymbol{X}^\intercal$ and $\boldsymbol{S}^{-1}$,

$$\hat{\boldsymbol{Z}} \sim \text{MN}_{J \times R}(\hat{\boldsymbol{R}} \boldsymbol{Z}, \hat{\boldsymbol{R}}, \boldsymbol{C}), \tag{4.3.1}$$

where $\boldsymbol{Z} = \sqrt{N} \boldsymbol{B} \boldsymbol{S}^{-1}$ represent an unobserved $J \times R$ matrix of standardized true effects; $\hat{\boldsymbol{R}} = \frac{1}{N} \boldsymbol{X}^\intercal \boldsymbol{X}$ is the sample correlation matrix among variants; $\boldsymbol{C} = \boldsymbol{S}^{-1} \boldsymbol{V} \boldsymbol{S}^{-1}$ is a correlation matrix that accounts for correlations among the measurements in the R phenotypes. With only one phenotype, the model (4.3.1) reduces to the well-known model for z scores for one phenotype (3.1.1) (e.g. Kichaev et al., 2014; Hormozdiari et al., 2014; Chen et al., 2015; Benner et al., 2016; Zhu and Stephens, 2017).

Provided that the LD matrix $\hat{\boldsymbol{R}}$ is invertible, the density for $\hat{\boldsymbol{Z}}$ leads to the RSS likelihood for $\boldsymbol{Z}$,

$$L(\boldsymbol{Z}; \hat{\boldsymbol{Z}}, \hat{\boldsymbol{R}}, \boldsymbol{C}) \coloneqq \exp\Big\{ -\frac{1}{2} \text{tr}\left[\boldsymbol{C}^{-1} \left(\boldsymbol{Z}^\intercal \hat{\boldsymbol{R}} \boldsymbol{Z} - 2 \boldsymbol{Z}^\intercal \hat{\boldsymbol{Z}} \right) \right] \Big\}. \tag{4.3.2}$$

For the reasons we have discussed in Section 3.2, we use the RSS likelihood (4.3.2)

even when $\hat{\boldsymbol{R}}$ is non-invertible.

We derive the multi-trait model for summary statistics (4.3.1) using the usual marginal z scores from simple linear regressions. The model will hold approximately even if $\boldsymbol{Y}$ is not normally distributed, e.g. $\boldsymbol{Y}$ is binary traits. It seems natural to expect that a similar model could be derived for summary statistics from logistic regressions, but we have not done this.

4.3.1 The Multivariate Single Effect Regression using Summary Statistics

Analogous to *mvSuSiE*, the building block for our approach is the "multivariate single effect regression using summary statistics" model (*MSER-RSS*), in which exactly one of the J variants has a non-zero effect in some phenotypes. The *MSER-RSS* model is

$$\hat{\boldsymbol{Z}} \sim \mathrm{MN}_{J\times R}(\hat{\boldsymbol{R}}\boldsymbol{Z}, \hat{\boldsymbol{R}}, \boldsymbol{C}) \tag{4.3.3}$$

$$\boldsymbol{Z} = \boldsymbol{\gamma} \otimes \boldsymbol{z} \tag{4.3.4}$$

$$\boldsymbol{\gamma} \sim \mathrm{Mult}(1, \boldsymbol{\pi}) \tag{4.3.5}$$

$$\boldsymbol{z} \sim \sum_{k=1}^{K} \omega_k N(\boldsymbol{0}, \sigma_0^2 \boldsymbol{U}_k). \tag{4.3.6}$$

The "single effect matrix" $\boldsymbol{Z} \in \mathbb{R}^{J\times R}$ has exactly one non-zero row, whose elements are given by $\boldsymbol{z}$. We notice that the RSS likelihood (4.3.2) is a special case of the likelihood (4.2.1) with $\boldsymbol{X}^\intercal\boldsymbol{X} = \hat{\boldsymbol{R}}$, $\boldsymbol{X}^\intercal\boldsymbol{Y} = \hat{\boldsymbol{Z}}$, $\boldsymbol{V} = \boldsymbol{C}$. Therefore, we can apply

the posterior computation derived in Proposition 4 and 5 by making use of $\hat{\boldsymbol{b}}_j = \hat{\boldsymbol{z}}_j$ (the j-th row of $\hat{\boldsymbol{Z}}$), $\hat{\boldsymbol{S}} = \boldsymbol{C}$. As in univariate *SER-RSS*, the posterior computation is independent of the LD matrix.

4.3.2 The Multivariate Sum of Single Effects Regression model using Summary Statistics

Similar to *mvSuSiE*, we parameterize the effects matrix $\boldsymbol{Z}$ as a sum of "single effect matrices" to allow multiple effect variants. The *mvSuSiE-RSS* model is

$$\hat{\boldsymbol{Z}} \sim \mathrm{MN}_{J\times R}(\hat{\boldsymbol{R}}\boldsymbol{Z}, \hat{\boldsymbol{R}}, \boldsymbol{C}) \tag{4.3.7}$$

$$\boldsymbol{Z} = \sum_{l=1}^{L} \boldsymbol{Z}^l \tag{4.3.8}$$

$$\boldsymbol{Z}^l = \boldsymbol{\gamma}^l \otimes \boldsymbol{z}^l \tag{4.3.9}$$

$$\boldsymbol{\gamma}^l \sim \mathrm{Mult}(1, \boldsymbol{\pi}) \tag{4.3.10}$$

$$\boldsymbol{z}^l \sim \sum_{k=1}^{K} \omega_k N(\boldsymbol{0}, \sigma_{0l}^2 \boldsymbol{U}_k). \tag{4.3.11}$$

The true standardized effect matrix $\boldsymbol{Z}$ has at most L rows containing non-zero elements.

We assume $\boldsymbol{\pi}$, $\boldsymbol{C}$, $\boldsymbol{\omega}$ and $\mathcal{U}$ are known prior to *mvSuSiE-RSS* model fitting. We fit the *mvSuSiE-RSS* model using the IBSS algorithm in Algorithm 3 with inputs $\boldsymbol{X}^\intercal\boldsymbol{X} = \hat{\boldsymbol{R}}$, $\boldsymbol{X}^\intercal\boldsymbol{Y} = \hat{\boldsymbol{Z}}$ and $\boldsymbol{V} = \boldsymbol{C}$, because the RSS likelihood (4.3.2) is a special case of the likelihood (4.2.1).

4.3.3 Estimating the residual correlation matrix

The $R \times R$ correlation matrix $\boldsymbol{C}$ accounts for residual correlations among measurements in R phenotypes. The measured z scores could be correlated across R phenotypes because there are uncorrected environmental effects in GWAS pre-processing steps. Failing to capture the residual correlations in $\boldsymbol{C}$ creates false discoveries because the model attempts to match the residual correlations using effect covariance matrices. The increase in false discoveries is clear from our simulations in Section 4.5.

If the phenotype matrix $\boldsymbol{Y}$ is available, we estimate $\boldsymbol{C}$ using the empirical correlation matrix of phenotypes after removing the covariates effects. If the phenotype matrix $\boldsymbol{Y}$ is unavailable, we estimate this correlation matrix using the fact that $\boldsymbol{C}$ is the correlation matrix of the z scores under the null ($\boldsymbol{Z} = \boldsymbol{0}$). Suppose there are P fine-mapping regions in total, let $\hat{\boldsymbol{Z}}^p$ denotes the observed z scores in region p. For each fine-mapping region p, we identify the variants that close to null, i.e. the variants with (absolute) z score < 2 in all phenotypes, $\mathcal{N}_p := \{j : \max_r |\hat{z}^p_{jr}| < 2\}$. The variants in $\mathcal{N}_p$ could be highly correlated due to LD. Thus, we randomly select a small number of variants from $\mathcal{N}_p$, $\mathcal{I}_p \subseteq \mathcal{N}_p$. By pooling $\mathcal{I}_p$ across all fine-mapping regions, we obtain a list of variants that are close to null and nearly independent. We estimate $\boldsymbol{C}$ as the empirical correlation matrix of the z scores for the pooled variants,

$$\boldsymbol{C} = \frac{1}{\sum_{p=1}^{P} |\mathcal{I}_p|} \sum_{p=1}^{P} \sum_{j \in \mathcal{I}_p} \hat{\boldsymbol{z}}^p_j \hat{\boldsymbol{z}}^{p\intercal}_j, \tag{4.3.12}$$

where $\hat{\boldsymbol{z}}^p_j$ is the j-th row of $\hat{\boldsymbol{Z}}^p$.

4.3.4 Generating prior covariance matrices

By default, $\boldsymbol{\omega}$ is fixed to uniform weight and $\mathcal{U}$ is fixed to the canonical covariance matrices, which are described in detail in the Section "Generate canonical covariance matrices $\boldsymbol{U}_k$" in Urbut et al. (2019). The canonical covariance matrices include the identity matrix (representing independent effects), a matrix of all ones (representing equal effects in all conditions), matrices representing trait-specific effects, etc.

The prior covariance matrices can also include data-driven matrices. We use a strategy similar to Urbut et al. (2019) to generate data-driven covariance matrices. We first identify the top SNP for each fine-mapping region, which is the SNP with the highest value of $\hat{z}_j^{p,\max} = \max_{1\leqslant r\leqslant R} |\hat{z}_{jr}^p|$. Let $\tilde{\boldsymbol{Z}}$ denotes the $P \times R$ matrix of z scores for top SNPs, P is the number of fine-mapping regions. To extract the main patterns in $\tilde{\boldsymbol{Z}}$, we fit a mixture of multivariate normal distributions to $\tilde{\boldsymbol{Z}}$ using the Extreme Deconvolution (ED) algorithm from Bovy et al. (2011) and obtain estimates of $\mathcal{U}$ and $\boldsymbol{\omega}$. We initialize the ED algorithm using data-driven covariance matrices based on factors from Principal Component Analysis and sparse matrix factorization (FLASH, Wang and Stephens (2018)) on $\tilde{\boldsymbol{Z}}$.

4.4 Posterior inference

Based on the posterior distribution of $\boldsymbol{B}$ or $\boldsymbol{Z}$, *mvSuSiE*, *mvSuSiE-suff* and *mvSuSiE-RSS* compute the posterior inclusion probability (PIP) for each SNP. We provide formulae in terms of *mvSuSiE-RSS* model, the formulae for *mvSuSiE* and

mvSuSiE-suff are similar. The PIP for each variant is

$$\text{PIP}_j \coloneqq P(\boldsymbol{z}_j \neq \mathbf{0}|\hat{\boldsymbol{Z}}, \hat{\boldsymbol{R}}) = 1 - \prod_{l=1}^{L}(1 - \alpha_j^l), \tag{4.4.1}$$

where α_j^l is the PIP of SNP j in the l-th single effect, $\alpha_j^l = P(\boldsymbol{\gamma}_j^l = 1|\hat{\boldsymbol{Z}}, \hat{\boldsymbol{R}}, \boldsymbol{C})$. The PIP_j gives the probability of SNP j having non-zero effect in at least one of the traits.

Similar to univariate *SuSiE*/*SuSiE-RSS*, a Credible Set (*CS*) for the l-th single effect can be computed as $CS(\boldsymbol{\alpha}^l; \rho)$ (2.1.14). The $CS(\boldsymbol{\alpha}^l; \rho)$ gives a subset of SNPs that has probability $\geqslant \rho$ of containing one SNP with non-zero effect in at least one phenotype. However, the $CS(\boldsymbol{\alpha}^l; \rho)$ does not contain information on whether the SNPs have non-zero effects in a specific phenotype. For this reason, we assess the significance of $CS(\boldsymbol{\alpha}^l; \rho)$ for each phenotype r. We first define the conditional local false sign rate (*clfsr*) for variant j in single effect l, trait r as

$$\mathit{clfsr}_{jr}^l \coloneqq 1 - \max[p_{\text{post}}(z_{jr}^l > 0|\gamma_j^l = 1), p_{\text{post}}(z_{jr}^l < 0|\gamma_j^l = 1)], \tag{4.4.2}$$

where p_{post} is the posterior distribution. The *clfsr* (Stephens, 2017) measures how confident we can be in the sign of variant j in single effect l, trait r given that variant j has non-zero effect. The significance of the l-th single effect credible set $CS(\boldsymbol{\alpha}^l; \rho)$ in trait r is then summarized using the weighted mean of *clfsr* across variants,

$$\mathit{lfsr}_r^l \coloneqq \sum_j \alpha_j^l \, \mathit{clfsr}_{jr}^l. \tag{4.4.3}$$

For each single effect, the Credible Set $CS(\boldsymbol{\alpha}^l; \rho)$ has a "significance level" ($lfsr_r^l$) in each phenotype r. Therefore, for each phenotype, we have the corresponding significant *CSs*.

The multi-trait fine-mapping methods *mvSuSiE*, *mvSuSiE-suff* and *mvSuSiE-RSS* are available at `https://github.com/stephenslab/mvsusieR`.

4.5 Numerical Experiments

To investigate the performance of *mvSuSiE-RSS*, we performed simulations using real genotype data from 248,980 UK Biobank unrelated White British individuals. We randomly selected 600 non-overlapping regions. The regions vary from 400-Kb to 1.6-Mb; each region contains 1,000 to 5,000 SNPs (including imputed SNPs) with MAF (minor allele frequency) > 0.001 and INFO (imputation quality score) > 0.6. The details about samples and regions are in Section 4.6. For each region, $\boldsymbol{X}$ is a matrix of column standardized genotype data, in which each row corresponds to an individual, and each column corresponds to a genetic variant. Standardizing the genotype data corresponds to assuming the SNPs with lower MAF have larger effects in the original genotype scale, and we have the same power to identify the common causal SNPs as the rare causal SNPs (Wakefield, 2009). We simulated response $\boldsymbol{Y}$ under the multivariate regression model (4.1.1). To mimic the real fine-mapping regions in UK Biobank, we set the maximum PVE among traits as 0.05%. Given priors on effects, and residual correlation matrix $\boldsymbol{C}$, the simulation scheme is as follows:

1. Sample the number of causal SNPs, S (see details below for each simulated

scenario).

2. Sample the indices of the S causal SNPs uniformly from $\{1, \cdots, J\}$.

3. For each causal SNP, draw $\boldsymbol{b}$ from $\sum_{k=1}^{K} \omega_k N(\boldsymbol{0}, \boldsymbol{U}_k)$.

4. Set σ^2 such that the maximum PVE among traits achieves the specified PVE, i.e., solve for σ^2 in PVE $= \frac{\text{Var}(\boldsymbol{X}\boldsymbol{b}_r)}{\sigma^2 + \text{Var}(\boldsymbol{X}\boldsymbol{b}_r)}$, where $r = \arg\max_r \text{Var}(\boldsymbol{X}\boldsymbol{b}_r)$ and Var(.) is the sample variance.

5. Draw $\boldsymbol{Y} \sim \text{MN}(\boldsymbol{X}\boldsymbol{B}, \boldsymbol{I}, \sigma^2\boldsymbol{C})$.

Since the sample size is large, we computed marginal z scores for each trait using plink 2.0 (Chang et al., 2015; Purcell and Chang, 2019). We computed the in-sample LD matrix $\hat{\boldsymbol{R}}$ using LDstore (Benner et al., 2017) for each region.

We simulated data under three scenarios with different priors and residual correlation matrices:

Scenario 1 Artificial mixture in 20 traits. The prior is a mixture of canonical patterns of sharing (Figure 4.1), which includes trait-specific effects, effects sharing in 2 traits, effects sharing in block of traits and effects sharing in all traits with different level of correlations. The residual correlation matrix, $\boldsymbol{C}$, is the identity matrix.

Scenario 2 UK Biobank 16 Blood Cell traits patterns. The prior is a mixture of data-driven pattern from 16 UK Biobank Blood Cell traits (Figure 4.2). The residual correlation matrix, $\boldsymbol{C}$, is the empirical correlation between 16 blood cell traits.

Scenario 3 Independent effects in 2 traits. The effect is generated from the identity matrix. The residual correlation matrix, $\boldsymbol{C}$, is the identity matrix.

In Scenario 1 and 2, each region could have 1, 2, 3, 4, 5 causal SNPs with probability 0.3, 0.3, 0.2, 0.1, 0.1. In Scenario 3, each region contains 2 causal SNPs.

We compare our method using summary statistics, *mvSuSiE-RSS*, with *mvSuSiE-suff*, *SuSiE-suff* and *SuSiE-RSS* (Chapter 3) under Scenario 1 and 2. We use the simple scenario with only 2 traits (Scenario 3) to compare *mvSuSiE-RSS* with *PAINTOR* version 3.1 (Kichaev et al., 2017), since *PAINTOR* is computationally intensive with a large number of traits. Because the sample size is large, we use *mvSuSiE-suff* and *SuSiE-suff*, instead of *mvSuSiE* and *SuSiE*, to save memory and running time. *SuSiE-suff* and *SuSiE-RSS* are designed for trait-specific fine-mapping, so we applied them for each trait separately. We set the maximum number of causal SNPs to 10 ($L = 10$) in *SuSiE-suff*, *SuSiE-RSS*, *mvSuSiE-suff* and *mvSuSiE-RSS*.

PAINTOR assumes the causal SNPs are shared across traits with independent effects and independent residuals. *PAINTOR* integrates the functional annotation data into the prior inclusion probabilities. Since we ran *PAINTOR* without any annotation data, we created a "dummy" annotation file for each region with all 1's. Using the *PAINTOR* `mcmc` option, the posterior inclusion probability is always 0 in several test data sets. The same issue is reported in Github (`https://github.com/gkichaev/PAINTOR_V3.0/issues/5`). Therefore, we set *PAINTOR* to enumerate all possible configurations up to 2 causal variants, which is the oracle number of causal variants in the Scenario 3 simulated data.

We applied *mvSuSiE-suff* with oracle mixture of priors and residual covariance

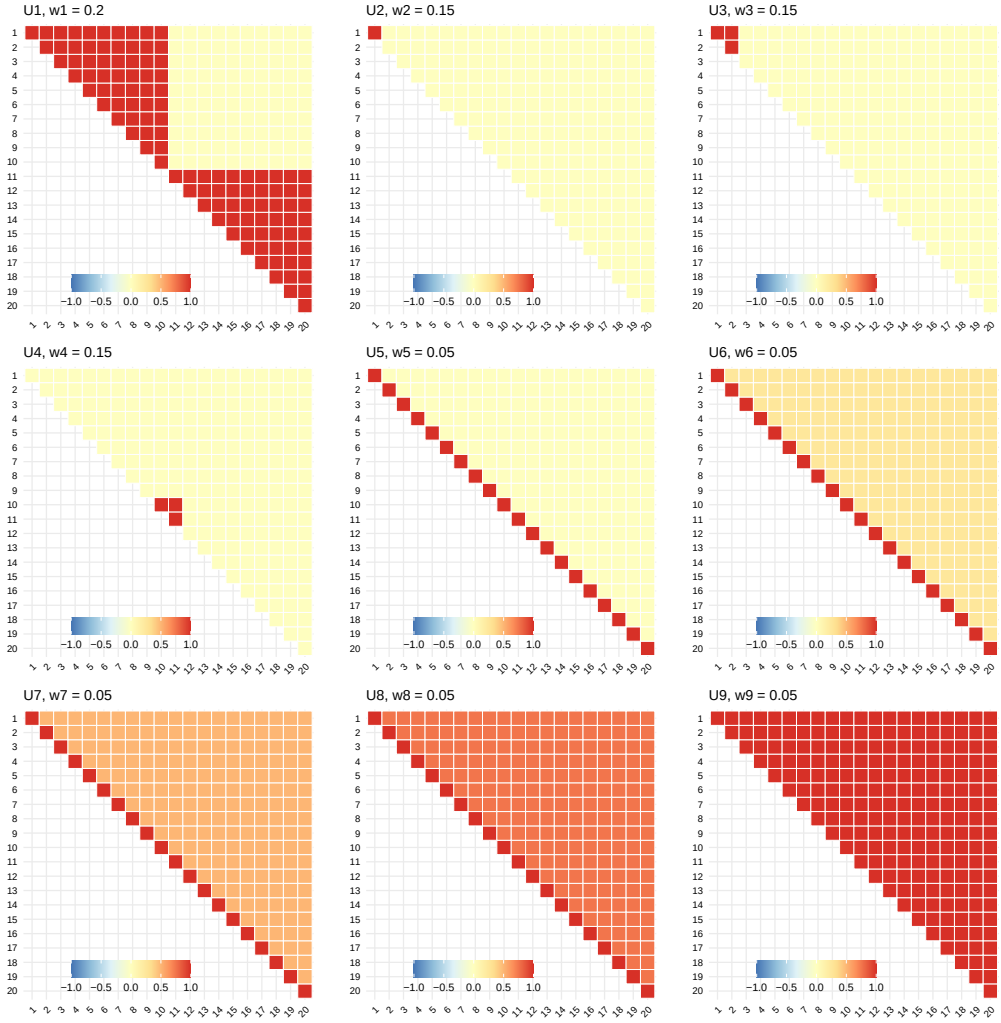


Figure 4.1: **Simulated structure for the "artificial mixture" in 20 traits for Scenario 1.** Each heatmap represents a covariance matrix $\boldsymbol{U}_k$, w_k gives the relative frequency of $\boldsymbol{U}_k$. The simulated signal has 20% chance to be shared in block ($\boldsymbol{U}_1$), 15% chance to be specific in trait 1 ($\boldsymbol{U}_2$), 30% chance to be shared in 2 traits ($\boldsymbol{U}_3$ and $\boldsymbol{U}_4$ with equal weights), 25% chance to be shared across traits with different heterogeneity ($\boldsymbol{U}_5$ - $\boldsymbol{U}_9$ with equal weights).

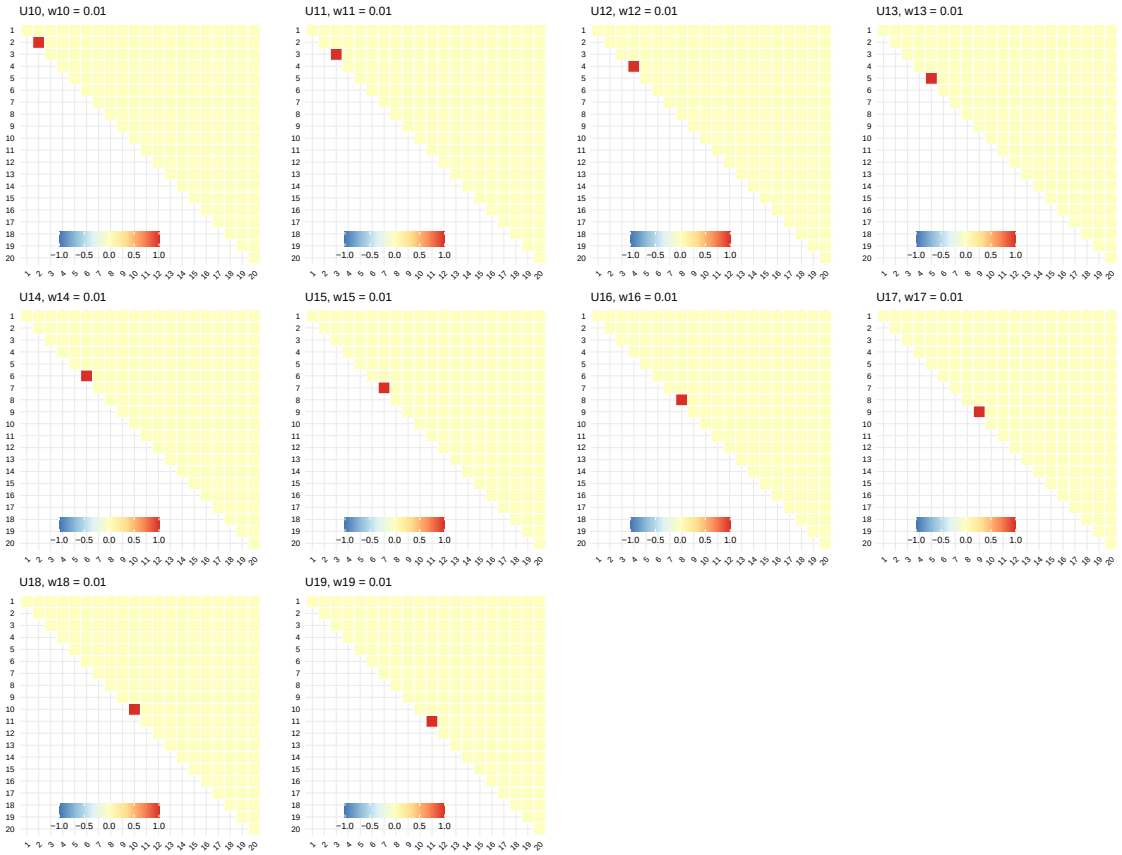


Figure 4.1: **Simulated structure for the "artificial mixture" in 20 traits for Scenario 1 (cont.).** The simulated signal has 10% chance to be specific in trait other than the first one ($\boldsymbol{U}_{10}$ - $\boldsymbol{U}_{19}$ with equal weights).

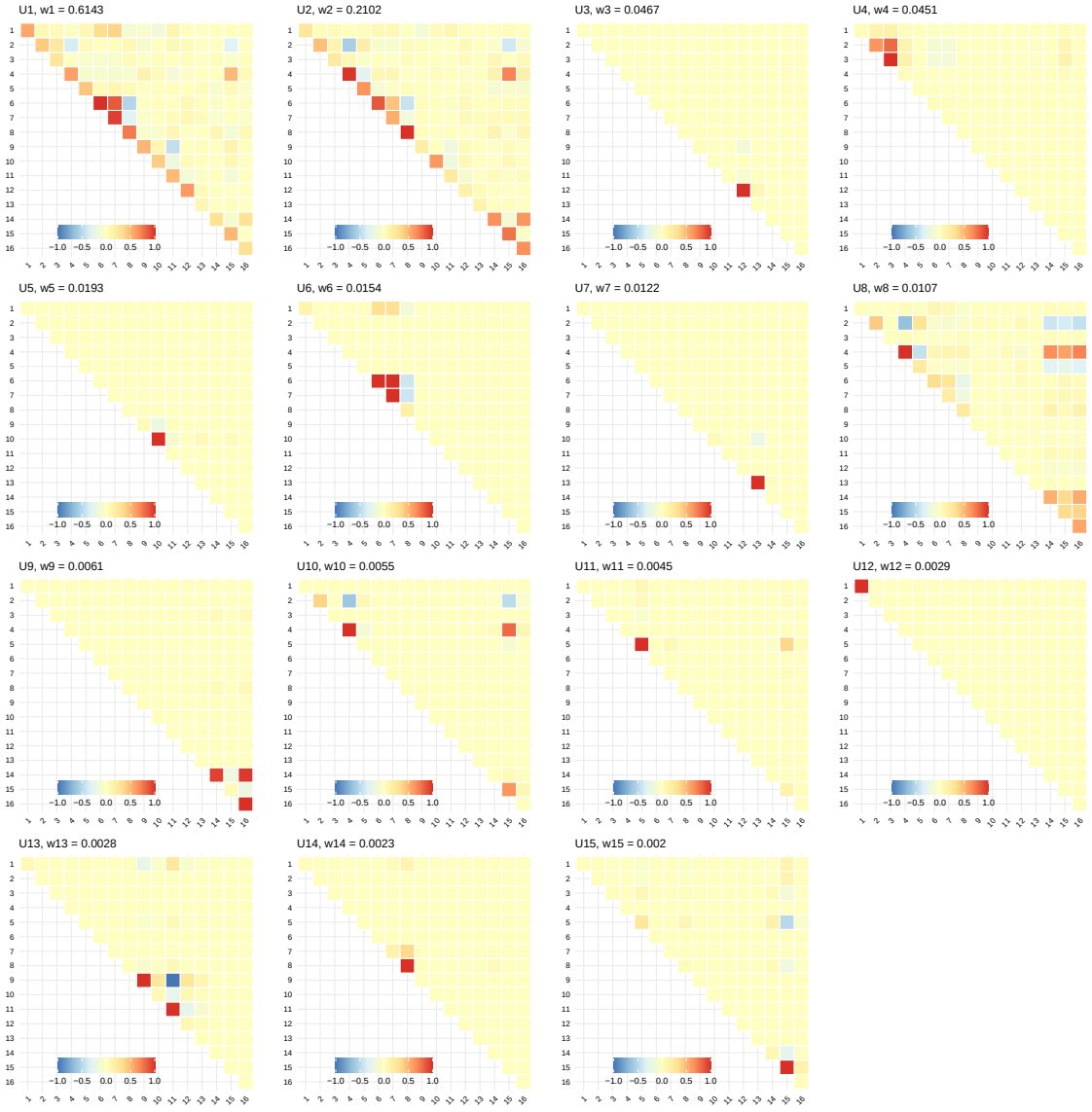


Figure 4.2: **Simulated structure for the "UK Biobank Blood Cells" in 16 traits for Scenario 2.** Each heatmap represents a covariance matrix $\boldsymbol{U}_k$, w_k gives the relative frequency of $\boldsymbol{U}_k$.

matrix. We assessed the performance of *mvSuSiE-RSS* under different priors and residual correlation matrices. We compared performance using the following priors: 1. oracle prior, the prior generating the signals; 2. default prior, a fixed mixture of canonical structure regardless of the simulation scenario; 3. random effect prior, the identity matrix, which represents independent effects in different traits; 4. fixed effect prior, a matrix of all ones, which represents identical effects among all traits; 5. ED prior, the data-driven prior using the procedure described in Section 4.3.4. The ED prior successfully identify the main signal patterns in Scenario 1 and 2 (Figure 4.3 for Scenario 1, Figure 4.4 for Scenario 2). For example, in Scenario 1, the ED prior captures the sharing in block pattern with relative frequency 0.27, the sharing between trait 10 and 11 with relative frequency 0.1. The trait 1 specific pattern and the sharing between trait 1 and 2 are collapsed into one covariance matrix, $\boldsymbol{U}_3$ in Figure 4.3. The simulated patterns about sharing across all traits are collapsed into $\boldsymbol{U}_1$ in Figure 4.3. The patterns captured in ED prior agree with the patterns in the oracle prior. We also checked performance using 4 different residual correlation matrices: 1. the oracle residual correlation matrix; 2. the identity matrix; 3. the empirical correlation matrix from the simulated traits; 4. the empirical correlation matrix from z scores (4.3.12).

The simulation was conducted using DSC. The simulation code are in `https://github.com/gaow/mvarbvs/tree/master/dsc/mnm_prototype`, the results are available at `https://github.com/zouyuxin/mmbr-rss-dsc`.

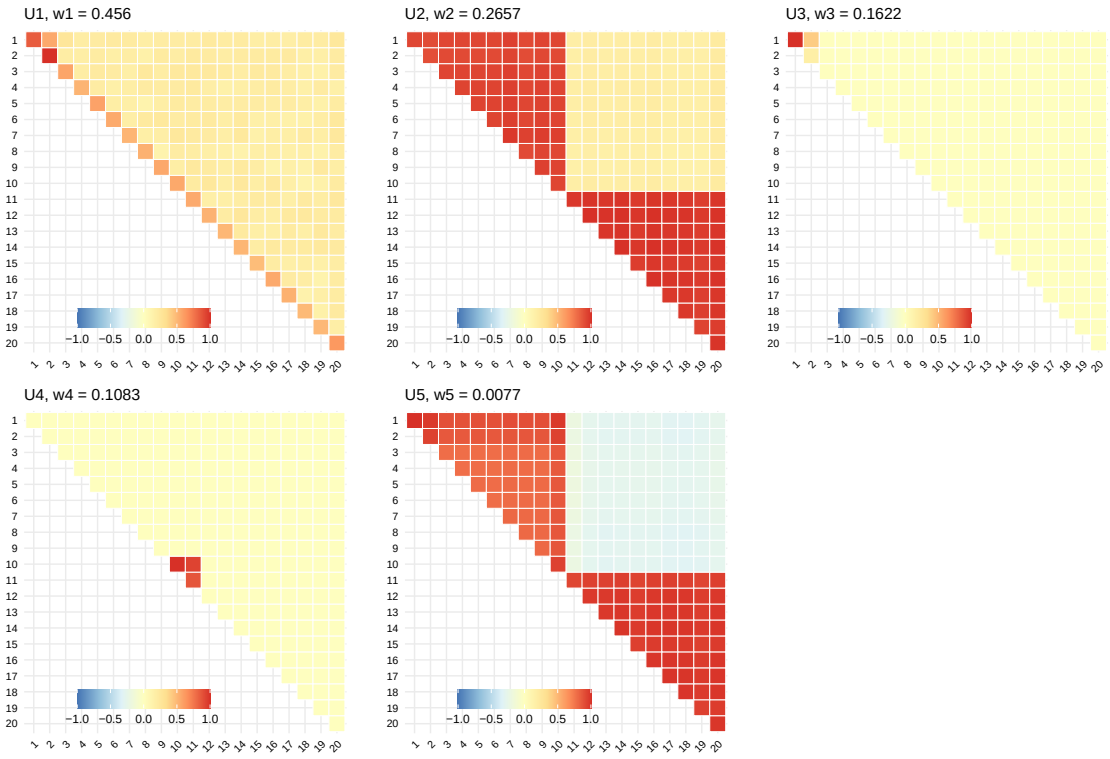


Figure 4.3: **Estimated "artificial mixture" prior for 20 traits via ED in `mashr` package.** Each heatmap represents a covariance matrix $\boldsymbol{U}_k$, w_k gives the relative frequency of $\boldsymbol{U}_k$.

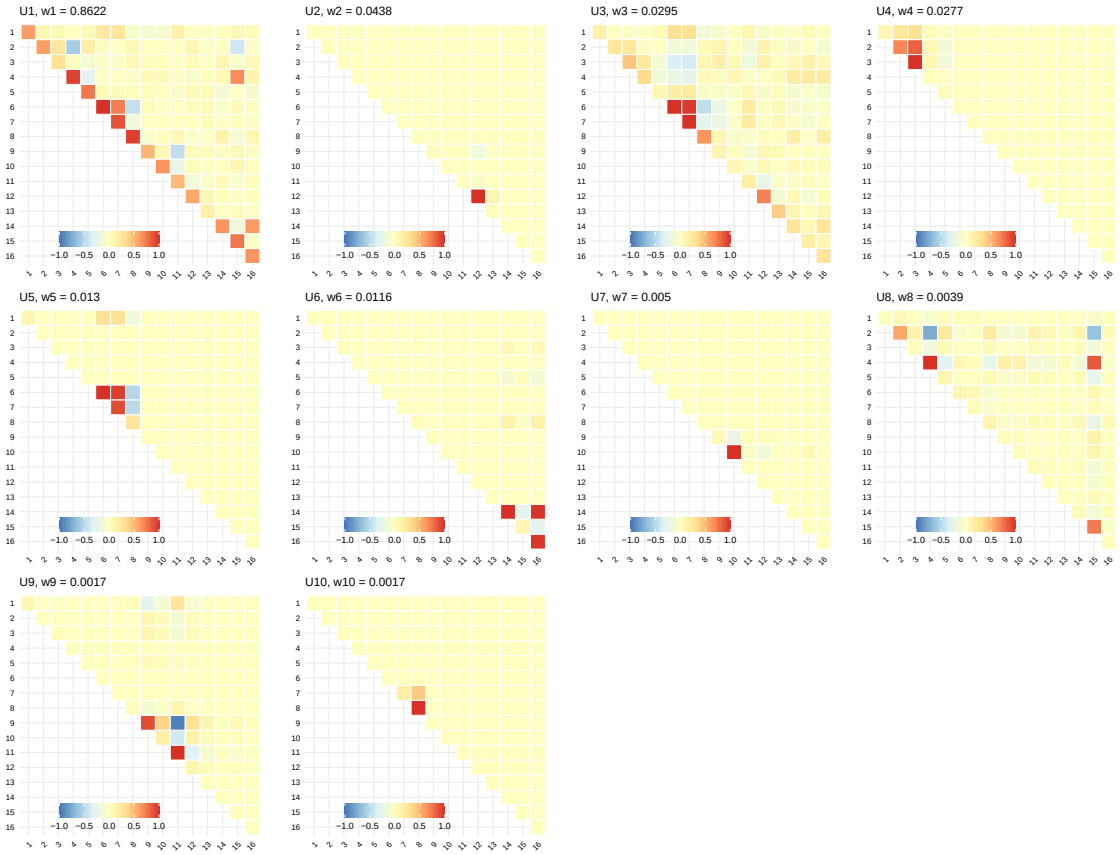


Figure 4.4: **Estimated "UK Biobank Blood Cell traits" prior via ED in `mashr` package.** Each heatmap represents a covariance matrix $\boldsymbol{U}_k$, w_k gives the relative frequency of $\boldsymbol{U}_k$.

4.5.1 Posterior Inclusion Probability

The joint fine-mapping methods, *mvSuSiE-suff*, *mvSuSiE-RSS* and *PAINTOR*, provide global PIP, which is the probability of a SNP having non-zero effect in at least one trait. In contrast, the univariate fine-mapping methods, *SuSiE-suff* and *SuSiE-RSS*, provide trait-specific PIP, which represents the probability of a SNP having non-zero effect in one specific trait. To obtain global PIP for *SuSiE-suff* and *SuSiE-RSS*, we use the maximum PIP among traits,

$$\mathrm{PIP}_j = \max_r \mathrm{PIP}_{jr} = P(\boldsymbol{z}_{jr} \neq 0 | \hat{\boldsymbol{z}}_r, \hat{\boldsymbol{R}}). \tag{4.5.1}$$

We use the maximum rather than deriving the PIP_j under the assumption that the traits are independent, because the independence assumption leads to over-estimated PIP. In this subsection, we assess the power, false discovery rate and calibration of PIP.

Power and False Discovery Rate

We first examine the power and false discovery rate (FDR) in discovering non-zero effects using PIPs.

Using Scenario 3, we compare *mvSuSiE-RSS* with *PAINTOR*. The *mvSuSiE-RSS* method was applied with default prior and residual correlation matrix estimated from z scores. *mvSuSiE-RSS* has higher power and lower FDR than *PAINTOR* (Figure 4.5). The mean running time for *PAINTOR* is 1785.22 seconds, which is much slower than *mvSuSiE-RSS* (average running time is 32.76 seconds). Therefore, we exclude

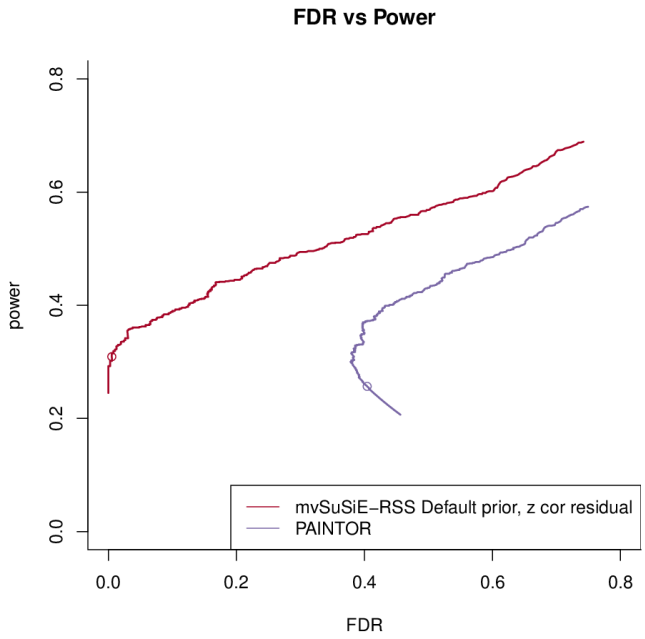


Figure 4.5: **Power versus FDR in Scenario 3.** *mvSuSiE-RSS* was fitted using default prior and residual correlation matrix estimated from z scores.

PAINTOR in the following assessments.

In Scenario 1 and 2, our joint fine-mapping methods have higher power than univariate fine-mapping methods (Figure 4.6). This suggests that leveraging association strength across related traits increases the power to detect weakly associated causal variants in single traits.

With oracle prior and residual covariance/correlation matrix, the model using summary statistics, *mvSuSiE-RSS*, performs similarly to *mvSuSiE-suff*, which uses complete information from genotype and phenotype data. This is expected because both *mvSuSiE-suff* and *mvSuSiE-RSS* use the same in-sample LD matrix, and the sample size is large in the simulation, so the information lost in the summary statistics

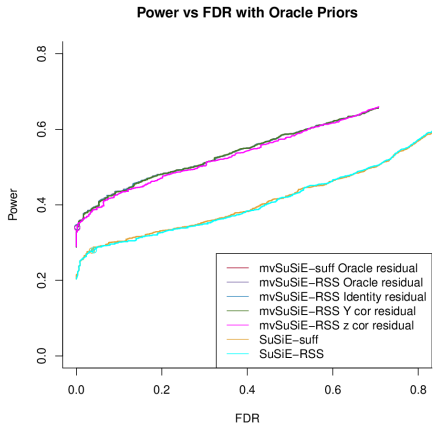
is negligible.

Using the identity matrix as residual correlation matrix increases FDR dramatically in Scenario 2, in which the simulated residual correlation is a dense matrix. The identity matrix ignores the residual correlation between traits. Since the residual correlations between traits are ignored in the residual part of the model, the model attempts to include these residual correlations in the signal part, which induces false positives. Using the residual correlation matrix estimated from simulated phenotypes or z scores close to null, the results are similar. The empirical correlation matrix of phenotypes performs slightly better than the one from z scores, because it contains more information than z scores.

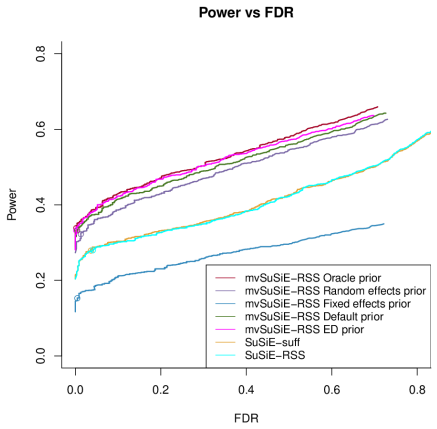
The performance of *mvSuSiE-RSS* using the ED prior is similar to the oracle prior. The *mvSuSiE-RSS* model with fixed effect prior has the lowest power in both scenarios. This is because the signals are simulated from a mixture of covariance structures. The fixed effect prior can only capture the signals that share identically among all traits, which fails to capture other signal patterns. Indeed, in Scenario 2, there is no fixed effect pattern in the prior structure.

PIP calibration

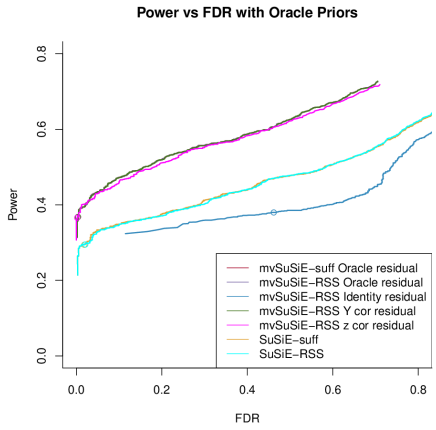
Next, we assess PIP calibration for *mvSuSiE-RSS* with different priors and residual correlation matrices. We group SNPs across all simulated data sets into 10 bins according to their reported PIP. Then we compute the proportion of SNPs with non-zero effects in at least one phenotype in each bin (i.e. observed frequency). For a well-calibrated method, the observed frequency is approximately equal to the average



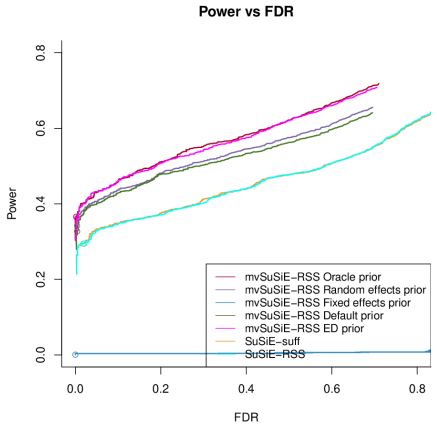
(a) **Artificial Mixture 20 traits. Results for different residual correlation matrices.** The prior mixture is fixed as oracle.



(b) **Artificial Mixture 20 traits. Results for different priors.** The residual covariance matrix is fixed as the empirical correlation from z scores.



(c) **UK Biobank 16 Blood Cell traits. Results for different residual correlation matrices.** The prior mixture is fixed as oracle.



(d) **UK Biobank 16 Blood Cell traits. Results for different priors.** The residual covariance matrix is fixed as the empirical correlation from z scores.

Figure 4.6: **Comparison of Power and FDR for different method in Scenario 1 and 2.**

PIP for each bin.

With appropriate priors and residual correlation matrix, the PIPs from *mvSuSiE-suff* and *mvSuSiE-RSS* are well-calibrated in Scenario 1 and 2 (Figure 4.7, 4.8). The PIPs from *mvSuSiE-RSS* using the identity residual correlation matrix in Scenario 2 are anti-conservative as expected.

Using fixed effect prior in Scenario 2, the estimated observed frequencies have large errors (the left plot on the third line in Figure 4.8). This is because the PIPs from *mvSuSiE-RSS* with fixed effect prior are too conservative in Scenario 2, there are not enough SNPs with PIP greater than 0.1 to estimate the observed frequency in each bin. Therefore, the estimated observed frequencies are not accurate.

4.5.2 Credible Sets

From *mvSuSiE-suff* and *mvSuSiE-RSS*, we obtain multiple *CSs*, each aimed at capturing one effect SNP with non-zero effect in at least one trait. *PAINTOR* does not produce *CS*, so it is excluded in the following comparisons. We assess the 95% *CSs* produced by *mvSuSiE-RSS* and *mvSuSiE-suff* using empirical coverage, power, median size of *CS* and median purity. The empirical coverage is the proportion of *CSs* that contain an effect variant with non-zero effect in at least one trait. The empirical power is the proportion of effect variants (in at least one trait) included in a *CS*.

With appropriate priors and residual correlation matrix, the *CSs* from *mvSuSiE-suff* and *mvSuSiE-RSS* have high coverage and high power (Figure 4.9). The *CSs* from *mvSuSiE-RSS* with misspecified residual correlation (identity matrix in Sce-

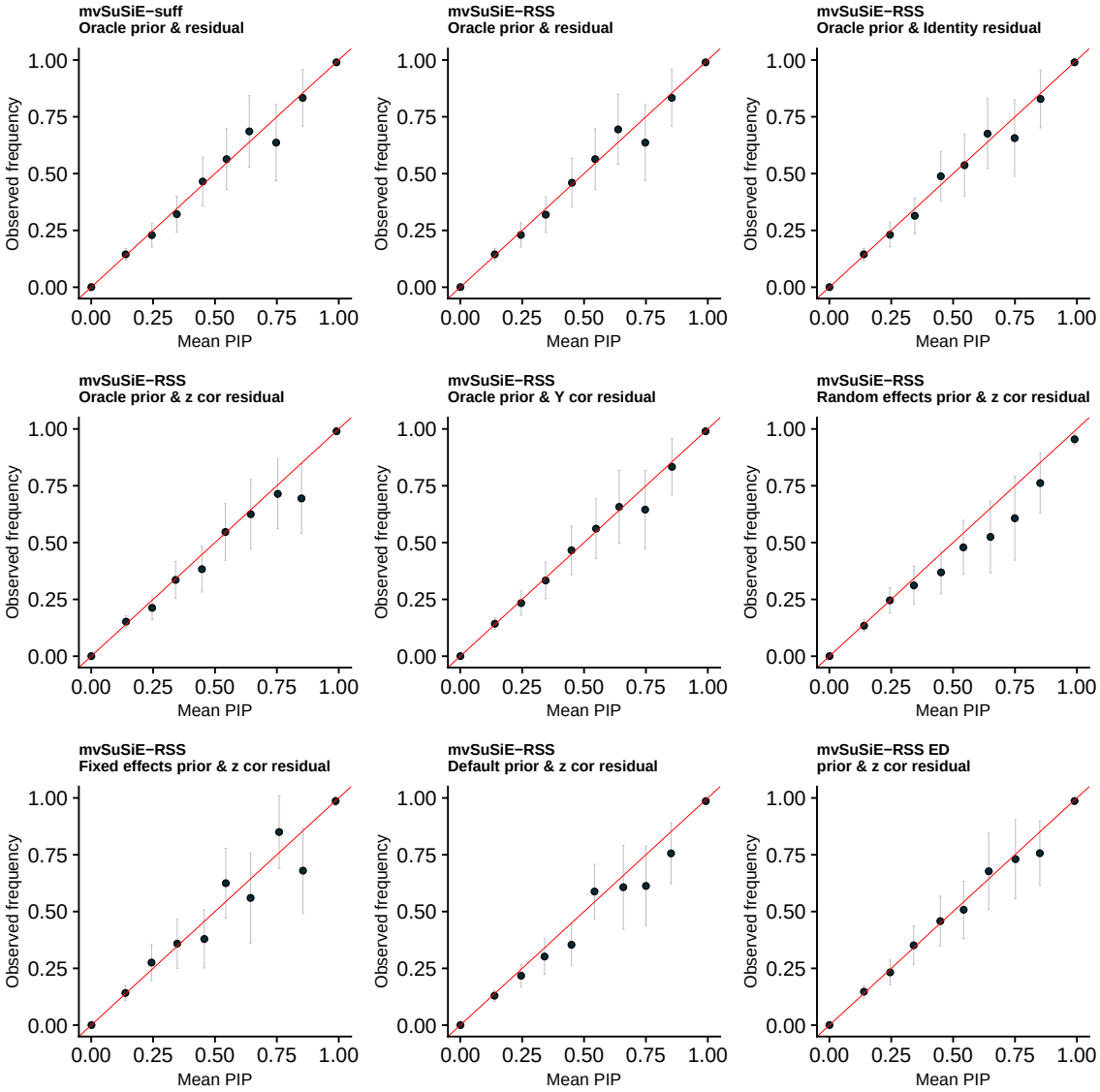


Figure 4.7: **Calibration of PIP for *mvSuSiE-suff* and *mvSuSiE-RSS* in Scenario 1.** The plots show the proportion of effect SNPs versus the mean PIP for each bin. We expect all points are aligned in the diagonal line for a well-calibrated method. The gray error bars show ± 2 standard errors. Points below the diagonal line imply the corresponding PIPs are anti-conservative and points above the diagonal line imply the PIPs are conservative.

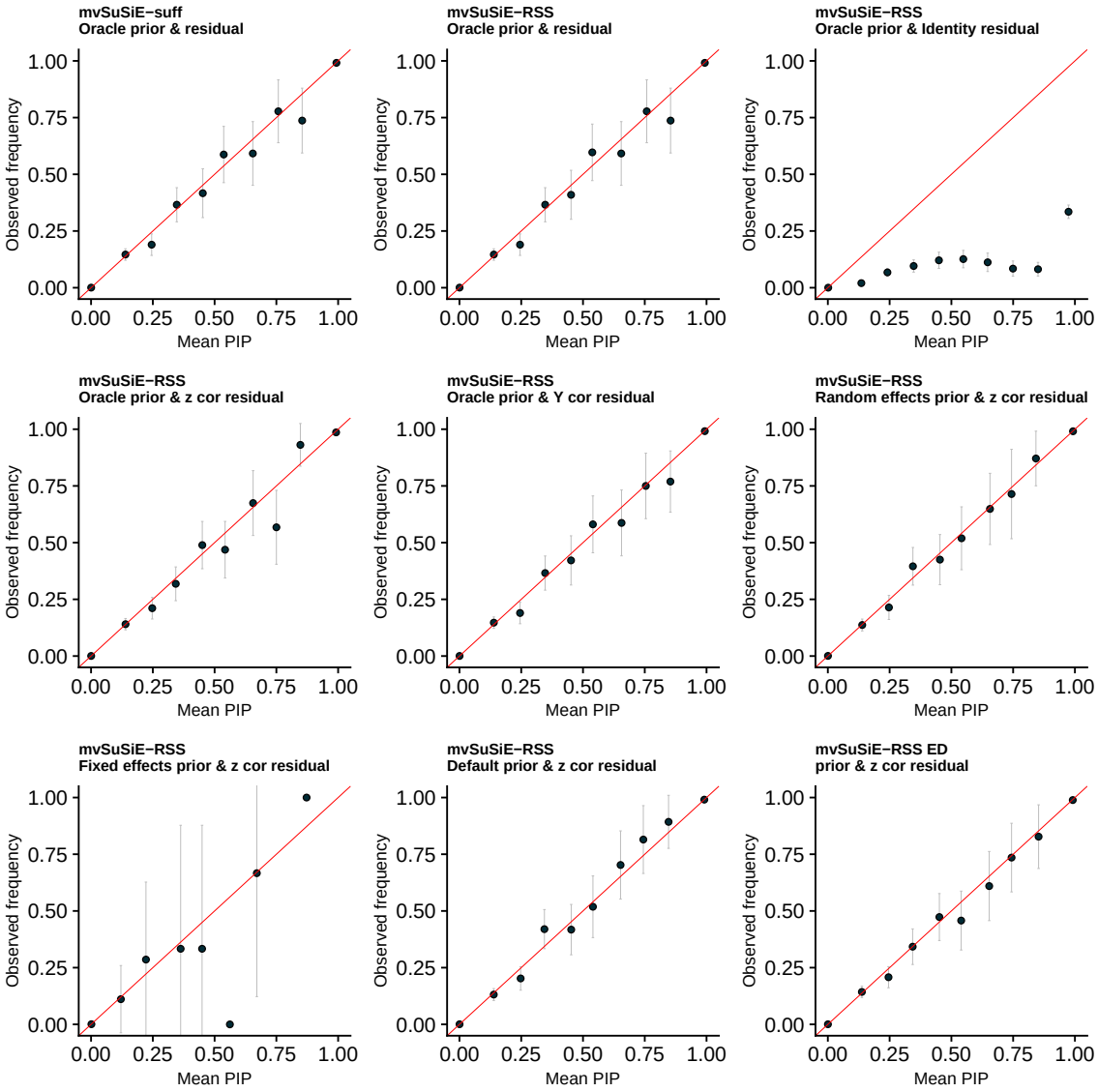


Figure 4.8: **Calibration of PIP for *mvSuSiE-suff* and *mvSuSiE-RSS* in Scenario 2.**

nario 2) or mismatched prior (fixed effect prior) have lower coverage and power.

To compare with trait-specific *CSs* from *SuSiE-suff* and *SuSiE-RSS*, we assess the significance of the *CSs* produced by *mvSuSiE-suff* and *mvSuSiE-RSS* using *lfsr* (4.4.3) for each trait. For each trait, we obtain the significant *CSs* ($lfsr < 0.05$). The trait-specific *CSs* from *mvSuSiE-suff* and *mvSuSiE-RSS* outperform *SuSiE-suff* and *SuSiE-RSS* in power, size and purity (Figure 4.10). The joint fine-mapping model produces a smaller, purer *CS* than univariate fine-mapping. The simulation demonstrates that fine-mapping resolution is improved by combining information across traits.

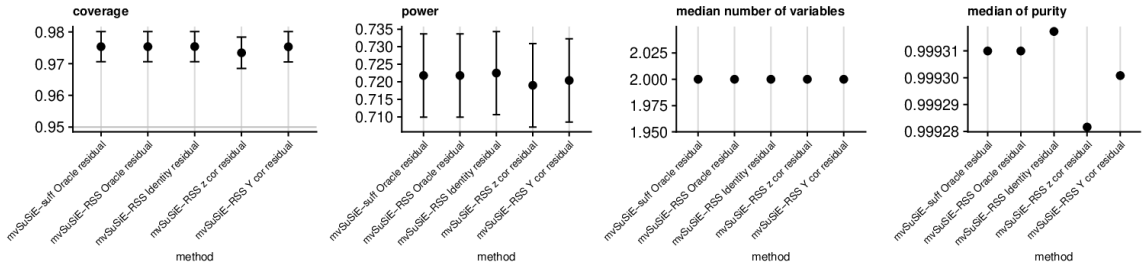
4.5.3 Runtimes

The running time for different methods in Scenario 1 are similar to one another, and it is summarized in Table 4.1.

Table 4.1: **Runtimes in seconds with 20 traits.** *mvSuSiE-suff* and *mvSuSiE-RSS* were fitted using default priors. The residual covariance matrix in *mvSuSiE-suff* is the empirical covariance matrix of phenotypes. The residual covariance matrix in *mvSuSiE-RSS* is the empirical correlation matrix of from z scores (4.3.12).

method	mean	min.	max.
mvSuSiE-suff	75.21	20.66	333.86
mvSuSiE-RSS	75.58	17.95	279.09
SuSiE-suff	105.76	15.44	305.50
SuSiE-RSS	113.55	14.46	311.09

(a) **Artificial Mixture 20 traits. Results for different residual correlation matrices.** The prior mixture is fixed as oracle.



(b) **Artificial Mixture 20 traits. Results for different priors.** The residual covariance matrix is fixed as the empirical correlation from z scores.

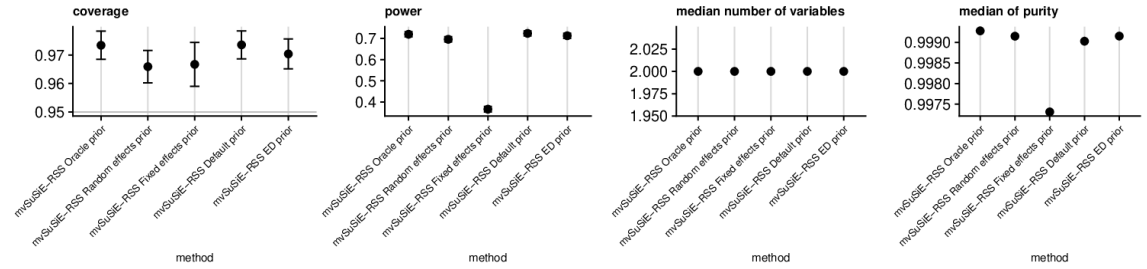
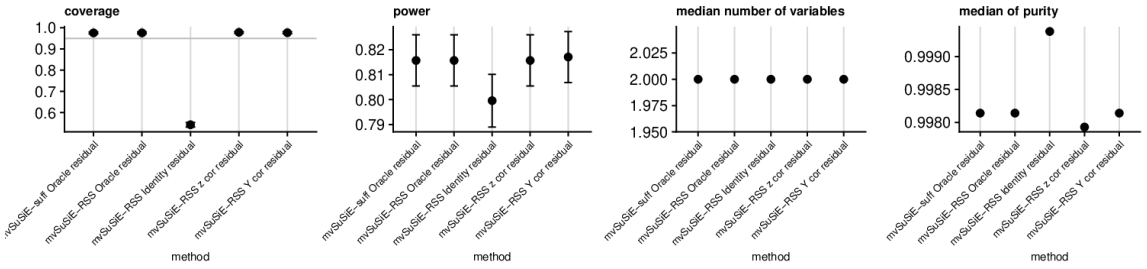


Figure 4.9: **Compare 95% Credible Sets (*CSs*) from *mvSuSiE-suff* and *mvSuSiE-RSS*.** The coverage, power, size and purity are computed using all *CSs* in all data sets. The error bar is computed as $2 \times$ standard error. The panel (a) and (b) are from simulation Scenario 1.

(c) **UK Biobank 16 Blood Cell traits. Results for different residual correlation matrices.** The prior mixture is fixed as oracle.



(d) **UK Biobank 16 Blood Cell traits. Results for different priors.** The residual covariance matrix is fixed as the empirical correlation from z scores.

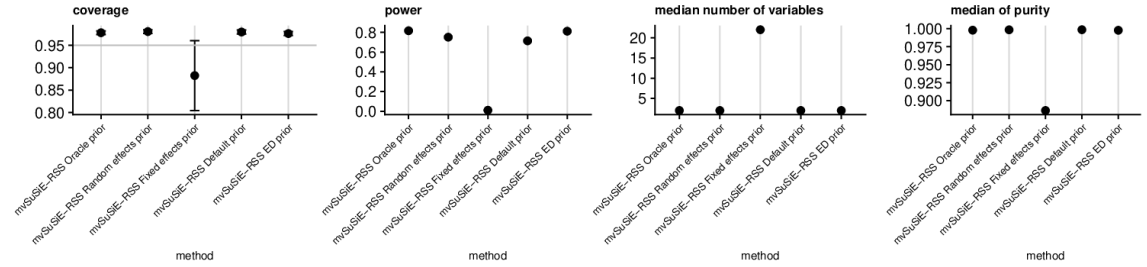
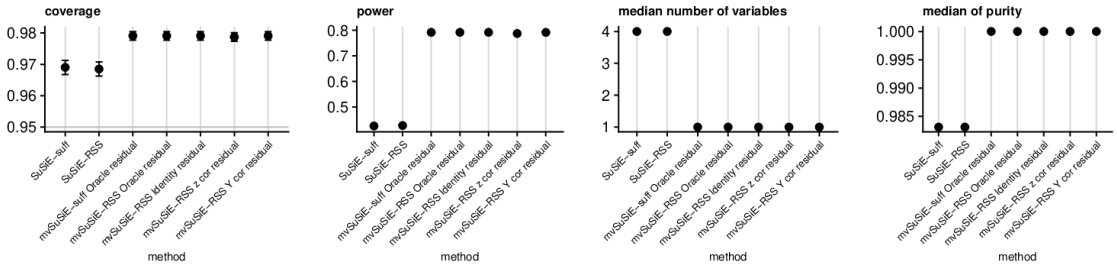


Figure 4.9: **Compare 95% Credible Sets (*CSs*) from *mvSuSiE-suff* and *mvSuSiE-RSS* (cont.).** The panel (c) and (d) are from simulation Scenario 2.

(a) **Artificial Mixture 20 traits. Results for different residual correlation matrices.** The prior mixture is fixed as oracle.



(b) **Artificial Mixture 20 traits. Results for different priors.** The residual covariance matrix is fixed as the empirical correlation from z scores.

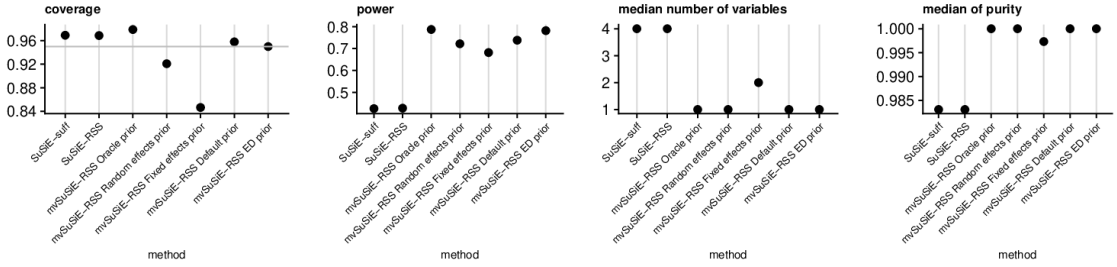
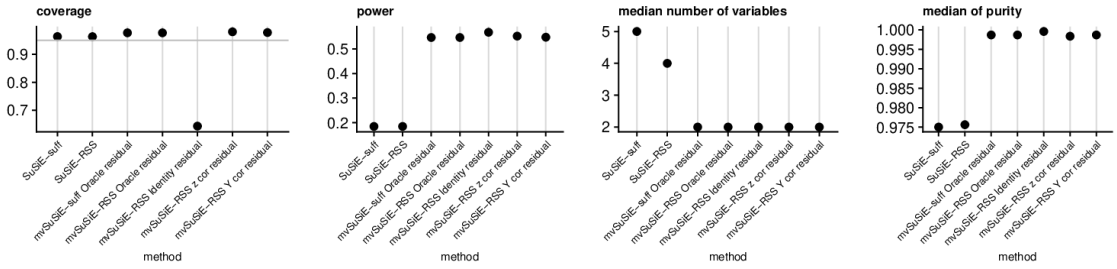


Figure 4.10: **Compare 95% trait-specific Credible Sets (*CSs*) from *SuSiE-suff*, *SuSiE-RSS*, *mvSuSiE-suff* and *mvSuSiE-RSS*.** The coverage, power, size and purity are computed using all *CSs* in all traits in all data sets. The error bar is computed as 2 × standard error. The panel (a) and (b) are from simulation Scenario 1.

(c) **UK Biobank 16 Blood Cell traits. Results for different residual correlation matrices.** The prior mixture is fixed as oracle.



(d) **UK Biobank 16 Blood Cell traits. Results for different priors.** The residual covariance matrix is fixed as the empirical correlation from z scores.

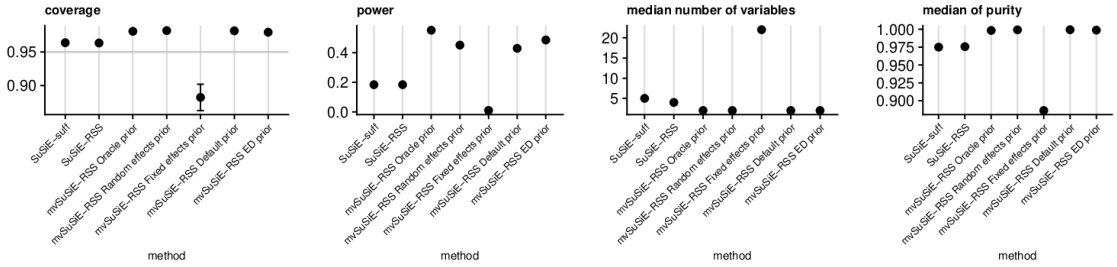


Figure 4.10: **Compare 95% trait-specific Credible Sets (*CSs*) from *SuSiE-suff*, *SuSiE-RSS*, *mvSuSiE-suff* and *mvSuSiE-RSS* (cont.).** The panel (c) and (d) are from simulation Scenario 2.

4.6 Fine-mapping on UK Biobank Blood Cell traits

To evaluate *mvSuSiE-RSS* on a real fine-mapping problem, we jointly analyzed 16 blood cell traits in UK Biobank. In this section, we describe the data preparation process, summarize the *mvSuSiE-RSS* results and illustrate *mvSuSiE-RSS* on some specific example regions.

4.6.1 UK Biobank Data

UK Biobank is a prospective cohort study with data on approximately 500,000 individuals from the United Kingdom, aged between 40 and 69 at recruitment (Sudlow et al., 2015; Bycroft et al., 2018). We choose 16 blood cell traits because they tend to be all measured for the same samples and there are some fine-mapping results for some of the blood cell traits that we can compare with (Astle et al., 2016; Ulirsch et al., 2019; Vuckovic et al., 2020). The 16 traits are summarized in Table 4.2. The cell type for each trait is based on Vuckovic et al. (2020).

There are 248,980 White British unrelated individuals after removing individuals with missing traits, mismatches between self-reported and genetic sex, pregnancy and any of the following diseases in hospital in-patient data: leukemia, lymphoma, bone marrow transplant, chemotherapy, myelodysplastic syndrome, anemia, HIV, end-stage kidney disease, dialysis, cirrhosis, multiple myeloma, lymphocytic leukemia, myeloid leukemia, polycythaemia vera, haemochromatosis. We also excluded outliers defined by UK Biobank. The traits were inverse transformed to a standard normal distribution. Because we would jointly model the 16 blood cell traits, we

Table 4.2: **UK Biobank blood cell traits.**

Abbreviation	Long Name	Cell Type
WBC#	White blood cell count	Compound white cell
RBC#	Red blood cell count	Mature red cell
HGB	Haemoglobin concentration	Mature red cell
MCV	Mean corpuscular volume	Mature red cell
RDW	Red blood cell distribution width	Mature red cell
PLT#	Platelet count	Platelet
PCT	Plateletcrit	Platelet
PDW	Platelet distribution width	Platelet
LYMPH%	Lymphocyte %	Compound white cell
MONO%	Monocyte %	Compound white cell
NEUT%	Neutrophill %	Compound white cell
EO%	Eosinophill %	Compound white cell
BASO%	Basophill %	Compound white cell
RET%	Reticulocyte %	Immature red cell
MSCV	Mean sphered cell volume	Mature red cell
HLR%	High light scatter reticulocyte %	Immature red cell

discarded outliers in the multivariate normal distribution. We measured the Mahalanobis distance between the observation and the multivariate normal $N(\mathbf{0}, \mathbf{\Sigma})$, where $\mathbf{\Sigma}$ is the empirical covariance matrix of the blood cell traits. We discarded samples with Mahalanobis distance falling into the upper 0.01 quantile of χ^2_{16} distribution. We included SNPs with INFO > 0.6 and MAF > 0.001 in association studies (Weissbrod et al., 2020). We performed a full GWAS using plink2.0 (Chang et al., 2015; Purcell and Chang, 2019) controlling for the top 10 PCs, sex, age, age^2, UK Biobank assessment centre and genotype measurement batch.

To select regions for fine-mapping in multiple traits, we first selected regions for each trait. For each trait, we derived regions as $\pm$ 250 kb centered at the top SNP, until we included all significant SNPs ($p < 5 \times 10^{-8}$). We merged the overlapping

regions together. The HLA region was excluded (chr6: 25Mb - 36Mb) because of the complexity of this region (e.g. Consortium et al., 1999; Horton et al., 2004). To define regions for multivariate fine-mapping, we used all regions from each phenotype and merged overlapping regions, which produced 975 regions in total. The regions varied from 400 Kb to 8.7 Mb, each contains 93 to 36,605 SNPs. For each region, we computed the in-sample LD matrix $\hat{\boldsymbol{R}}$ using LDstore (Benner et al., 2017).

4.6.2 Multivariate fine-mapping

We applied *mvSuSiE-RSS* to fine-map each region. To specify the residual correlation matrix, $\boldsymbol{C}$, we randomly selected 2 variants in each region, the selected variants have (absolute) z scores < 2 in all traits. We estimated the residual correlation matrix $\boldsymbol{C}$ by (4.3.12). The prior covariance matrices and the corresponding mixture weights were estimated from the data using the ED procedure described in Section 4.3.4. The ED algorithm was initialized using both data-driven and canonical covariance matrices.

Figure 4.11 summarizes the identified patterns from ED procedure. The identified primary pattern shows effects are correlated among red blood cells (Figure 4.12), MCV is positively correlated with MSCV, but negatively correlated with RBC#. The second prior component captures correlations among effects for compound white cells and platelet cells. The most common types of white blood cells are neutrophill and lymphocyte. The NEUT% and LYMPH% are negatively correlated, because the traits are about the percentages of white blood cells, a genetic variant that increases one trait naturally decreases the other. The third covariance matrix shows the effects

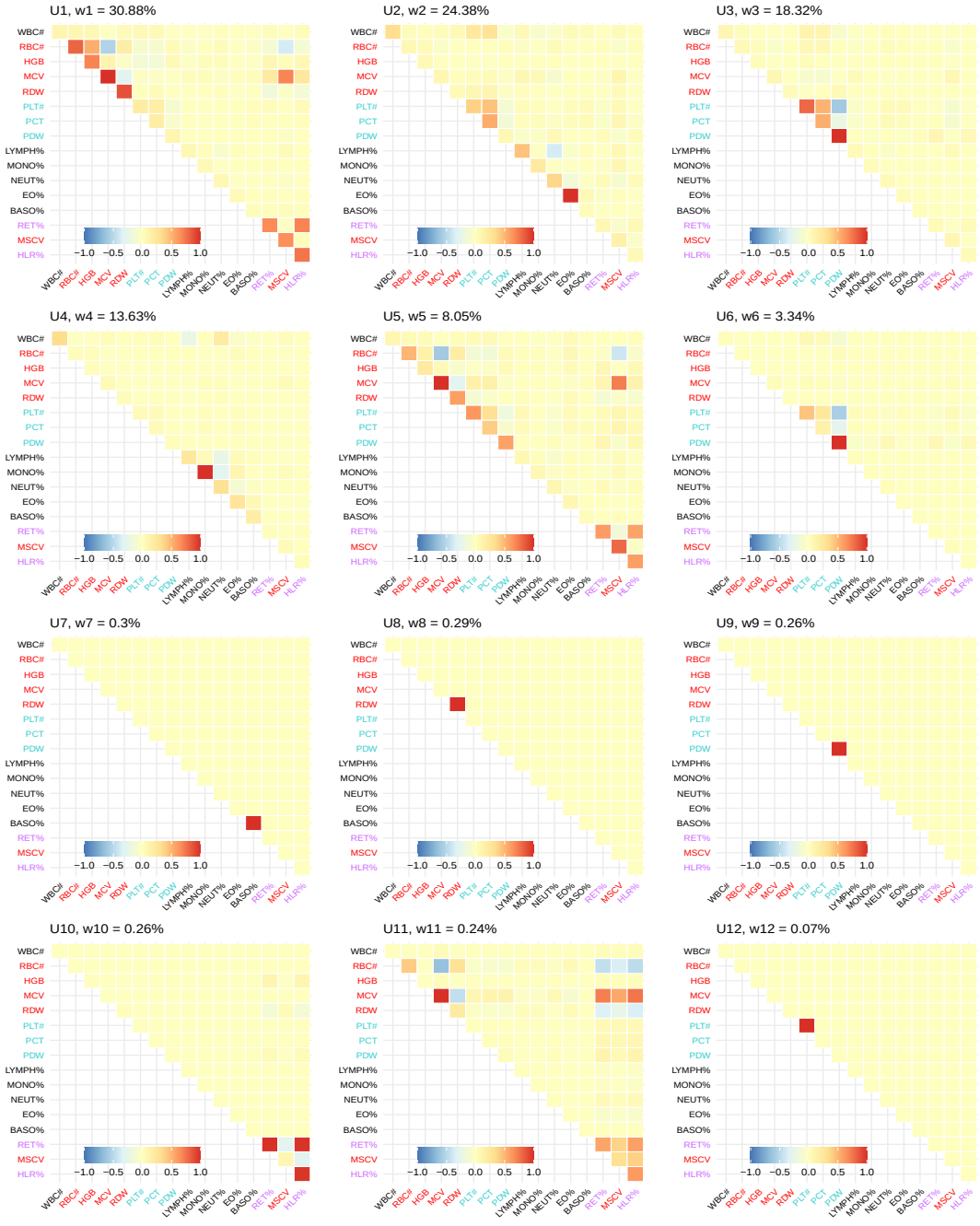
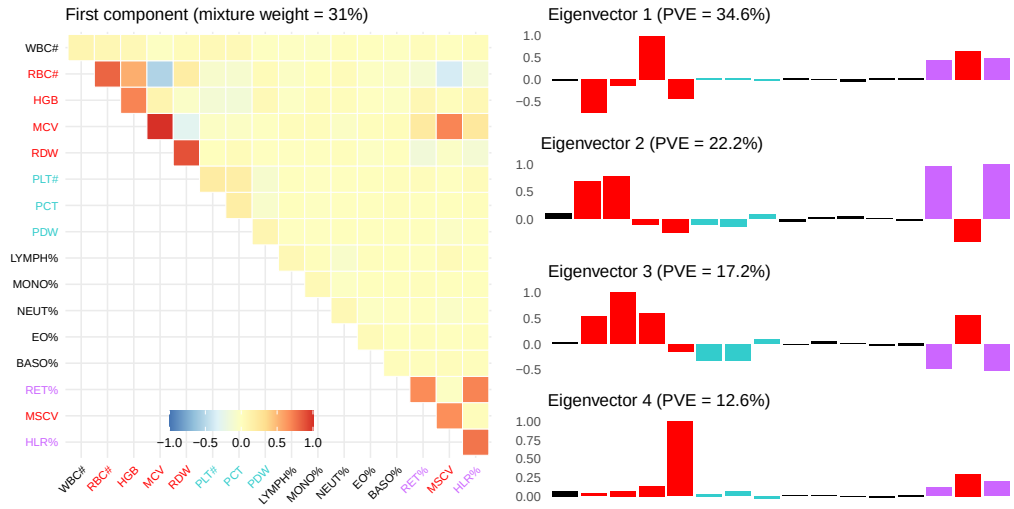


Figure 4.11: **Estimated UK Biobank 16 Blood Cell traits prior via ED in `mashr` package.** Each heatmap represents a covariance matrix $\boldsymbol{U}_k$, w_k gives the relative frequency of $\boldsymbol{U}_k$. Traits are color-coded by cell types.

(a) Prior Component 1



(b) Prior Component 2

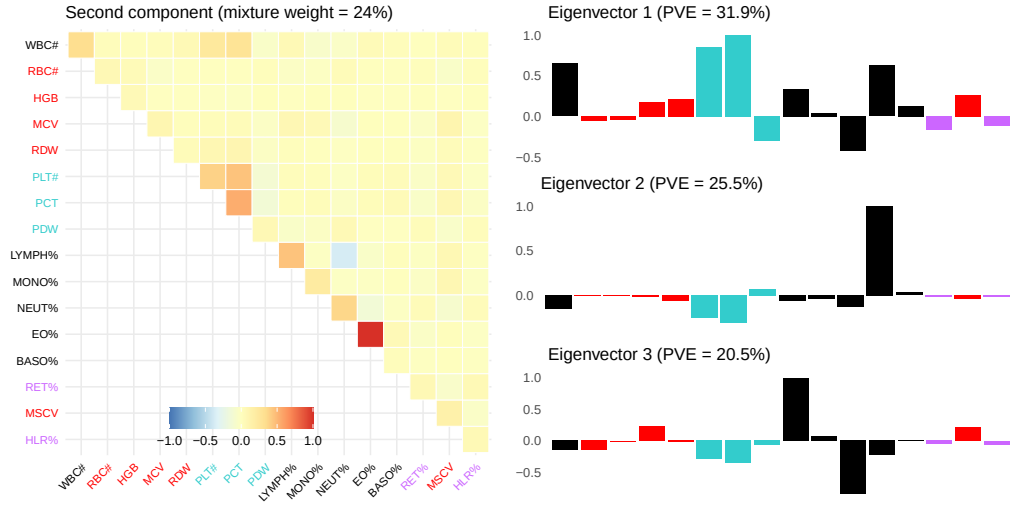
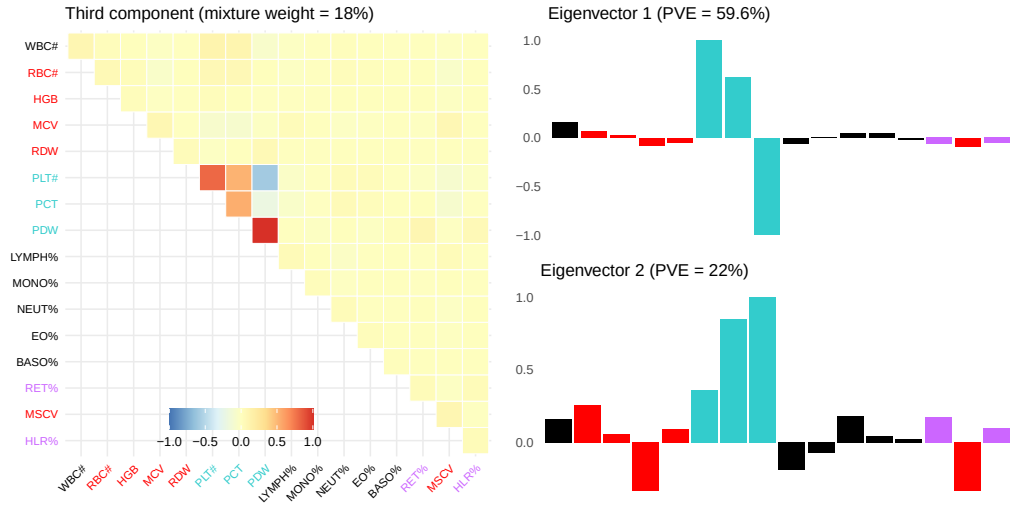


Figure 4.12: **Summary of patterns identified by ED in `mashr` package.** For each covariance matrix $\boldsymbol{U}_k$, the figure shows the heatmap of $\boldsymbol{U}_k$, and bar plots of the top eigenvectors of $\boldsymbol{U}_k$. Traits are color-coded by cell types. Component (a) reflects effects sharing among red cells. Component (b) captures correlations among compound white cells and platelet.

(c) Prior Component 3



(d) Prior Component 4

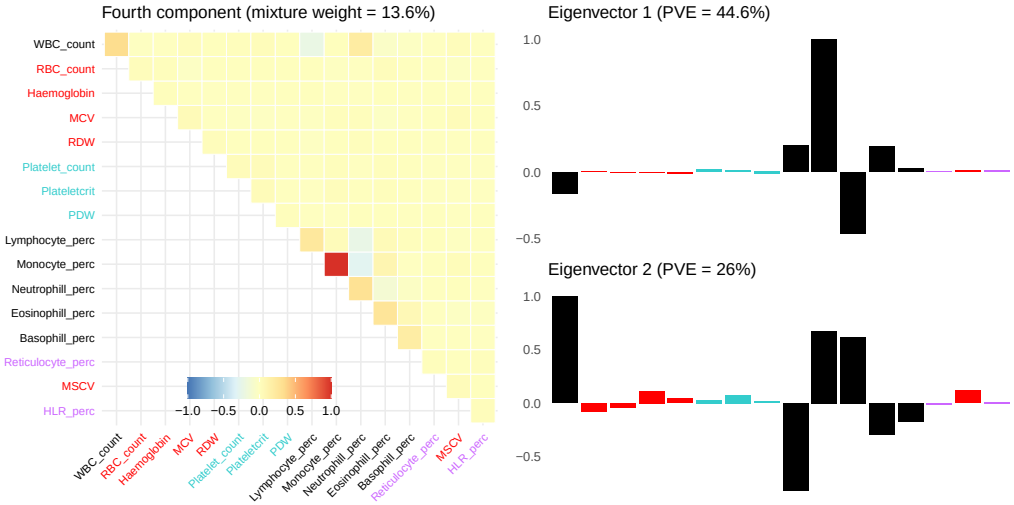


Figure 4.12: **Summary of patterns identified by ED in `mashr` package (cont.).** Component (c) captures platelet effects. Component (b) captures compound white cells effects.

are correlated among platelet; the platelet count is negatively correlated with platelet distribution width. Other prior components capture less prevalent patterns, such as trait-specific effects.

From *mvSuSiE-RSS*, there are 954 regions, out of 975, contain at least one 95% *CS*. The 21 regions that did not contain a CS are mostly corresponding to borderline significant signals, which led to impure *CS* and we filtered those *CS* out. There are 3,870 95% *CSs* in total, 767 contain exactly one variant. The median size of a *CS* is 7, and the median purity is 0.97.

To investigate sharing of signals among blood cell traits, we further assess the significance of each *CS* in different blood cell traits using *lfsr* (4.4.3) with threshold 0.01. Thus, we have a list of significant *CSs* for each trait. Figure 4.13 shows that the majority of the *CSs* are shared among a subset of blood cell traits. Figure 4.14 summarizes the proportion of significant *CSs* are being shared for each pair of blood cell traits. There are particularly high sharing in several cell types, e.g. platelet cells, mature red cells, immature red cells.

4.6.3 Comparison with single-trait fine-mapping

To compare the multivariate fine-mapping result with single-trait fine-mapping result, we applied *SuSiE-RSS* for each trait separately. The *CSs* information for each trait are summarized in Table 4.3. Consistent with simulations, multivariate fine-mapping provides a reduction in credible set size and a slight increment in credible set purity, compared to single-trait fine-mapping. Furthermore, multivariate fine-mapping finds more signals. The number of identified *CSs* from *mvSuSiE-RSS* is

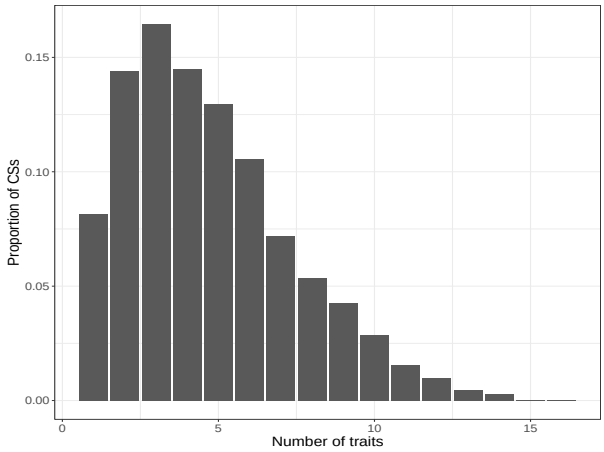


Figure 4.13: **Number of traits share a *CS* from *mvSuSiE-RSS*.** The histogram shows number of traits in which the *CS* are significant in.

much more than any of the single-trait fine-mapping results. For any trait, the multivariate fine-mapping also has more significant *CSs* than single-trait analysis (Figure 4.15). In single-trait fine-mapping, MSCV has fewer *CSs* than MCV, whereas multivariate fine-mapping produces similar number of significant *CSs* for MSCV and MCV. This is because we have learned MCV shares signals with MSCV from the data and we include the information in the prior (Figure 4.12(a)); the posterior on effects are therefore tightly correlated. Once we are certain about the signals in MCV, we become certain of the signals in MSCV as well, and the power to detect signals in MSCV increases.

4.6.4 Examples

We illustrate the gains from *mvSuSiE-RSS* on two examples, the regions around AK3 and GLIS3. In both regions, *mvSuSiE-RSS* successfully identified the validated

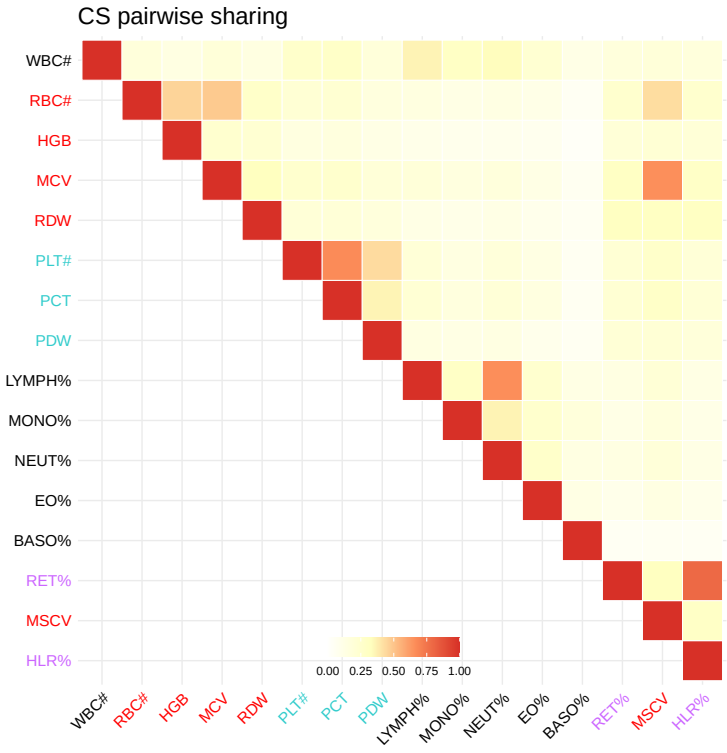


Figure 4.14: **Pairwise sharing of significant *CS* among blood cell traits.** For each pair of blood traits, we consider the *CSs* that are significant ($lfsr < 0.01$) in at least one of the two blood traits, and plot the proportion of these that are shared.

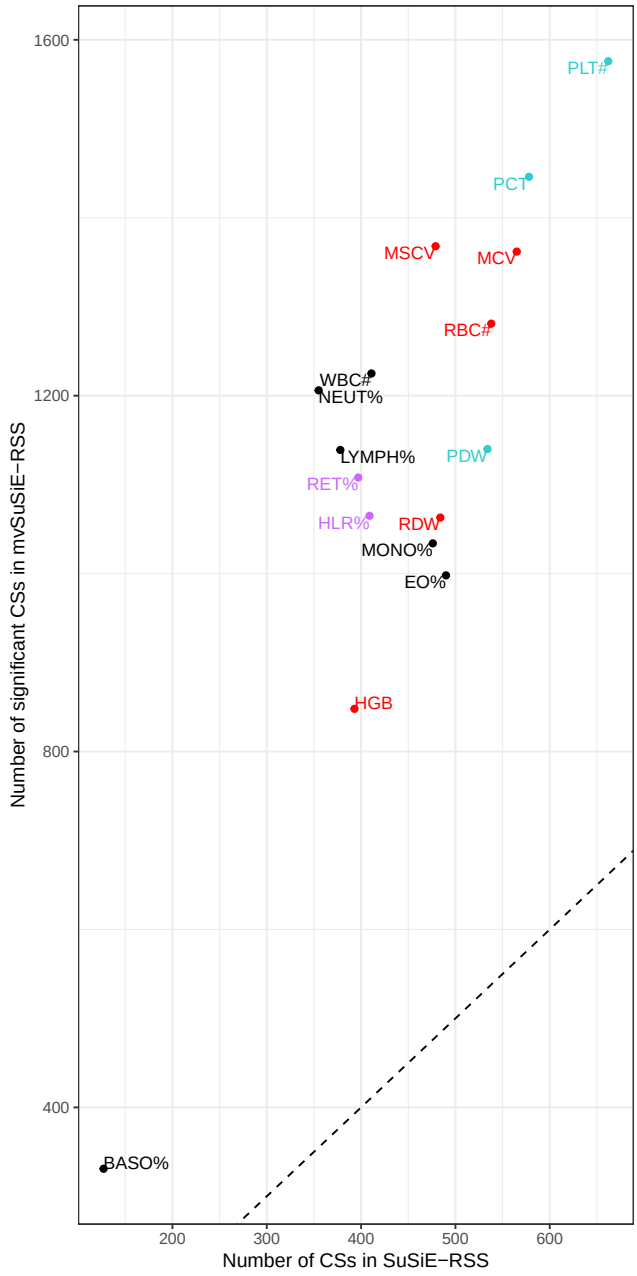


Figure 4.15: **Comparison of number of significant *CSs* in *mvSuSiE-RSS* vs number of *CSs* in *SuSiE-RSS*.** The traits are colored coded by cell types. The dashed line represents $y = x$.

Table 4.3: **Summary of single-trait fine-mapping results from *SuSiE-RSS*.**

traits	number of *CSs*	median of *CSs* size	median of *CSs* purity
WBC#	411	15	0.906
RBC#	538	11	0.926
HGB	393	12	0.918
MCV	565	8	0.948
RDW	484	9.5	0.941
PLT#	662	10	0.952
PCT	578	10	0.950
PDW	534	10	0.950
LYMPH%	378	13	0.906
MONO%	476	9	0.948
NEUT%	355	12	0.910
EO%	490	10.5	0.937
BASO%	127	8	0.947
RET%	397	9	0.932
MSCV	479	8	0.953
HLR%	409	9	0.936

causal variants for some blood traits. In addition, *mvSuSiE-RSS* uses learned patterns of sharing among traits to improve the fine-mapping resolution and the power to detect weakly associated putative causal variants.

AK3 region

At region around AK3 gene, *mvSuSiE-RSS* identified 3 *CSs* (see Figure 4.16, 4.17 and 4.18 to visualize results). There are two putative causal variants (rs12005199, rs409950) with strong signals in platelet cells and weak signals in red blood cells (RBC#, RDW, MCV, MSCV) and white blood cells (WBC#, NEUT%, LYMPH%, MONO%). The evidence for strong signals in platelet cells is consistent with previous studies (Astle et al., 2016; Guo et al., 2017; Ulirsch et al., 2019; Vuckovic et al.,

2020). These two variants decrease PLT#, PCT, NEUT% and WBC#, but increase LYMPH%, MONO%, MCV and MSCV. The two putative causal variants are located close to AK3 and ECM1P1, where AK3 is involved in the pathway for megakaryocyte development and platelet production (Gieger et al., 2011). Another *CS* identified in the region contains 20 SNPs with minimum pairwise correlation 0.968. Because of the high correlation between SNPs, it is difficult to tell them apart. These 20 SNPs have effects in red cells (RBC#, HGB and RET%), and they are all mapped to gene CDC37L1. CDC37L1 is a co-chaperone protein that binds to numerous proteins and promotes their interaction with HSP90 (Scholz et al., 2001). The chaperone HSP90 helps mature the hemoglobin (Hb-β, Hb-γ) in erythroid cells (Ghosh et al., 2018). The posterior effects for these 20 SNPs are shrunk to zero in white blood cells and platelet cells.

From single-trait fine-mapping analysis, there are no *CSs* for RET%; there is one 95% *CS* for HGB with 67 SNPs (Figure 4.19 left panel). In contrast, there are two significant *CSs* for both RET% and HGB using *mvSuSiE-RSS* (Figure 4.19 right panel), because we have included the sharing of signals in red blood cells in *mvSuSiE-RSS* prior. Furthermore, for HGB, the *CS* size reduces from 67 to 20 SNPs in multivariate fine-mapping.

This example shows that *mvSuSiE-RSS* improved fine-mapping resolution (i.e. reduces *CS* size) and identified *CSs* for traits with weak signals by combining information in related traits.

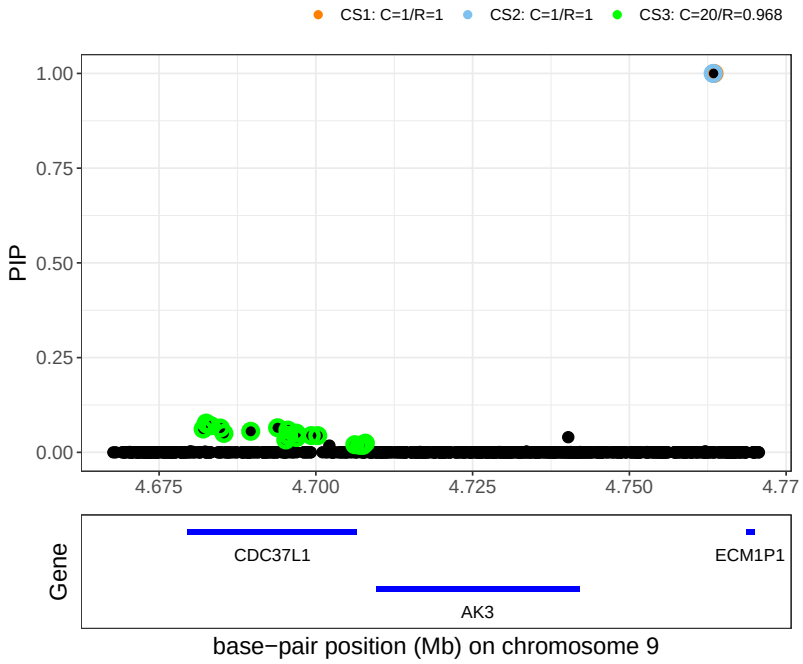


Figure 4.16: **Cross-trait Posterior Inclusion Probability from *mvSuSiE-RSS* for AK3 region.** The *CSs* are color coded. The *CS* 1 contains rs12005199 (chr9: 4763491), which is very close to rs409950 (chr9: 4763368) in *CS* 2. The *CS* 3 contains 20 SNPs with minimum pairwise correlation 0.968.

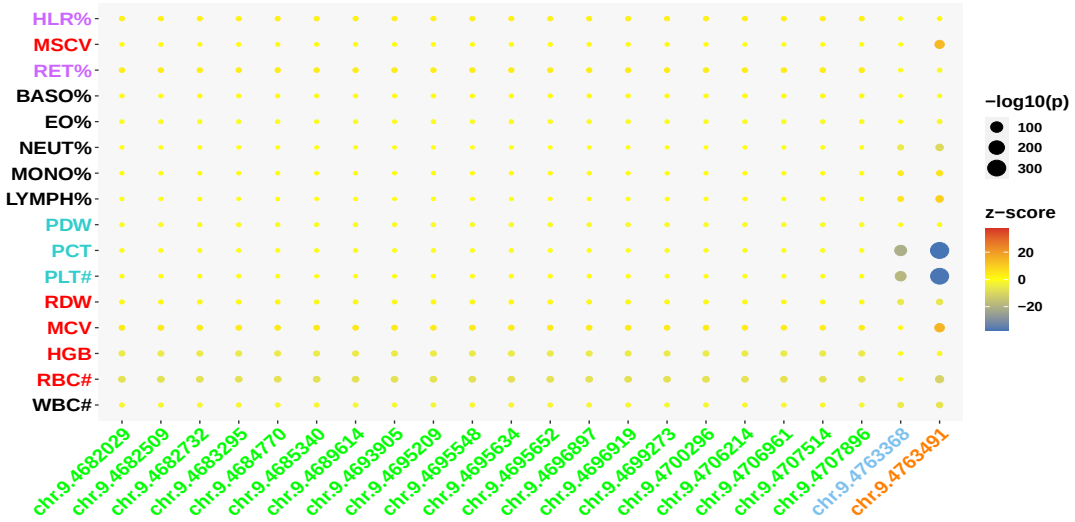


Figure 4.17: **Observed z scores for SNPs in the identified *CSs* for AK3 region.** The locations for variants are color coded by *CSs*. The color of bubble represents effect size and the size of bubble represents $-\log_{10}$(p value).

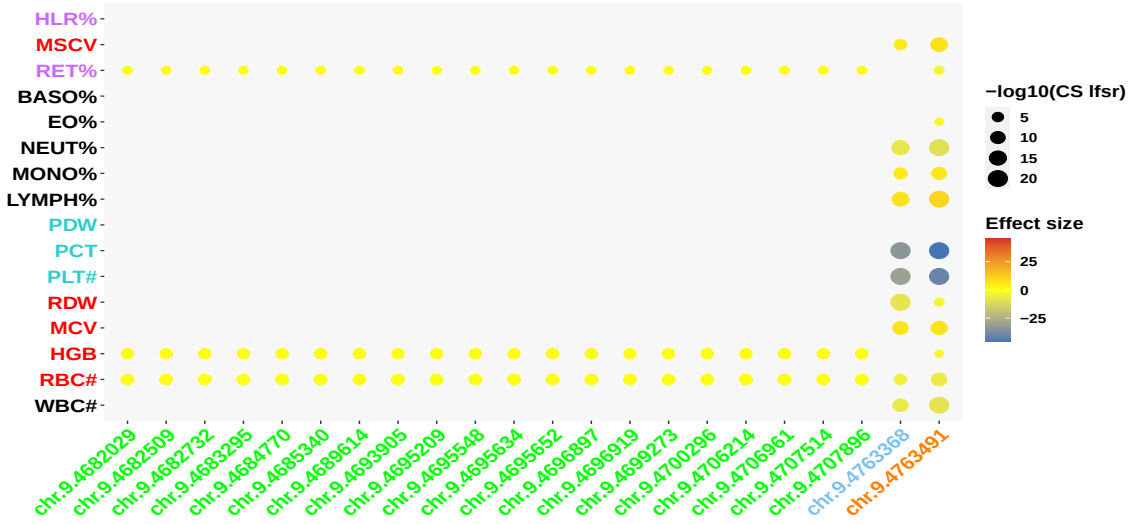


Figure 4.18: **Posterior effects for SNPs in the identified *CSs* for AK3 region.** The locations for variants are color coded by *CSs*. The color of bubble represents posterior effect size and the size of bubble represents $-\log_{10}(CS\ lfsr)$. For each trait, we plot bubbles with significant *CSs*.

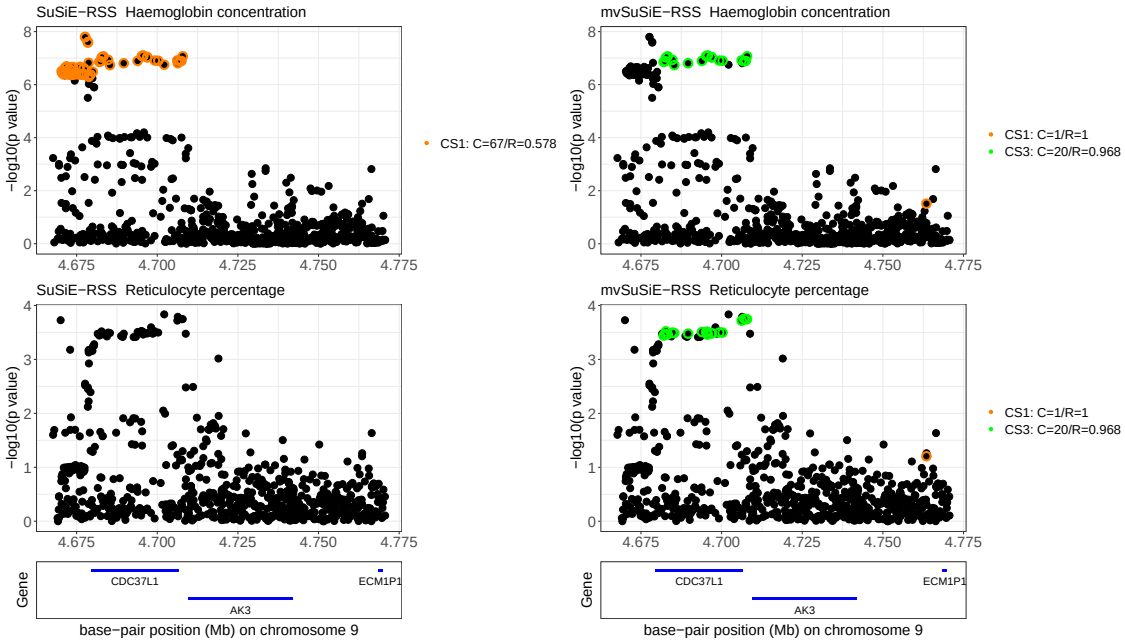


Figure 4.19: **Associations for haemoglobin and reticulocyte percentage from original GWAS for AK3 region.** The left panel shows the identified *CSs* using *SuSiE-RSS*. The right panel shows the identified *CSs* using *mvSuSiE-RSS*.

GLIS3 region

In single-trait fine-mapping, *SuSiE-RSS* identified the known causal variant rs6415788 associated with RBC# and HGB (Astle et al., 2016; Vuckovic et al., 2020); there are no *CSs* for other blood traits. However, in multivariate fine-mapping, there are two more *CSs* besides the *CS* with rs6415788 (Figure 4.20, 4.21 and 4.22). The *CS* 2 and 3 provide putative causal variants for platelet and white blood cells. The variant rs7033677 in *CS* 2 has been previously found evidence for association with lymphocyte cells (Vuckovic et al., 2020). By combining information among traits, the variant rs7033677 has shared effects among white blood cells (WBC#, LYMPH%, NEUT%), platelet cells, MCV and MSCV.

All three *CSs* are around the gene GLIS3, one is significant in red blood cells, the other two are significant in platelet and white blood cells. The discordance of the significant traits among *CSs* is likely explained by the role of GLIS3. GLIS3 has an essential role in thyroid hormone biosynthesis and thyroid gland growth (Rurale et al., 2018). Thyroid dysfunction induces changes in both red blood cell count and white blood cell count (Dorgalaleh et al., 2013).

This example shows that *mvSuSiE-RSS* is able to identify novel putative causal variants by leveraging signals association strength across all traits.

4.7 Discussion

We have introduced an efficient multi-trait fine-mapping method that accounts for complicated effect heterogeneity across phenotypes. Our method outperforms the

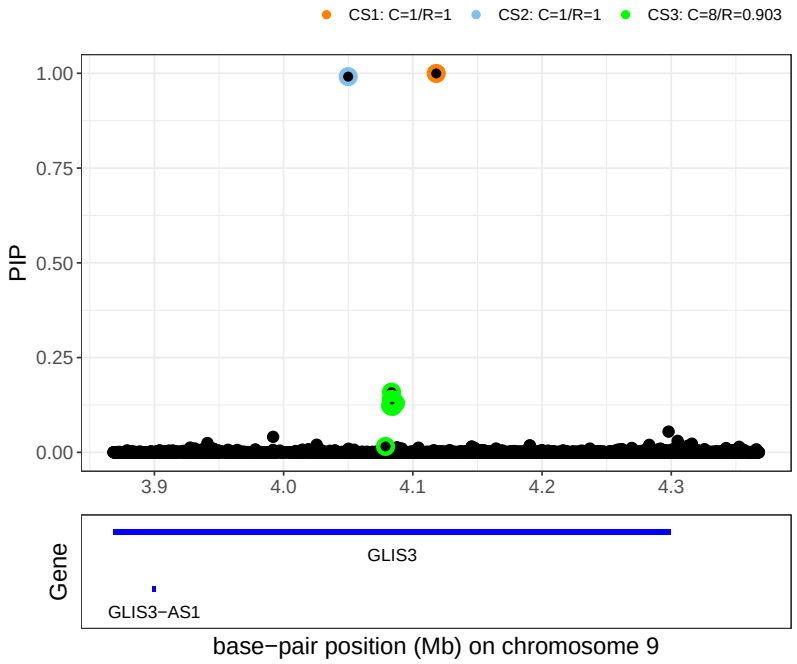


Figure 4.20: **Cross-trait Posterior Inclusion Probability from *mvSuSiE-RSS* for GLIS3 region.** The *CSs* are color coded. The *CS* 1 contains rs6415788 (chr9: 4118111). The *CS* 2 contains rs7033677 (chr9: 4049942). The *CS* 3 contains 8 SNPs with minimum pairwise correlation 0.903.

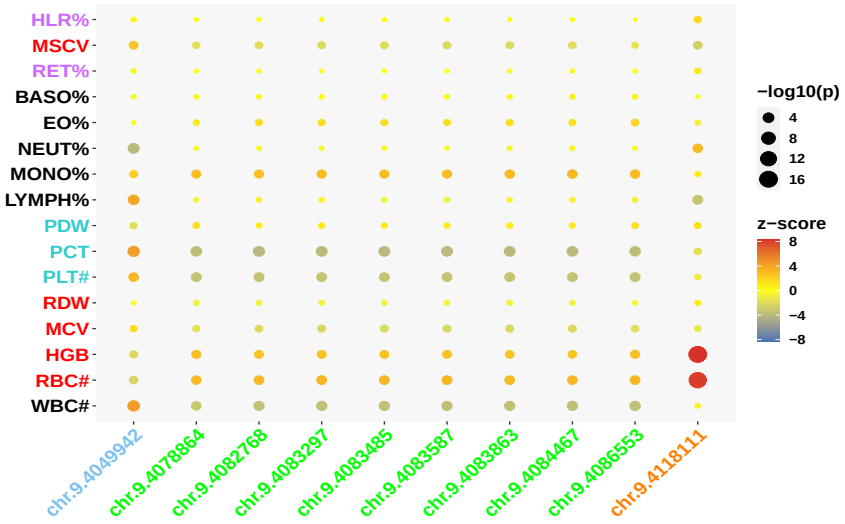


Figure 4.21: **Observed z scores for SNPs in the identified *CSs* for GLIS3 region.**

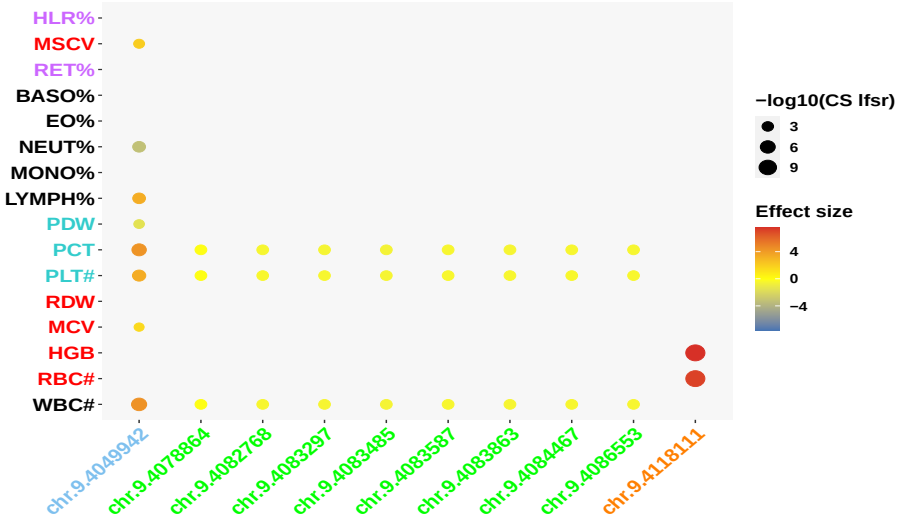


Figure 4.22: **Posterior effects for SNPs in the identified *CSs* for GLIS3 region.**

single phenotype fine-mapping methods in both power and the resolution to fine map causal effects. Compared to the multi-trait fine-mapping method *PAINTOR*, our method is magnitudes faster and more powerful. *mvSuSiE* and *mvSuSiE-RSS* are flexible, allowing for both trait-specific and shared effects among phenotypes. It includes the fixed effects model, which assumes equal effects in all traits, and the random effects model, which allows for different effect sizes among traits (Han and Eskin, 2011), as special cases.

One special application of *mvSuSiE-RSS* is the colocalization problem. Colocalization is aimed at determining whether two traits share a causal variant in a genomic region. Giambartolomei et al. (2014) proposed coloc and tested colocalization between pairs of traits under an assumption that at most one causal variant per trait exists in the region. The single causal variant assumption is convenient but not realistic. To allow multiple causal variants per trait, there are methods using exhaustive search (Hormozdiari et al., 2016), deterministic approximation of posteriors (Wen et al., 2017) or conditioning analysis (Wallace, 2020). Recently, Wallace (2021) combined *SuSiE* with coloc. They decomposed multiple signals using *SuSiE* for each trait and applied multiple coloc comparisons. Using *mvSuSiE-RSS*, we can parameterize the colocalization problem as a fine-mapping problem using 2 traits with some specific priors (e.g. shared effect with different heterogeneity).

The present model focuses on fine-mapping trait measurements that are performed in a single population. It would be useful to extend our method to do cross-population fine-mapping. The distinct LD structures in each population pose a challenge in cross-population fine-mapping. The current approaches assume shared sig-

nals across all populations, and account for the heterogeneity of effect sizes between populations using a random effects model (Kichaev and Pasaniuc, 2015; LaPierre et al., 2020). Extending our method to do cross-population fine-mapping could further improve the power and precision to detect sharing/population-specific effects.

4.8 Author Contribution

G.W. developed and implemented *mvSuSiE* in mvsusieR package. Y.Z. derived and implemented *mvSuSiE* with missing values in phenotypes. Y.Z. developed and implemented *mvSuSiE-suff* and *mvSuSiE-RSS*. Y.Z. conducted simulations and analyses. P.C. provided suggestions for data process steps. M.S. supervised the research.

CHAPTER 5

ENHANCEMENTS FOR MASH

Multivariate adaptive shrinkage (`mash`) is a method to estimate and compare many effects across multiple conditions jointly (Urbut et al., 2019). This method allows for arbitrary correlations in effect sizes among conditions, and adapts to the patterns present in the data set being analyzed. It provides estimates of effect sizes with measures of uncertainty. In this chapter, we introduce two enhancements for `mash`. The two enhancements are all related to the error correlations in the model.

The effects measured in different conditions could be correlated because of sample overlap among conditions, non-removed environmental effects etc. Failing to include the error correlations in the model increases false discoveries (Section 4.3.3). In `mash`, there is a parameter to include the error correlations among effects in different conditions. Urbut et al. (2019) estimated the parameter using a simple ad hoc method. It is estimated as the empirical correlation matrix of the z scores for those effects close to null. Our first enhancement is to consider an alternative way to estimate the error correlation matrix. The estimated error correlation matrix from the enhancement method provides a better `mash` fit.

The `mash` model is useful when there is an obvious way to define an "effect" in each condition, e.g. effects of expression quantitative trait loci in multiple tissues. When there is no obvious "effect" in each condition, it is common to estimate the change in some quantity computed in multiple conditions over a common baseline level. Such analyses are common in differential gene expression studies (Katsel et al., 2005; McCarthy et al., 2012; Tang et al., 2015). However, comparing expressions in

all conditions to the same baseline level induces correlations in effect errors. Urbut (2017) extended the `mash` model to account for the correlation induced by comparing all conditions with the same control condition, the method is called `mash commonbaseline`. Our second enhancement is to implement `mash commonbaseline` in the mashr package `https://github.com/stephenslab/mashr`. We further extend `mash commonbaseline` to compare all conditions with the mean over conditions.

A quick review for `mash` is in Section 5.1. The first enhancement is in Section 5.2. The enhancement about `mash commonbaseline` is in Section 5.3.

5.1 A review of the `mash` model

Let b_{jr} ($j = 1, \cdots, J$; r $= 1, \cdots, R$) denote the true effect of gene j in condition r. Further let $\hat{b}_{jr}$ denote the observed estimate of this effect, and $\hat{s}_{jr}$ denote the standard error of this estimate. Let $\hat{\boldsymbol{B}}$, $\boldsymbol{B}$ and $\hat{\boldsymbol{S}}$ denote the corresponding $J \times R$ matrices, and let $\boldsymbol{b}_j$, $\hat{\boldsymbol{b}}_j$ denote the j-th row of $\boldsymbol{B}$ and $\hat{\boldsymbol{B}}$.

The `mash` model assumes the genes are independent, the vector $\hat{\boldsymbol{b}}_j$ is normally distributed about the true effects $\boldsymbol{b}_j$ with variance-covariance matrix $\hat{\boldsymbol{S}}_j \boldsymbol{C} \hat{\boldsymbol{S}}_j$. The true effects follow a mixture of multivariate normals. That is,

$$\hat{\boldsymbol{b}}_j | \boldsymbol{b}_j, \hat{\boldsymbol{S}}_j \sim N_R(\boldsymbol{b}_j, \hat{\boldsymbol{S}}_j \boldsymbol{C} \hat{\boldsymbol{S}}_j), \tag{5.1.1}$$

$$\boldsymbol{b}_j | \boldsymbol{\pi}, \mathcal{U} \sim \sum_{k=1}^{K} \sum_{l=1}^{L} \pi_{kl} N_R(\boldsymbol{0}, \omega_l \boldsymbol{U}_k). \tag{5.1.2}$$

Each $\boldsymbol{U}_k$ is a covariance matrix that captures a pattern of effects; each ω_k is a scaling parameter that corresponds to a different effect size; the scaling parameters

$\omega_1, \cdots, \omega_L$ are fixed on a dense grid; $\boldsymbol{C}$ is a full rank correlation matrix that accounts for error correlations among the measurements in the R conditions; $\hat{\boldsymbol{S}}_j$ is the $R \times R$ diagonal matrix with diagonal elements $(\hat{s}_{j1}, \cdots, \hat{s}_{jR})$. The residual correlation matrix $\boldsymbol{C}$ is assumed known or has been estimated from the data (see Section 5.2).

The prior covariance matrices, $\boldsymbol{U}_k$, can be any matrix that represents a potential pattern of effects, including no effect matrix ($\boldsymbol{U}_k = \boldsymbol{0}$). We use a list of both canonical and data-driven covariance matrices to capture the pattern of effects (see Section 4.3.4).

Given $\hat{\boldsymbol{B}}$, $\hat{\boldsymbol{S}}$, $\mathcal{U} = (\boldsymbol{U}_1, \cdots, \boldsymbol{U}_K)$, $\boldsymbol{C}$, `mash` first estimates the unknown weights $\hat{\boldsymbol{\pi}}$ for covariance matrices by maximizing likelihood, which is a convex problem. The posterior distribution for true effect of gene j, $p(\boldsymbol{b}_j|\hat{\boldsymbol{b}}_j, \hat{\boldsymbol{S}}_j, \mathcal{U}, \hat{\boldsymbol{\pi}}, \boldsymbol{C})$, is then computed using Bayes' theorem. The posterior of $\boldsymbol{b}_j$ is

$$\boldsymbol{b}_j|\hat{\boldsymbol{b}}_j, \hat{\boldsymbol{S}}_j, \mathcal{U}, \hat{\boldsymbol{\pi}}, \boldsymbol{C} \sim \sum_{k=1}^{K}\sum_{l=1}^{L} \tilde{\pi}_{jkl} N(\tilde{\boldsymbol{\mu}}_{jkl}, \tilde{\boldsymbol{\Sigma}}_{jkl}), \tag{5.1.3}$$

where

$$\tilde{\boldsymbol{\Sigma}}_{jkl} \coloneqq \omega_l \boldsymbol{U}_k (\boldsymbol{I} + \hat{\boldsymbol{S}}_j^{-1} \boldsymbol{C}^{-1} \hat{\boldsymbol{S}}_j^{-1} \omega_l \boldsymbol{U}_k)^{-1} \tag{5.1.4}$$

$$\tilde{\boldsymbol{\mu}}_{jkl} \coloneqq \tilde{\boldsymbol{\Sigma}}_{jkl} \hat{\boldsymbol{S}}_j^{-1} \boldsymbol{C}^{-1} \hat{\boldsymbol{S}}_j^{-1} \hat{\boldsymbol{b}}_j \tag{5.1.5}$$

$$\tilde{\pi}_{jkl} \coloneqq \frac{\pi_{kl} N_R(\hat{\boldsymbol{b}}_j; \boldsymbol{0}, \hat{\boldsymbol{S}}_j \boldsymbol{C} \hat{\boldsymbol{S}}_j + \omega_l \boldsymbol{U}_k)}{\sum_{k',l'} \pi_{k'l'} N_R(\hat{\boldsymbol{b}}_j; \boldsymbol{0}, \hat{\boldsymbol{S}}_j \boldsymbol{C} \hat{\boldsymbol{S}}_j + \omega_{l'} \boldsymbol{U}_{k'})}. \tag{5.1.6}$$

To measure "significance" of an effect b_{jr}, we use the local false sign rate (lfsr),

which is defined as

$$\text{lfsr}_{jr} = \min\{p(b_{jr} \geqslant 0|\hat{\boldsymbol{b}}_j, \hat{\boldsymbol{S}}_j, \hat{\boldsymbol{\pi}}, \mathcal{U}), p(b_{jr} \leqslant 0|\hat{\boldsymbol{b}}_j, \hat{\boldsymbol{S}}_j, \hat{\boldsymbol{\pi}}, \mathcal{U})\}. \tag{5.1.7}$$

The lfsr is the probability that we would get the sign of effect incorrect if we were to use our best guess of the sign. Therefore, a small lfsr indicates high confidence in determining the sign of an effect.

5.2 Estimating the residual correlation matrix

Urbut et al. (2019) estimated the residual correlations $\boldsymbol{C}$ using the fact that $\boldsymbol{C}$ is the correlation matrix of the z scores under the null ($\boldsymbol{b}_j = \boldsymbol{0}$). The z score for effect j in condition r is $\hat{z}_{jr} = \hat{b}_{jr}/\hat{s}_{jr}$. The estimated $\boldsymbol{C}$ is

$$\boldsymbol{C} = \frac{1}{|\mathcal{N}|} \sum_{j \in \mathcal{N}} \hat{\boldsymbol{z}}_j \hat{\boldsymbol{z}}_j^\intercal, \tag{5.2.1}$$

where $\mathcal{N} = \{j : \max_r |\hat{z}_{jr}| < 2\}$.

This approach is very simple, but ad hoc. We provide an alternative way to estimate $\boldsymbol{C}$ in `mash` model. We estimate $\boldsymbol{C}$ using an EM algorithm (Dempster et al., 1977).

The complete log-likelihood for $\boldsymbol{C}$ is

$$\begin{aligned}
&\log L(\boldsymbol{C}; \hat{\boldsymbol{B}}, \boldsymbol{B}) \\
&:= \sum_{j=1}^{J} \left[\log N_R(\hat{\boldsymbol{b}}_j; \boldsymbol{b}_j, \hat{\boldsymbol{S}}_j \boldsymbol{C} \hat{\boldsymbol{S}}_j) + \log \sum_{k=1}^{K} \sum_{l=1}^{L} \pi_{kl} N_R(\boldsymbol{b}_j; \boldsymbol{0}, \omega_l \boldsymbol{U}_k) \right] \qquad (5.2.2) \\
&= \sum_{j=1}^{J} -\frac{1}{2} \log |\boldsymbol{C}| - \frac{1}{2} (\hat{\boldsymbol{b}}_j - \boldsymbol{b}_j)^T \hat{\boldsymbol{S}}_j^{-1} \boldsymbol{C}^{-1} \hat{\boldsymbol{S}}_j^{-1} (\hat{\boldsymbol{b}}_j - \boldsymbol{b}_j) + \text{constant}, \qquad (5.2.3)
\end{aligned}$$

where constant denotes all terms that do not depend on $\boldsymbol{C}$.

At E-step, we set the first and second moments of $\boldsymbol{b}_j$,

$$\begin{aligned}
\tilde{\boldsymbol{\mu}}_j := \mathbb{E}(\boldsymbol{b}_j | \hat{\boldsymbol{b}}_j) &= \sum_{k=1}^{K} \sum_{l=1}^{L} \tilde{\pi}_{jkl} \tilde{\boldsymbol{\mu}}_{jkl} \qquad (5.2.4) \\
\mathbb{E}(\boldsymbol{b}_j \boldsymbol{b}_j^\intercal | \hat{\boldsymbol{b}}_j) &= \sum_{k=1}^{K} \sum_{l=1}^{L} \tilde{\pi}_{jkl} (\tilde{\boldsymbol{\Sigma}}_{jkl} + \tilde{\boldsymbol{\mu}}_{jkl} \tilde{\boldsymbol{\mu}}_{jkl}^\intercal) \qquad (5.2.5) \\
Q_j &:= \mathbb{E}((\hat{\boldsymbol{b}}_j - \boldsymbol{b}_j)(\hat{\boldsymbol{b}}_j - \boldsymbol{b}_j)^\intercal | \hat{\boldsymbol{b}}_j) \qquad (5.2.6) \\
&= \hat{\boldsymbol{b}}_j \hat{\boldsymbol{b}}_j^\intercal - \hat{\boldsymbol{b}}_j \tilde{\boldsymbol{\mu}}_j^\intercal - \tilde{\boldsymbol{\mu}}_j \hat{\boldsymbol{b}}_j^\intercal + \mathbb{E}(\boldsymbol{b}_j \boldsymbol{b}_j^\intercal | \hat{\boldsymbol{b}}_j). \qquad (5.2.7)
\end{aligned}$$

Taking expectations of (5.2.3), we have

$$\mathbb{E} \log L(\boldsymbol{C}; \hat{\boldsymbol{B}}, \boldsymbol{B}) = -\frac{J}{2} \log |\boldsymbol{C}| - \frac{1}{2} \sum_{j=1}^{J} \operatorname{tr} \left(\boldsymbol{C}^{-1} \hat{\boldsymbol{S}}_j^{-1} Q_j \hat{\boldsymbol{S}}_j \right) + \text{constant}. \qquad (5.2.8)$$

At M-step, we update $\boldsymbol{C}$ by maximizing (5.2.8) under the constraint that the diagonal of $\boldsymbol{C}$ must be 1, since it is a correlation matrix. However, because of the constraint, finding the optimal value is nontrivial. We describe two approaches below

to do M step update.

5.2.1 Exact update in M step

The constraint on $\boldsymbol{C}$ can be removed by different parameterizations (Pinheiro and Bates, 1996). It is easier to solve an unconstrained optimization problem than constrained problem. We use the spherical parametrization on $\boldsymbol{C}$, which is based on the Cholesky decomposition, $\boldsymbol{C} = \boldsymbol{L}^T\boldsymbol{L}$, where $\boldsymbol{L}$ is an upper triangular matrix. Let $\boldsymbol{L}_i$ be the ith column of $\boldsymbol{L}$, and $\boldsymbol{l}_i$ be the spherical coordinates of the first i elements of $\boldsymbol{L}_i$. We can write the spherical parameterization as

$$\boldsymbol{L}_{i,1} = \boldsymbol{l}_{i,1}\cos(\boldsymbol{l}_{i,2}) \tag{5.2.9}$$

$$\boldsymbol{L}_{i,2} = \boldsymbol{l}_{i,1}\sin(\boldsymbol{l}_{i,2})\cos(\boldsymbol{l}_{i,3}) \tag{5.2.10}$$

$$\boldsymbol{L}_{i,3} = \boldsymbol{l}_{i,1}\sin(\boldsymbol{l}_{i,2})\sin(\boldsymbol{l}_{i,3})\cos(\boldsymbol{l}_{i,4}) \tag{5.2.11}$$

$$\vdots$$

$$\boldsymbol{L}_{i,i-1} = \boldsymbol{l}_{i,1}\sin(\boldsymbol{l}_{i,2})\cdots\cos(\boldsymbol{l}_{i,i}) \tag{5.2.12}$$

$$\boldsymbol{L}_{i,i} = \boldsymbol{l}_{i,1}\sin(\boldsymbol{l}_{i,2})\cdots\sin(\boldsymbol{l}_{i,i}), \tag{5.2.13}$$

where $\boldsymbol{l}_{i,j} \in (0,\pi)$, $i = 2,\cdots,R$, $j = 2,\cdots,i$. It follows that $\boldsymbol{C}_{i,i} = l_{i,1}^2$. Because the diagonal elements of $\boldsymbol{C}$ are one, $\boldsymbol{l}_{i,1} = 1$ for all $i = 1,\cdots,R$.

To have an unconstrained parameter, we define $\boldsymbol{\theta} \in \mathbb{R}^{\frac{(R-1)R}{2}}$ as follows,

$$\theta_{(i-2)(i-1)/2+(j-1)} = \log\left(\frac{\boldsymbol{l}_{i,j}}{\pi - \boldsymbol{l}_{i,j}}\right) \quad i = 2,\cdots,R,\ j = 2,\cdots,i. \tag{5.2.14}$$

Therefore, the parameter $\boldsymbol{C}$ is a function of $\boldsymbol{\theta}$. The optimization problem becomes finding $\boldsymbol{\theta}$ maximizes $\mathbb{E}\left(\log L(\boldsymbol{C}(\boldsymbol{\theta}); \hat{\boldsymbol{B}}, \boldsymbol{B})\right)$ (5.2.8) without any constraints. We solve the unconstrained optimization problem using numerical tools (we use the R function `optim`). We then convert the estimated $\boldsymbol{\theta}$ to $\boldsymbol{C}$.

The reparametrization converts the constrained optimization problem to an unconstrained optimization problem. Although we find the exact solution that maximizes the objective function (5.2.8), it is very slow to find a length $R(R-1)/2$ vector $\boldsymbol{\theta}$ that achieves the maximum. We propose another approach to do the M step.

5.2.2 Ad hoc update in M step

Because $\boldsymbol{C}$ is a correlation matrix, the diagonal of $\boldsymbol{C}$ is required to be 1. If we ignore this constraint on $\boldsymbol{C}$, maximizing (5.2.8) over $\boldsymbol{C}$ (the variance-covariance matrix) is easy. We then convert the estimated variance-covariance matrix to a correlation matrix. That is

$$\hat{\boldsymbol{V}} = \frac{1}{J}\left[\sum_{j=1}^{J} \hat{\boldsymbol{S}}_j^{-1} Q_j \hat{\boldsymbol{S}}_j^{-1}\right] \tag{5.2.15}$$

$$\hat{\boldsymbol{C}} = \boldsymbol{D}^{-1}\hat{\boldsymbol{V}}\boldsymbol{D}^{-1} \tag{5.2.16}$$

$$\boldsymbol{D} = \mathrm{diag}(\sqrt{\hat{\boldsymbol{V}}_{11}}, \cdots, \sqrt{\hat{\boldsymbol{V}}_{RR}}). \tag{5.2.17}$$

We implement the EM update for $\boldsymbol{C}$ using (5.2.16) in the mashr package, because it is simple. But the update does not guarantee to increase the objective function (5.2.8). This is because the objective function (5.2.8) achieves maximum at $\hat{\boldsymbol{V}}$, not

$\hat{\boldsymbol{C}}$. Therefore, we stop the EM algorithm before the objective function (5.2.8) drops.

5.2.3 Numerical Comparisons

We randomly generated 20 error correlation matrices. For each error correlation matrix $\boldsymbol{C}$, we generated effects for 4000 genes in 5 conditions,

$$\hat{\boldsymbol{b}}_j|\boldsymbol{b}_j \sim N_5(\boldsymbol{b}_j, \boldsymbol{C}) \tag{5.2.18}$$

$$\boldsymbol{b}_j \sim \frac{1}{4}\delta_0 + \frac{1}{4}N_5(\boldsymbol{0}, \boldsymbol{e}_1\boldsymbol{e}_1^\intercal) + \frac{1}{4}N_5(\boldsymbol{0}, \boldsymbol{I}) + \frac{1}{4}N_5(\boldsymbol{0}, \boldsymbol{x}\boldsymbol{x}^\intercal), \tag{5.2.19}$$

where δ_0 represents a point mass at 0, $\boldsymbol{x} = (0, 1, 1, 0, 0)^\intercal$, $\boldsymbol{e}_r$ is the unit vector with zeros everywhere except for element r.

We compare the `mash` results using oracle $\boldsymbol{C}$ and estimated $\boldsymbol{C}$ from the simple approach (5.2.1), the exact EM updates (Section 5.2.1), the ad hoc EM updates (5.2.16).

Likelihood

We first compare the `mash` log-likelihood using different $\boldsymbol{C}$. The `mash` log-likelihood is

$$\log P(\hat{\boldsymbol{B}}|\hat{\boldsymbol{S}}, \hat{\boldsymbol{\pi}}, \mathcal{U}, \hat{\boldsymbol{C}}) \coloneqq \sum_{j=1}^{J} \log \sum_{k=1}^{K}\sum_{l=1}^{L} \hat{\pi}_{kl} N(\hat{\boldsymbol{b}}_j; \boldsymbol{0}, \hat{\boldsymbol{S}}_j\hat{\boldsymbol{C}}\hat{\boldsymbol{S}}_j + \omega_l \boldsymbol{U}_k). \tag{5.2.20}$$

Figure 5.1 compares the `mash` log-likelihood using $\hat{\boldsymbol{C}}$ from EM updates against the $\hat{\boldsymbol{C}}$ from simple ad hoc approach (5.2.1). The likelihood ratios are positive for all

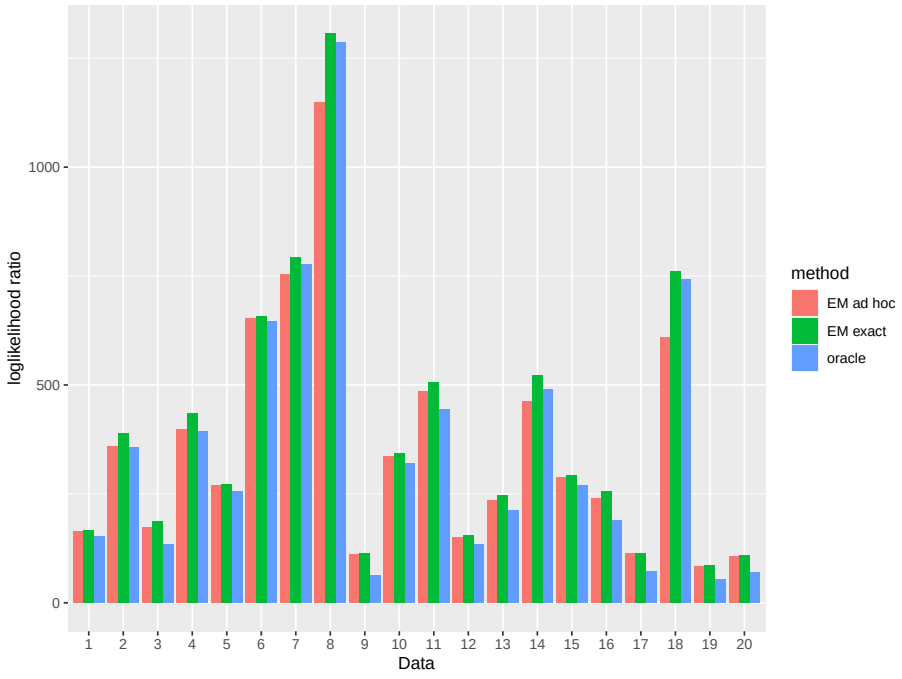


Figure 5.1: **Comparison of log-likelihood ratio using $\hat{C}$ from different methods.** The likelihood ratio compares the model with the one using simple (5.2.1) method.

simulated data sets, so the data favors the estimated $\hat{\boldsymbol{C}}$ from EM updates. The results from exact EM updates and ad hoc EM updates are similar.

Accuracy and Power

We first evaluate the accuracy of the estimated $\hat{\boldsymbol{C}}$ using Frobenius norm from the true value, $\|\hat{\boldsymbol{C}} - \boldsymbol{C}\|_F = \sqrt{\sum_{r,r'}(\hat{\boldsymbol{C}} - \boldsymbol{C})^2_{rr'}}$. The estimated $\hat{\boldsymbol{C}}$ from EM updates are closer to the true value (Table 5.1).

Then, we evaluate the accuracy of `mash` estimated effects by relative root mean squared error (RRMSE). It is defined as the RMSE of the `mash` estimates divided by

Table 5.1: **Frobenius norm between estimated $\hat{C}$ and the true value.**

method	mean of Frobenius norm
simple	0.67
EM exact	0.19
EM ad hoc	0.22

the RMSE achieved by simply using the original observed effects. It can be written as

$$\text{RRMSE} = \sqrt{\frac{\frac{1}{JR}\sum_{j,r}(\boldsymbol{b}_{jr} - \hat{\hat{\boldsymbol{b}}}_{jr})^2}{\frac{1}{JR}\sum_{j,r}(\boldsymbol{b}_{jr} - \hat{\boldsymbol{b}}_{jr})^2}}, \tag{5.2.21}$$

in which $\hat{\hat{\boldsymbol{b}}}_{jr}$ is the posterior mean of true effect. The accuracy of the estimated effects improves with the $\hat{\boldsymbol{C}}$ from EM updates (Figure 5.2a).

Finally, we evaluate the power using ROC curves. The True Positive Rate (TPR) and False Positive Rate (FPR) are computed at any given threshold t as

$$\text{True Positive Rate} = \frac{|CS \cap S|}{|T|} \quad \text{False Positive Rate} = \frac{|N \cap S|}{|N|}, \tag{5.2.22}$$

where S is the set of significant results at threshold t, CS is the set of correctly-signed results, T is the set of true (non-zero) effects and N is the set of null effects:

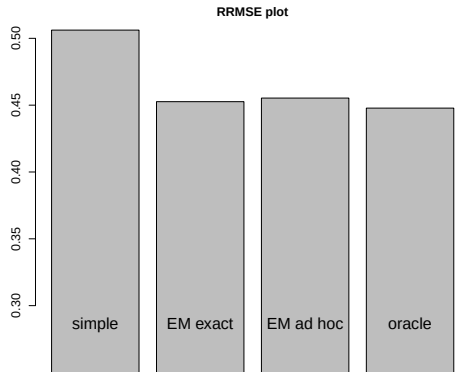
$$S = \{j, r : \text{lfsr}_{jr} \leqslant t\} \tag{5.2.23}$$

$$CS = \{j, r : E(\delta_{jr}|\hat{\Delta}) \times \delta_{jr} > 0\} \tag{5.2.24}$$

$$N = \{j, r : \delta_{jr} = 0\} \tag{5.2.25}$$

$$T = \{j, r : \delta_{jr} \neq 0\}. \tag{5.2.26}$$

(a) Relative Root Mean Square Error (RRMSE)



(b) ROC curve

Figure 5.2: **Power and Accuracy with different estimated $\boldsymbol{C}$.** The left plot compares RRMSE using different $\boldsymbol{C}$. The right plot gives ROC curves for detecting significant genes.

Table 5.2: **Runtimes in seconds.** The running time for the exact EM updates and ad hoc EM updates

method	mean	min	max
EM exact	1364.19	422.35	3087.34
EM ad hoc	96.75	20.74	210.94

The estimated $\hat{\boldsymbol{C}}$ from EM updates have higher power than the simple ad hoc approach (Figure 5.2b).

Runtimes

We compare the running time for the exact EM updates and ad hoc EM updates (Table 5.2). The ad hoc EM updates is roughly 15 times faster than the exact EM updates.

5.2.4 Discussion

The exact EM updates and the ad hoc EM updates give similar results. Because of the long runtimes for the exact EM updates, we implement the ad hoc EM updates in the mashr package (function `mash_estimate_corr_em`).

The ad hoc EM updates needs some time to converge as well, because the `mash` model is fitted at each E step. There are several things we can do to reduce the running time. First of all, we can use a good initial value for $\boldsymbol{C}$. We set it as the estimated $\boldsymbol{C}$ from the simple approach (5.2.1). Moreover, we can set the number of iterations to a small number. Because there is a large improvement in the log-likelihood within the first few iterations, running the algorithm with small number of iterations provides estimates of $\boldsymbol{C}$ that is better than the initial value. Finally, we can estimate $\boldsymbol{C}$ using a random subset of $\hat{\boldsymbol{b}}_j$, $j = 1, \cdots, J$, not the whole observed genes.

5.3 `mash commonbaseline`: Comparing multiple conditions with the same reference level

In this section, we describe the second enhancement for `mash`, `mash commonbaseline`. Suppose we observe estimates of gene expression in R conditions, and we want to estimate the changes in expressions in multiple conditions relative to a common baseline level.

One method to jointly model expression deviations across all conditions is `Cormotif` (Wei et al., 2015). `Cormotif` shares information across conditions to iden-

tify the main patterns of deviations and assigns each gene to one of these deviations patterns. However, `Cormotif` gives no information about deviation size and it assumes the expression deviations are uncorrelated among conditions. It does not include the error correlations induced by comparing with the same baseline level.

Urbut (2017) introduced `mash commonbaseline` to account for the correlation induced by comparing all conditions with the same control condition. Urbut (2017) assumed there is a common control condition in the study and introduced `mash commonbaseline` to estimate changes in multiple conditions relative to the common control condition. However, there might be no control condition in a study. To deal with this case, we define the baseline condition as the expression mean over different conditions. The expression deviation in any condition is then defined as the difference in expression over the mean. We extend `mash commonbaseline` to estimate deviations in multiple conditions relative to the mean.

We describe the `mash commonbaseline` model with a common control condition (Urbut, 2017) in Section 5.3.1. The `mash commonbaseline` model to estimate deviations over the mean is described in subsection 5.3.2. Section 5.3.3 and 5.3.4 show the improvement of the `mash commonbaseline` method through simulations. Section 5.3.5 discusses how to estimate deviations over the median. Section 5.3.6 shows the improvement of the `mash commonbaseline` in a real application.

5.3.1 `mash commonbaseline` with a common control condition

For each gene j, we observe a vector of uncentered noisy mean feature expression $\hat{\boldsymbol{m}}_j$ across R conditions,

$$\hat{\boldsymbol{m}}_j | \boldsymbol{m}_j \sim N_R(\boldsymbol{m}_j, \hat{\boldsymbol{S}}_j \boldsymbol{C} \hat{\boldsymbol{S}}_j), \tag{5.3.1}$$

where $\boldsymbol{m}_j$ represents the "true" means across R conditions. The "true" means $\boldsymbol{m}_j$ follows a mixture of multivariate normals which centered at an underlying mean, $\mu_j \mathbf{1}_R$. Each covariance matrix $\boldsymbol{U}_k$ represents the underlying covariance matrix from which the "true" expression $\boldsymbol{m}_j$ are thought to arise,

$$\boldsymbol{m}_j | \boldsymbol{\pi} \sim \mu_j \mathbf{1}_R + \sum_{k,l} \pi_{kl} N_R(\mathbf{0}, w_l \boldsymbol{U}_k). \tag{5.3.2}$$

Let $\boldsymbol{L}$ denotes the $R-1 \times R$ matrix of contrasts which removes expression in the control condition from each subsequent condition. Suppose the last condition is the control condition, the contrast matrix takes the form:

$$\boldsymbol{L} = \begin{pmatrix} 1 & 0 & 0 & \cdots & -1 \\ 0 & 1 & 0 & & -1 \\ \vdots & & \ddots & & \\ 0 & 0 & \cdots & 1 & -1 \end{pmatrix}_{R-1 \times R} . \tag{5.3.3}$$

Using the contrast matrix $\boldsymbol{L}$, we obtain the expression deviations over the control

condition,

$$\hat{\boldsymbol{\delta}}_j = \boldsymbol{L}\hat{\boldsymbol{m}}_j \sim N_{R-1}(\boldsymbol{L}\boldsymbol{m}_j, \boldsymbol{L}\hat{\boldsymbol{S}}_j\boldsymbol{C}\hat{\boldsymbol{S}}_j\boldsymbol{L}^\intercal). \tag{5.3.4}$$

The "true" deviations, $\boldsymbol{\delta}_j = \boldsymbol{L}\boldsymbol{m}_j$, can be expressed as a zero-centered mixture of multivariate normals,

$$\boldsymbol{\delta}_j|\boldsymbol{\pi} \sim \sum_{k,l} \pi_{kl} N_{R-1}(\boldsymbol{0}, w_l \boldsymbol{U}'_k), \tag{5.3.5}$$

where $\boldsymbol{U}'_k = \boldsymbol{L}\boldsymbol{U}_k\boldsymbol{L}^\intercal$ represents the underlying covariance matrix from which the "true" deviations $\boldsymbol{\delta}_j$ are thought to arise.

Given a matrix of observed mean expression $\hat{\boldsymbol{M}}$, the corresponding standard errors $\hat{\boldsymbol{S}}$, the contrast matrix $\boldsymbol{L}$, the prior covariance matrices $\mathcal{U}$, the correlation matrix $\boldsymbol{C}$, we estimate mixture weights $\hat{\boldsymbol{\pi}}$ by maximum likelihood. The likelihood for $\boldsymbol{\pi}$ is

$$L(\boldsymbol{\pi}) = \prod_{j=1}^{J} \sum_{k,l} \pi_{kl} N_{R-1}(\hat{\boldsymbol{\delta}}_j|\boldsymbol{0}, \boldsymbol{L}\hat{\boldsymbol{S}}_j\boldsymbol{C}\hat{\boldsymbol{S}}_j\boldsymbol{L}^\intercal + w_l\boldsymbol{U}'_k). \tag{5.3.6}$$

This step adapts to patterns present in the data. If most deviations are zero, this step puts most weight on zero effect matrix. If some prior matrices in $\mathcal{U}$ do not help capture patterns in the data, they will receive little weight. With the estimated weights $\hat{\boldsymbol{\pi}}$, we get the posterior distribution of true deviations $\boldsymbol{\delta}_j$, which can be used for estimation and inference of $\boldsymbol{\delta}_j$.

The $\hat{\boldsymbol{\delta}}_j$ and $\boldsymbol{\delta}_j$ can be treated as the observed effects $\hat{\boldsymbol{b}}_j$ and true effects $\boldsymbol{b}_j$ in the `mash` model. The critical step above is the residual covariance of the observed deviations. Even if the original residual covariance matrix ($\hat{\boldsymbol{S}}_j\boldsymbol{C}\hat{\boldsymbol{S}}_j$) is diagonal, and thus the observed noisy mean expression measurements in each condition are inde-

pendent, $\boldsymbol{L}\hat{\boldsymbol{S}}_j\boldsymbol{C}\hat{\boldsymbol{S}}_j\boldsymbol{L}^\intercal$ is not diagonal and thus accounts for the induced correlation in errors.

Dependence of $\hat{\boldsymbol{m}}_j$ on $\hat{\boldsymbol{S}}_j$

The model (5.3.1) assumes $\hat{\boldsymbol{m}}_j$ are independent of their standard errors $\hat{\boldsymbol{S}}_j$, and it is referred as the "exchangeable effects" (EE) model (Wen and Stephens, 2014). We can generalize this assumption that the expression may scale with standard error, so that expressions with larger standard error tend to be larger,

$$\hat{\boldsymbol{S}}_j^{-\alpha}\hat{\boldsymbol{m}}_j|\boldsymbol{m}_j \sim N_R(\hat{\boldsymbol{S}}_j^{-\alpha}\boldsymbol{m}_j, \hat{\boldsymbol{S}}_j^{1-\alpha}\boldsymbol{C}\hat{\boldsymbol{S}}_j^{1-\alpha}). \tag{5.3.7}$$

Setting $\alpha = 0$ yields (5.3.1). Setting $\alpha > 0$ implies that the expressions with larger standard error tend to be larger (in absolute value). The contrast matrix $\boldsymbol{L}$ can be applied on model (5.3.7). The posterior inference is for changes in the scaled quantity (by the standard error) in multiple conditions over a common control condition.

5.3.2 `mash commonbaseline` *without a common control condition*

In the case there is an obvious control group in the study, we estimate the deviation over the control condition. When there is no control group in the study, we estimate the deviation over the mean of different conditions. The contrast matrix takes the

form:

$$\boldsymbol{L} = \begin{pmatrix} \frac{R-1}{R} & -\frac{1}{R} & -\frac{1}{R} & \cdots & -\frac{1}{R} \\ -\frac{1}{R} & \frac{R-1}{R} & -\frac{1}{R} & & -\frac{1}{R} \\ \vdots & & \ddots & & \\ -\frac{1}{R} & -\frac{1}{R} & \cdots & \frac{R-1}{R} & -\frac{1}{R} \end{pmatrix}_{R-1 \times R} . \tag{5.3.8}$$

Note that the contrast matrix $\boldsymbol{L}$ has only $R-1$ rows. This requirement is necessary, because the `mash` framework requires the correlation matrix among observed deviations to be full rank. When there is a control condition, the deviation for the control condition is always zero. When we compare the expression with the mean, any deviation can be expressed using the remaining deviations, i.e. $\hat{\delta}_{j,i} := \hat{m}_{j,i} - \bar{\hat{m}}_j = -\sum_{r=1,r\neq i}^{R}(\hat{m}_{j,r} - \bar{\hat{m}}_j) = -\sum_{r=1,r\neq i}^{R}\hat{\delta}_{j,r}$. Therefore, we must discard the deviation in one condition to have a non-degenerate model.

The contrast matrix $\boldsymbol{L}$ defined in (5.3.8) discards the deviation in the last condition. The deviations are $\hat{m}_{j,1} - \bar{\hat{m}}_j, \hat{m}_{j,2} - \bar{\hat{m}}_j, \cdots, \hat{m}_{j,R-1} - \bar{\hat{m}}_j$. However, the contrast matrix $\boldsymbol{L}$ can discard any condition from $\hat{m}_{j,1} - \bar{\hat{m}}_j, \cdots, \hat{m}_{j,R} - \bar{\hat{m}}_j$, and the results are similar. (For the canonical priors, the results are identical in principle; for data-driven priors, the result depends on the way the priors are learnt.)

The posterior distribution is only for the deviations in the first $R-1$ conditions. Using a linear transformation of the posteriors, we obtain the posterior for all R

conditions. The linear transformation corresponding to contrast matrix (5.3.8) is

$$\boldsymbol{A} = \left(\begin{array}{c} \boldsymbol{I}_{R-1} \\ \hline -1 \quad \cdots \quad -1 \end{array} \right)_{R\times(R-1)} . \tag{5.3.9}$$

Using the linear transformation, we have the posterior for all R conditions,

$$\boldsymbol{A}\boldsymbol{\delta}_j = \left(\delta_{j,1}, \cdots, \delta_{j,R-1}, -\sum_{r=1}^{R-1} \delta_{j,r} \right). \tag{5.3.10}$$

There is a drawback when we estimate the deviation over the mean. We discuss it in detail in Section 5.3.5, and propose methods to estimate the deviation over the median.

5.3.3 Simple Simulation with a control condition

Urbut (2017) demonstrated that failing to account for the correlations induced by comparing with the same control condition inflates false discoveries. We conducted similar simulations as in Urbut (2017) with our `mash commonbaseline` implementation in mashr package. We compare the results from our method `mash commonbaseline` with the one from `mash` ignoring the induced correlations, we call this `independent mash` model. The `independent mash` model is analogous to `Cormotif`, which fails to account for the induced correlations.

In the following simulations, we treat the last condition as the control condition.

Simulation without Deviation

There is no true deviation exists in this simulation. We simulated samples with identical mean expression in all conditions.

$$\hat{\boldsymbol{m}}_j \sim N_{10}(\boldsymbol{m}_j, \frac{1}{2}\boldsymbol{I}), \tag{5.3.11}$$

$$\boldsymbol{m}_j = \mu_j \mathbf{1}_{10}. \tag{5.3.12}$$

Let $\boldsymbol{L}$ be the contrast matrix as defined in (5.3.3). The `mash commonbaseline` uses the following model,

$$\hat{\boldsymbol{\delta}}_j \sim N_9(\boldsymbol{\delta}_j, \frac{1}{2}\boldsymbol{L}\boldsymbol{L}^T). \tag{5.3.13}$$

However, one might subtract the expression in the control condition from every subsequent condition, and ignore the induced correlations, which leads to the `independent mash` model,

$$\hat{\boldsymbol{\delta}}_j \sim N_9(\boldsymbol{\delta}_j, \boldsymbol{I}), \tag{5.3.14}$$

where the variance of $\hat{\delta}_{jr}, r = 1, \cdots, R-1$, is calculated by

$$\mathrm{Var}(\hat{m}_{jr} - \hat{m}_{jR} | m_{jr}, m_{jR}) = \mathrm{Var}(\hat{m}_{jr}|m_{jr}) + \mathrm{Var}(\hat{m}_{jR}|m_{jR}) = \frac{1}{2} + \frac{1}{2} = 1. \tag{5.3.15}$$

The correlation between $\hat{\delta}_{jr}$ and $\hat{\delta}_{js'}$, $r \neq r'$, is ignored.

The `mash commonbaseline` model yields a much higher log-likelihood (-108173.1 vs -115816.6 from `independent mash` model). Including the induced correlations,

there are no discoveries. This is expected because the true deviations, $\boldsymbol{\delta}_j$, are zero for all samples. However, the `independent mash` model produces around 35% discoveries, which are all false discoveries.

In both `mash commonbaseline` and `independent mash`, we used a list of canonical covariance matrices as priors. The `mash commonbaseline` method correctly puts the majority of the mixture weights on the null matrix. In contrast, the `independent mash` model puts the majority of mixture weights on the equal effect matrix. This is caused by ignoring of correlations in errors. From this simulation, we see the improvement of false discoveries by `mash commonbaseline` clearly.

Simulation with Deviation

We added signals to a number of "non-null" genes such that there are deviations from the control group in at least one subgroup $r = 1, \cdots, R-1$,

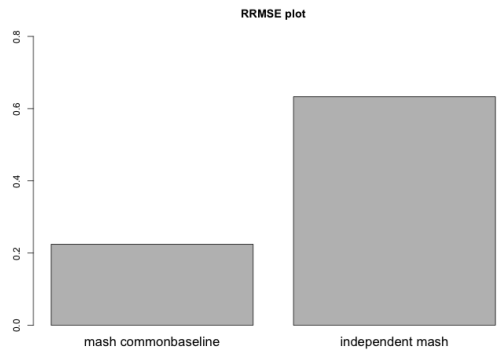
$$\hat{\boldsymbol{m}}_j | \boldsymbol{m}_j \sim N_R(\boldsymbol{m}_j, \frac{1}{2}\boldsymbol{I}), \tag{5.3.16}$$

$$\boldsymbol{m}_{j1\ldots(R-1)} = m_{jR}\mathbf{1}_{R-1} + \boldsymbol{\delta}_j. \tag{5.3.17}$$

We simulated data with 10 conditions and four different types of deviations $\boldsymbol{\delta}_j$: null ($\boldsymbol{\delta}_j = 0$), independent among conditions, condition-specific in condition 1 ($\delta_{j1} \neq 0$), and shared (equal deviations in all sub-conditions, $\delta_{j1\ldots9} = \mathbf{1}_9$). The data contained 10% non-null deviations,

$$\boldsymbol{\delta}_j \sim \frac{9}{10}N_9(\mathbf{0}, \mathbf{0}) + \frac{1}{30}N_9(\mathbf{0}, \boldsymbol{I}) + \frac{1}{30}N_9(\mathbf{0}, \mathbf{e}_1\mathbf{e}_1^T) + \frac{1}{30}N_9(\mathbf{0}, \mathbf{1}\mathbf{1}^T). \tag{5.3.18}$$

(a) Relative Root Mean Squared Error (RRMSE)



(b) ROC curve

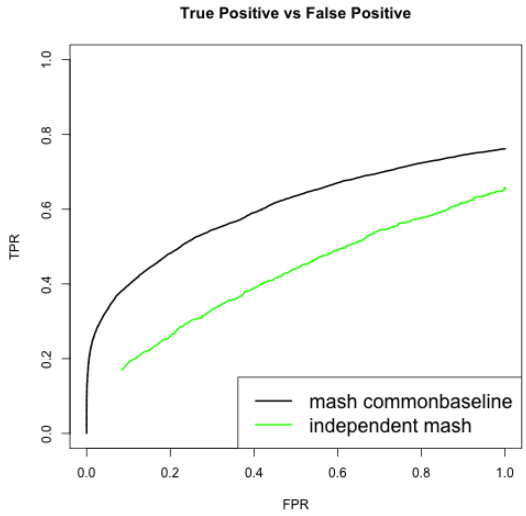


Figure 5.3: **Power and Accuracy for different methods.** The deviations are computed over the common control group. The left plot compares RRMSE from different models. The right plot gives ROC curves for detecting significant genes.

The `mash commonbaseline` outperforms the `independent mash` model in both power and accuracy (Figure 5.3).

5.3.4 Simple Simulation without a control condition

In this section, we show how the methods perform when every condition is compared to the mean of all conditions.

Simulation without Deviation

The simulation scheme is similar to Section 5.3.3. We simulated genes with identical mean expression in all conditions.

$$\hat{\boldsymbol{m}}_j|\boldsymbol{m}_j \sim N_R(\boldsymbol{m}_j, \boldsymbol{I}), \tag{5.3.19}$$

$$\boldsymbol{m}_j = \mu_j \mathbf{1}_R. \tag{5.3.20}$$

Let $\boldsymbol{L}$ be the contrast matrix as defined in (5.3.8), the `mash commonbaseline` has the model

$$\hat{\boldsymbol{\delta}}_j|\boldsymbol{\delta}_j \sim N_{R-1}(\boldsymbol{\delta}_j, \boldsymbol{L}\boldsymbol{L}^\intercal). \tag{5.3.21}$$

In the `independent mash` model, the correlation among $\hat{\delta}_{jr}$ and $\hat{\delta}_{jr'}$ is ignored,

$$\hat{\boldsymbol{\delta}}_j|\boldsymbol{\delta}_j \sim N_{R-1}(\boldsymbol{\delta}_j, \boldsymbol{S}). \tag{5.3.22}$$

where $\boldsymbol{S}$ is a diagonal matrix with diagonal elements

$$\mathrm{Var}(\hat{\boldsymbol{m}}_{j,r} - \bar{\hat{m}}_j|\boldsymbol{m}_j) = \frac{R-1}{R}. \tag{5.3.23}$$

With $R = 10$ simulated conditions, there are no discoveries from both models. The `mash commonbaseline` model yields higher log-likelihood (-116342.4 vs -123179 from `independent mash` model). However, with $R = 3$ simulated conditions, we observe the inflation in false discoveries for `independent mash`. There are no false

discoveries for `mash commonbaseline`, but the `independent mash` model produces around 1.4% false discoveries. This is because when we compute deviations over the mean, the error correlation between the deviations depends on the number of conditions.

Consider a simple example that the standard error is common among different conditions and the measurements in different conditions are independent,

$$\text{Var}(\hat{\boldsymbol{m}}_j|\boldsymbol{m}_j) = s^2\boldsymbol{I}_R. \tag{5.3.24}$$

The variance of the deviation over the mean is

$$\text{Var}(\hat{m}_{jr} - \bar{\hat{\boldsymbol{m}}}_j|\boldsymbol{m}_j) = \frac{R-1}{R}s^2. \tag{5.3.25}$$

The covariance between two deviations is

$$Cov(\hat{m}_{jr} - \bar{\hat{\boldsymbol{m}}}_j, \hat{m}_{jr'} - \bar{\hat{\boldsymbol{m}}}_j|\boldsymbol{m}_j) = -\frac{s^2}{R}, \tag{5.3.26}$$

which depends on the number of conditions. As the number of conditions increases, the induced error correlations become weaker and it becomes negligible.

Simulation with Deviation

We simulated data with 10 conditions, half of the samples had equal expression among conditions. In the remaining samples, half had higher and equal expression

in the first 2 conditions, half had higher expression in the last condition,

$$\boldsymbol{m}_j \sim \mu_j \mathbf{1}_{10} + \frac{1}{2} N_{10}(\mathbf{0}, \mathbf{0}) + \frac{1}{4} N_{10}(\mathbf{0}, 9 \begin{pmatrix} \mathbf{1}_2 \mathbf{1}_2^\intercal & \mathbf{0}_{2\times 8} \\ \mathbf{0}_{8\times 2} & \mathbf{0}_{8\times 8} \end{pmatrix}) + \frac{1}{4} N_{10}(\mathbf{0}, 9 \begin{pmatrix} \mathbf{0}_{9\times 9} & \mathbf{0}_9 \\ \mathbf{0}_9^\intercal & 1 \end{pmatrix}), \tag{5.3.27}$$

$$\hat{\boldsymbol{m}}_j | \boldsymbol{m}_j \sim N_{10}(\boldsymbol{m}_j, \boldsymbol{I}). \tag{5.3.28}$$

Let $\boldsymbol{L}$ be the contrast matrix in (5.3.8) that subtract the mean from each sample, the deviations are

$$\hat{\boldsymbol{\delta}}_j | \boldsymbol{\delta}_j \sim N_9(\boldsymbol{\delta}_j, \boldsymbol{L}\boldsymbol{L}^\intercal). \tag{5.3.29}$$

Half of the true deviations were zero, quarter of the deviations had correlation that the first two conditions were negatively correlated with the rest conditions. For the remaining quarter of the deviations, the first 9 conditions were negatively correlated with the last condition.

We applied three models on the simulated data.

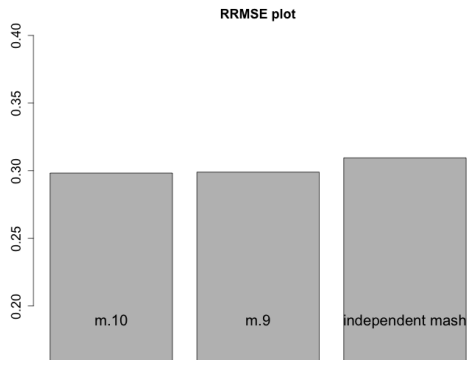
1. `m.10`: the `mash commonbaseline` model uses $\boldsymbol{L}$ (5.3.8) discarding the 10-th condition.

2. `m.9`: the `mash commonbaseline` model uses $\boldsymbol{L}$ (5.3.8) discarding the 9-th condition.

3. `independent mash` model.

To better capture the covariance structures, we fitted models with data-driven covariance matrices.

(a) Relative Root Mean Square Error (RRMSE)



(b) ROC curve

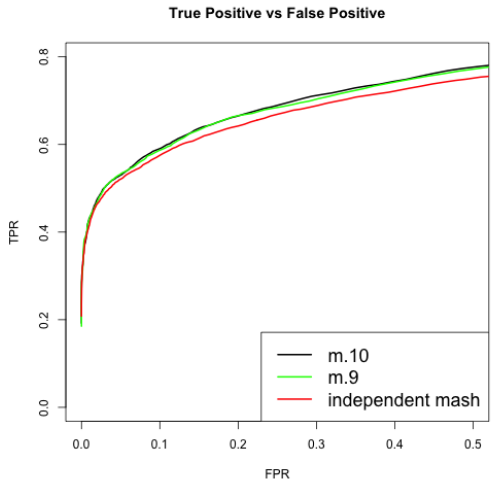


Figure 5.4: **Power and Accuracy for different methods.** The deviations are computed over the mean. The left plot compares RRMSE from different models. The right plot gives ROC curves for detecting significant genes.

From Figure 5.4, the results from `m.9` and `m.10` have similar accuracy and power, which confirms that `mash commonbaseline` is robust to the choice of the discarded condition. The `mash commonbaseline` model performs slightly better than `independent mash`. When the number of conditions is large, the difference is negligible.

5.3.5 Deviation from Median

From the simulations above, it is clear that correctly modeling the inherent correlations when comparing all conditions to a common control group can dramatically reduce false positives. However, when the comparison is made with the mean, the induced correlation becomes negligible as the number of conditions increases. The

reason is that the quantity subtracted from each condition is a summary statistic, which contains information from every condition. As the number of conditions increases, the correlation between deviations over the summary statistics becomes negligible.

However, there is one drawback of using the deviations over the mean. When some conditions have large positive deviations over the mean, the other conditions must have negative deviations. For instance, suppose the first condition has high gene expression level, and all the other conditions have near zero expression levels. Subtracting the mean from each condition leads to high positive deviation in the first condition, small negative deviations in all other conditions. Therefore, all conditions have deviations, but the deviations in the first condition have opposite sign than others. In this case, it is more parsimonious to conclude that the first condition is different from others. It is better to report "condition specific" effect than a shared effect at all but one condition. To achieve this parsimonious statement, we could estimate the change in the quantity computed in R conditions over their median.

Comparing the gene expression with the median among R conditions in the above example, we identify that the first condition has high gene expression level comparing with other conditions. Since subtracting the median would not cause the rank deficiency problem in the correlation matrix $\boldsymbol{C}$, we do not need to discard deviations in any condition. However, there is no contrast matrix exists to get the deviations over median. The variance of median in a multivariate normal distribution is hard to compute as well. To the best of our knowledge, there is no result about the explicit variance of the median exists in literature. Therefore, we cannot use the `mash`

`commonbaseline` model directly. There are several ways to analyze the deviation over the median using `mash`.

1. Subtract median directly from each condition.

   For deviations over the mean, we have observed that when the number of conditions is large, the variance of the deviation is similar to the variance of the observed expression (5.3.25); the covariance between deviations is negligible (5.3.26). We assume the similar property holds for deviations over the median.

   We treat median as a constant and subtract it from each condition without considering the variance and the correlation. We then identify deviations using `mash`, i.e. `independent mash` model.

2. Estimate deviations from posterior samples.

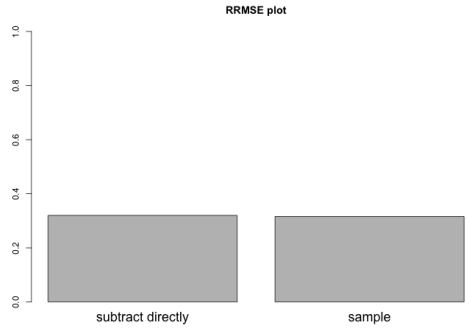
   We first get the posterior samples of deviations over the mean, $\boldsymbol{\delta}_j$, using `mash commonbaseline`. Then we estimate the posterior for deviation over the median using posterior samples,

$$\boldsymbol{m}_j - \text{median}(\boldsymbol{m}_j) = (\boldsymbol{m}_j - \bar{\boldsymbol{m}}_j) - \text{median}(\boldsymbol{m}_j - \bar{\boldsymbol{m}}_j) = \boldsymbol{\delta}_j - \text{median}(\boldsymbol{\delta}_j). \tag{5.3.30}$$

   We summarize the posterior information based on the samples.

We applied the methods above on the simulated data in Section 5.3.4. When there is no deviation in the data, the simulation scheme is same as (5.3.19) and (5.3.20). There is no false discovery using both methods with 10 conditions. We also simulated deviations using scheme (5.3.27), (5.3.28). Figure 5.5 shows the RRMSE and the

(a) Relative Root Mean Square Error (RRMSE)



(b) ROC curve

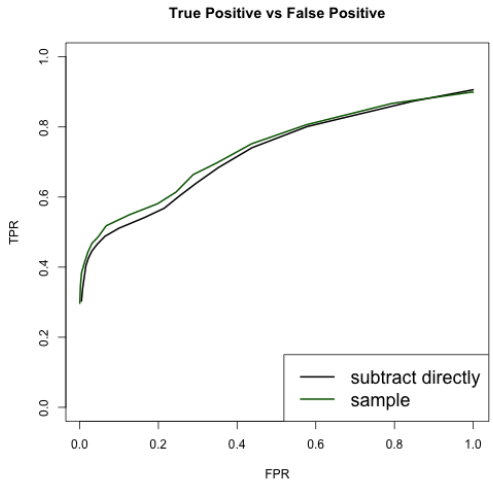


Figure 5.5: **Simulation with Deviation and compare with median.** (a) shows the accuracy of the estimated deviations; (b) shows the ROC curve.

ROC curve. Both methods have RRMSE < 1, indicating a substantial improvement in accuracy compared with the original observed effects $\hat{\boldsymbol{\delta}}_j$. Both methods perform similarly.

5.3.6 Application

Blischak et al. (2015) was interested in identifying genes that are differently expressed in human innate immune cells in response to infection with Mycobacterium tuberculosis stains and related mycobacterial species. The change in gene expression under each infection was defined relative to an uninfected control group. The expressions were measured at different post-infection timepoints. We applied our `mash commonbaseline` approach to data at 18 hours post-infection.

The gene expression readings represent batch-corrected $\log_2$ counts per million

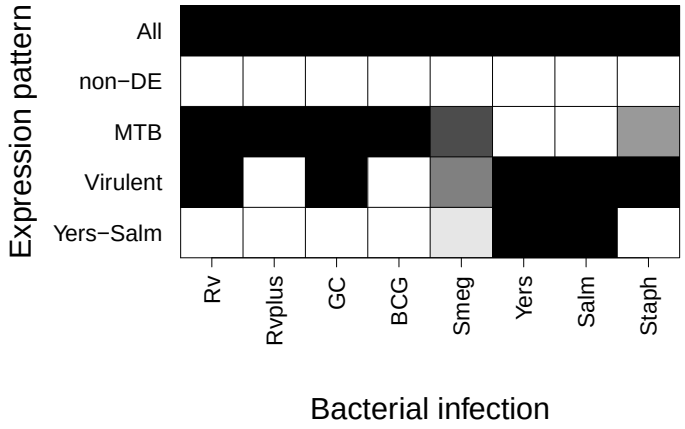


Figure 5.6: **Inferred Patterns of Sharing using `Cormotif`.** The plot shows the five patterns identified in Blischak et al. (2015).

for the 12,728 Ensemble genes, each has 6 samples across 8 infections (MTB H37Rv (`Rv`), heat-inactivated MTB H37Rv (`Rvplot`), MTB GC1237 (`GC`), bacillus Calmette-Guérin (`BCG`), Mycobacterium smegmatis (`Smeg`), Yersinia pseudotuberculosis (`Yers`), Salmonella typhimurium (`Salm`), Staphylococcus epidermidis (`Staph`)) and control (see Blischak et al., 2015, for details). There are 9 conditions in total including the uninfected control group. To obtain the mean expression in each condition for each gene, we used the Empirical Bayes linear model method, Limma (Smyth, 2004), to estimate $\hat{\boldsymbol{m}}_j$ of mean gene expression and corresponding standard errors. The matrix $\hat{\boldsymbol{M}} \in \mathbb{R}^{J \times R}$ contains mean gene expression of gene j in each condition r.

The matrix of observed deviations is $\hat{\boldsymbol{\Delta}} = \hat{\boldsymbol{M}}\boldsymbol{L}^{\intercal}$, where $\boldsymbol{L}$ is defined as in (5.3.3). We used a list of canonical and data-driven covariance matrices as prior and computed posteriors for all genes.

Figure 5.6 shows the patterns identified by `Cormotif` in Blischak et al. (2015). The four differential expressed patterns are collapsed into the primary patterns of

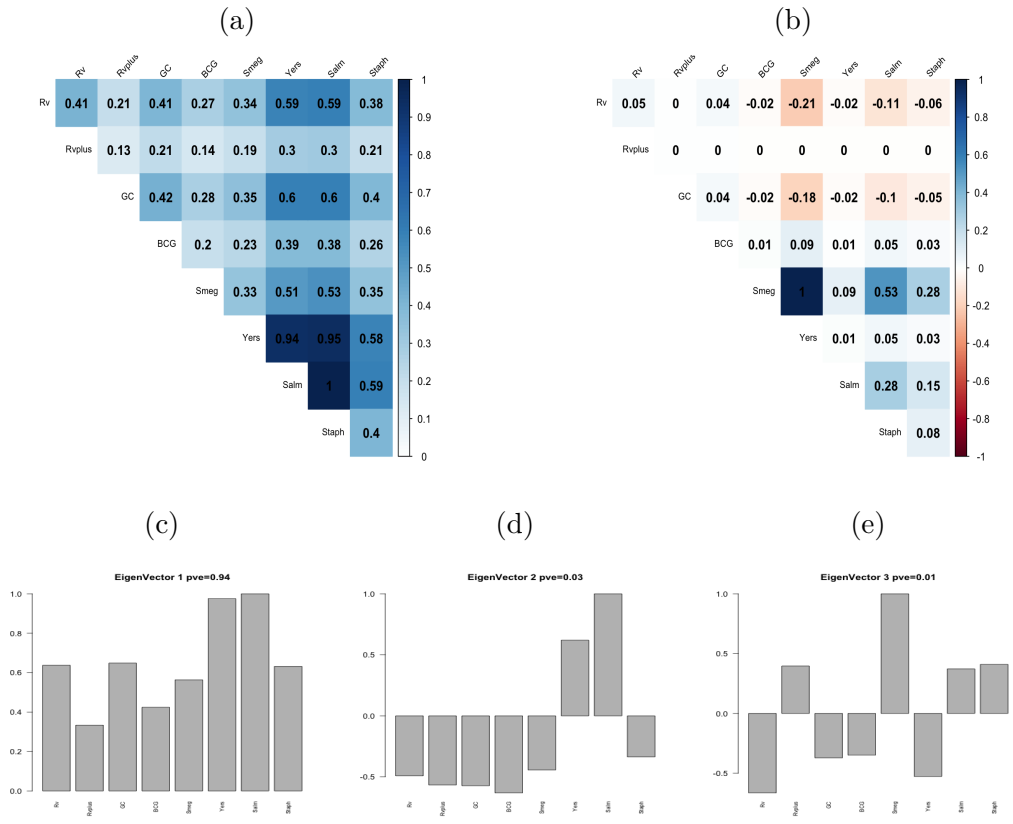
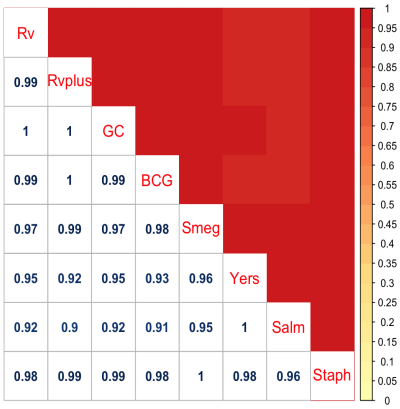


Figure 5.7: **Inferred Patterns of Sharing.** The plot (a) shows the most common pattern of sharing identified in `mash commonbaseline`. The corresponding first three eigenvectors are in (c), (d) and (e). The plot (b) shows another pattern we identified in `mash commonbaseline` model.

(a) Pairwise sharing by Sign



(b) Pairwise sharing by Magnitude

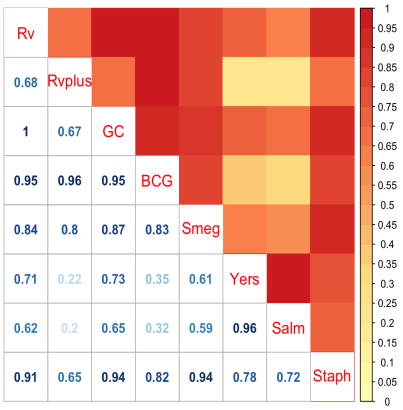


Figure 5.8: **Plots for similarity of deviations by sign and magnitude.** (a) shows the proportion of significant genes that are "shared in sign"; (b) shows the proportion of significant genes that are "shared in magnitude".

sharing in `mash commonbaseline` (Figure 5.7). In `mash commonbaseline`, the majority mixture weight falls on the pattern that reflect broad sharing of both sign and magnitude across infections. `Yers` and `Salm` share deviations very closely and they are strongly correlated with one another (see Figure 5.7a). We identify another pattern in `mash commonbaseline`, Figure 5.7b, which is missed in Blischak et al. (2015). The pattern shows strong differential expression with `Smeg` infection, and modest positive correlation between deviations from `Smeg` and `Salm` infections.

We assess the quantitative similarity of deviations by sign and magnitude. Urbut et al. (2019) defined two effects to be shared in magnitude if they have the same sign and an effect within a factor of 2 of one another. Figure 5.8 shows the proportion of differentially expressed genes with similar magnitude and sign in each pair of bacteria infections. Sharing by magnitude is necessarily lower because it implies sharing by

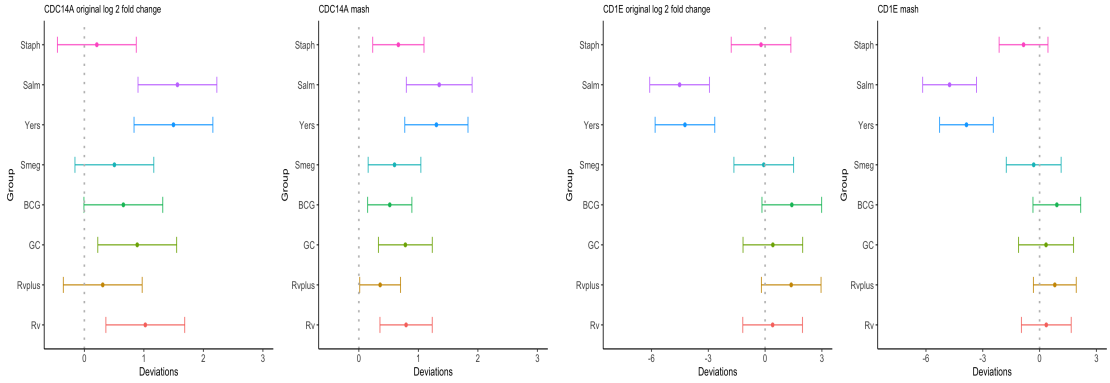
sign. Almost all genes share the deviation direction among infections. We notice that `Yers` and `Salm` share the magnitude very closely. Moreover, `BCG` tends to share effects with `GC`, `Rv` and `Rvplus`, which is biologically reasonable given they are all mycobacteria. These sharing patterns agree with the primary patterns from the `mash commonbaseline` model and `Cormotif` results from Blischak et al. (2015).

Comparing with `Cormotif` analysis, which restricting genes to a binary pattern, our method describes a pattern of continuous effects. A common binary configuration identified by `Cormotif` is `Yers-Salm`, which contains roughly 13% genes. This configuration includes genes with strong differential expression over controls in only these two infected conditions. In `mash commonbaseline` model, the identified pattern is more subtle: it has strong deviations in `Yers` and `Salm`, and weaker but positively correlated deviations in other infected conditions. In Figure 5.9a, we show one example that `Cormotif` classified the gene (CDC14A) to `Yers-Salm` configuration, while the raw data have some weak but positively correlated deviations in other infections. Our method learns patterns from data and borrows information across conditions in the posterior calculation, so it recognizes that the gene is differentially expressed in all conditions with strong deviations in `Yers` and `Salm` infections. In another example (Figure 5.9b), `mash commonbaseline` preserves the `Yers-Salm` conditions specific pattern and shrinks other deviations to zero.

The other advantage of `mash commonbaseline` is it learns patterns of effects from data and it identifies the pattern with strong `Smeg` effect. Therefore, we identify genes that are strongly differentially expressed with `Smeg` infection and weaker correlated deviation in `Salm`. Figure 5.9c shows one example. `Cormotif` classified this gene as

(a) Sharing in all conditions (Yers-Salm cluster in `Cormotif`)

(b) Conditions specific (Yers-Salm cluster in `Cormotif`)



(c) `Smeg` effect (non-DE cluster in `Cormotif`)

(d) No deviations (All cluster in `Cormotif`)

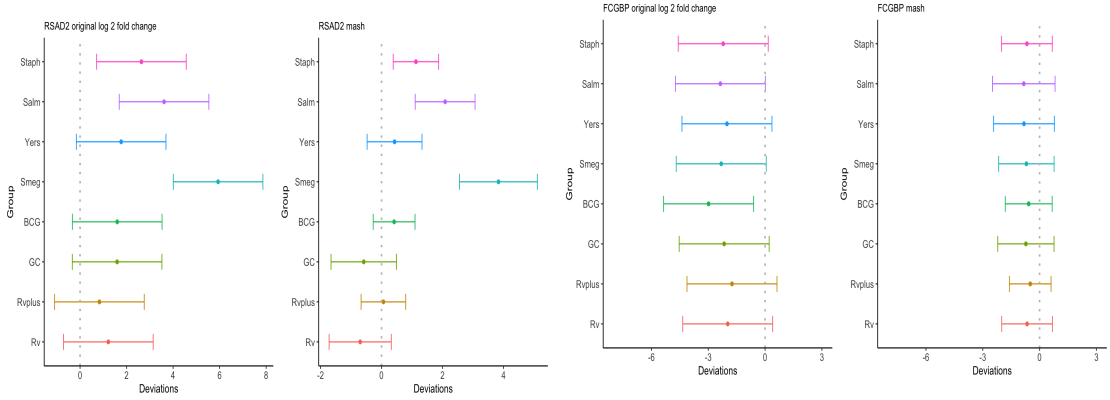


Figure 5.9: **`mash commonbaseline` examples.** `mash commonbaseline` uses learned patterns of sharing to capture more subtle patterns. For each subfigure, the dots in the left plot are raw deviations for each infection with bars indicating ± 2 se. The dots in the right plot are `mash commonbaseline` posterior estimated deviations with bars indicating ± 2 se.

non-DE, because it fails to recognize the pattern with strong deviation in `Smeg`.

Lastly, our method takes into account the correlation induced by comparing all expressions to the same control group. In contrast, `Cormotif` ignores the induced correlation in the error structure, which could yield possible false discoveries. Figure 5.9d shows one example in which `Cormotif` assigned the gene to all differentially expressed pattern while the raw deviations are small and correlated. In `mash commonbaseline`, we include the induced correlation in the model, so the corresponding `mash commonbaseline` estimates shrink all deviations to zero.

5.3.7 Discussion

The `mash commonbaseline` model jointly analyzes differential expressions in multiple conditions. It takes into account the correlations induced by comparing all conditions to a common baseline, which is commonly ignored in other methods. The false positives reduce by including the induced correlations. It provides quantitative estimation and assessment of deviations, rather than binary configurations for pattern of sharing.

We extend the `mash commonbaseline` model from Urbut (2017) to compare the quantity with the mean or median. Comparing with median could provide a more parsimonious conclusion. We did not find an easy method to compute the variance of median in literature. So we describe two ways to estimate the deviations over median, 1. subtract median from all conditions and use `mash`, 2. estimate from posteriors of deviations over mean. These two methods perform similarly. We recommend the first one, because it is simpler and faster than the sampling method.

References

Albert, A. (1972). *Regression and the Moore-Penrose pseudoinverse*. Elsevier.

Allen, H. L., Estrada, K., Lettre, G., Berndt, S. I., Weedon, M. N., Rivadeneira, F., Willer, C. J., Jackson, A. U., Vedantam, S., Raychaudhuri, S., et al. (2010). Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*, 467(7317):832–838.

Astle, W. J., Elding, H., Jiang, T., Allen, D., Ruklisa, D., Mann, A. L., Mead, D., Bouman, H., Riveros-Mckay, F., Kostadima, M. A., et al. (2016). The allelic landscape of human blood cell trait variation and links to common complex disease. *Cell*, 167(5):1415–1429.

Benner, C., Havulinna, A. S., Järvelin, M.-R., Salomaa, V., Ripatti, S., and Pirinen, M. (2017). Prospects of fine-mapping trait-associated genomic regions by using summary statistics from genome-wide association studies. *The American Journal of Human Genetics*, 101(4):539–551.

Benner, C., Spencer, C. C., Havulinna, A. S., Salomaa, V., Ripatti, S., and Pirinen, M. (2016). FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics*, 32(10):1493–1501.

Berk, R., Brown, L., Buja, A., Zhang, K., Zhao, L., et al. (2013). Valid post-selection inference. *The Annals of Statistics*, 41(2):802–837.

Blischak, J. D., Tailleux, L., Mitrano, A., Barreiro, L. B., and Gilad, Y. (2015). Mycobacterial infection induces a specific human innate immune response. *Scientific reports*, 5:16882.

Bottolo, L., Petretto, E., Blankenberg, S., Cambien, F., Cook, S. A., Tiret, L., and Richardson, S. (2011). Bayesian detection of expression quantitative trait loci hot spots. *Genetics*, 189(4):1449–1459.

Bottolo, L., Richardson, S., et al. (2010). Evolutionary stochastic search for bayesian model exploration. *Bayesian Analysis*, 5(3):583–618.

Bovy, J., Hogg, D. W., Roweis, S. T., et al. (2011). Extreme deconvolution: Inferring complete distribution functions from noisy, heterogeneous and incomplete observations. *The Annals of Applied Statistics*, 5(2B):1657–1677.

Brent, R. P. (2002). *Algorithms for minimization without derivatives*. Dover, Mineola, NY.

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., et al. (2018). The uk biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209.

Canela-Xandri, O., Rawlik, K., and Tenesa, A. (2018). An atlas of genetic associations in uk biobank. *Nature genetics*, 50(11):1593–1599.

Carbonetto, P., Stephens, M., et al. (2012). Scalable variational inference for bayesian variable selection in regression, and its accuracy in genetic association studies. *Bayesian analysis*, 7(1):73–108.

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., and Lee, J. J. (2015). Second-generation plink: rising to the challenge of larger and richer datasets. *Gigascience*, 4(1):s13742–015.

Chen, H., Wang, C., Conomos, M. P., Stilp, A. M., Li, Z., Sofer, T., Szpiro, A. A., Chen, W., Brehm, J. M., Celedón, J. C., et al. (2016). Control for population structure and relatedness for binary traits in genetic association studies via logistic mixed models. *The American Journal of Human Genetics*, 98(4):653–666.

Chen, W., Larrabee, B. R., Ovsyannikova, I. G., Kennedy, R. B., Haralambieva, I. H., Poland, G. A., and Schaid, D. J. (2015). Fine mapping causal variants with an approximate bayesian method using marginal test statistics. *Genetics*, 200(3):719–736.

Chen, W., Wu, Y., Zheng, Z., Qi, T., Visscher, P. M., Zhu, Z., and Yang, J. (2020). Improved analyses of gwas summary statistics by reducing data heterogeneity and errors. *bioRxiv*.

Consortium, . G. P. et al. (2015). A global reference for human genetic variation. *Nature*, 526(7571):68.

Consortium, M. S. et al. (1999). Complete sequence and gene map of a human major histocompatibility complex. *Nature*, 401(6756):921–923.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22.

Donnelly, P. (2008). Progress and challenges in genome-wide association studies in humans. *Nature*, 456(7223):728–731.

Dorgalaleh, A., Mahmoodi, M., Varmaghani, B., et al. (2013). Effect of thyroid dysfunctions on blood cell count and red blood cell indice. *Iranian journal of pediatric hematology and oncology*, 3(2):73.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, 80(1):27–38.

Flister, M. J., Tsaih, S.-W., O'Meara, C. C., Endres, B., Hoffman, M. J., Geurts, A. M., Dwinell, M. R., Lazar, J., Jacob, H. J., and Moreno, C. (2013). Identifying multiple causative genes at a single gwas locus. *Genome research*, 23(12):1996–2002.

Flutre, T., Wen, X., Pritchard, J., and Stephens, M. (2013). A statistical framework for joint eqtl analysis in multiple tissues. *PLoS Genet*, 9(5):e1003486.

Friedman, J., Hastie, T., Tibshirani, R., et al. (2001). *The elements of statistical learning*, volume 1. Springer series in statistics New York.

Fritsche, L. G., Igl, W., Bailey, J. N. C., Grassmann, F., Sengupta, S., Bragg-Gresham, J. L., Burdon, K. P., Hebbring, S. J., Wen, C., Gorski, M., et al. (2016). A large genome-wide association study of age-related macular degeneration highlights contributions of rare and common variants. *Nature genetics*, 48(2):134–143.

George, E. I. and McCulloch, R. E. (1997). Approaches for bayesian variable selection. *Statistica sinica*, pages 339–373.

Ghosh, A., Garee, G., Sweeny, E. A., Nakamura, Y., and Stuehr, D. J. (2018). Hsp90 chaperones hemoglobin maturation in erythroid and nonerythroid cells. *Proceedings of the National Academy of Sciences*, 115(6):E1117–E1126.

Giambartolomei, C., Vukcevic, D., Schadt, E. E., Franke, L., Hingorani, A. D., Wallace, C., and Plagnol, V. (2014). Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet*, 10(5):e1004383.

Gieger, C., Radhakrishnan, A., Cvejic, A., Tang, W., Porcu, E., Pistis, G., Serbanovic-Canic, J., Elling, U., Goodall, A. H., Labrune, Y., et al. (2011). New gene functions in megakaryopoiesis and platelet formation. *Nature*, 480(7376):201–208.

Guan, Y. and Stephens, M. (2011). Bayesian variable selection regression for genome-wide association studies and other large-scale problems. *The Annals of Applied Statistics*, pages 1780–1815.

Guo, M. H., Nandakumar, S. K., Ulirsch, J. C., Zekavat, S. M., Buenrostro, J. D., Natarajan, P., Salem, R. M., Chiarle, R., Mitt, M., Kals, M., et al. (2017). Comprehensive population-based genome sequencing provides insight into hematopoietic regulatory mechanisms. *Proceedings of the National Academy of Sciences*, 114(3):E327–E336.

Hakonarson, H., Grant, S. F., Bradfield, J. P., Marchand, L., Kim, C. E., Glessner, J. T., Grabs, R., Casalunovo, T., Taback, S. P., Frackelton, E. C., et al. (2007). A genome-wide association study identifies kiaa0350 as a type 1 diabetes gene. *Nature*, 448(7153):591–594.

Han, B. and Eskin, E. (2011). Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. *The American Journal of Human Genetics*, 88(5):586–598.

Han, B., Kang, H. M., and Eskin, E. (2009). Rapid and accurate multiple testing correction and power estimation for millions of correlated markers. *PLoS Genet*, 5(4):e1000456.

Heinze, G. and Schemper, M. (2002). A solution to the problem of separation in logistic regression. *Statistics in medicine*, 21(16):2409–2419.

Hoggart, C. J., Whittaker, J. C., De Iorio, M., and Balding, D. J. (2008). Simultaneous analysis of all snps in genome-wide and re-sequencing association studies. *PLoS Genet*, 4(7):e1000130.

Hormozdiari, F., Kostem, E., Kang, E. Y., Pasaniuc, B., and Eskin, E. (2014). Identifying causal variants at loci with multiple signals of association. *Genetics*, 198(2):497–508.

Hormozdiari, F., Van De Bunt, M., Segre, A. V., Li, X., Joo, J. W. J., Bilow, M., Sul, J. H., Sankararaman, S., Pasaniuc, B., and Eskin, E. (2016). Colocalization of gwas and eqtl signals detects target genes. *The American Journal of Human Genetics*, 99(6):1245–1260.

Horton, R., Wilming, L., Rand, V., Lovering, R. C., Bruford, E. A., Khodiyar, V. K., Lush, M. J., Povey, S., Talbot, C. C., Wright, M. W., et al. (2004). Gene map of the extended human mhc. *Nature Reviews Genetics*, 5(12):889–899.

Kanai, M., Akiyama, M., Takahashi, A., Matoba, N., Momozawa, Y., Ikeda, M., Iwata, N., Ikegawa, S., Hirata, M., Matsuda, K., et al. (2018). Genetic analysis of quantitative traits in the japanese population links cell types to complex human diseases. *Nature genetics*, 50(3):390–400.

Katsel, P., Davis, K., Gorman, J., and Haroutunian, V. (2005). Variations in differential gene expression patterns across multiple brain regions in schizophrenia. *Schizophrenia research*, 77(2-3):241–252.

Kichaev, G. and Pasaniuc, B. (2015). Leveraging functional-annotation data in trans-ethnic fine-mapping studies. *The American Journal of Human Genetics*, 97(2):260–271.

Kichaev, G., Roytman, M., Johnson, R., Eskin, E., Lindstroem, S., Kraft, P., and Pasaniuc, B. (2017). Improved methods for multi-trait fine mapping of pleiotropic risk loci. *Bioinformatics*, 33(2):248–255.

Kichaev, G., Yang, W.-Y., Lindstrom, S., Hormozdiari, F., Eskin, E., Price, A. L., Kraft, P., and Pasaniuc, B. (2014). Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS Genet*, 10(10):e1004722.

Kim, Y., Carbonetto, P., Stephens, M., and Anitescu, M. (2020). A fast algorithm for maximum likelihood estimation of mixture proportions using sequential quadratic programming. *Journal of Computational and Graphical Statistics*, 29(2):261–273.

Köttgen, A., Albrecht, E., Teumer, A., Vitart, V., Krumsiek, J., Hundertmark, C., Pistis, G., Ruggiero, D., O'Seaghdha, C. M., Haller, T., et al. (2013). Genome-wide association analyses identify 18 new loci associated with serum urate concentrations. *Nature genetics*, 45(2):145–154.

Kuonen, D. (1999). Miscellanea. saddlepoint approximations for distributions of quadratic forms in normal variables. *Biometrika*, 86(4):929–935.

LaPierre, N., Taraszka, K., Huang, H., He, R., Hormozdiari, F., and Eskin, E. (2020). Identifying causal variants by fine mapping across multiple studies. In *International Conference on Research in Computational Molecular Biology*, pages 257–258. Springer.

Ledoit, O. and Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of multivariate analysis*, 88(2):365–411.

Lee, C., Eskin, E., and Han, B. (2017). Increasing the power of meta-analysis of genome-wide association studies to detect heterogeneous effects. *Bioinformatics*, 33(14):i379–i388.

Lee, D., Bigdeli, T. B., Riley, B. P., Fanous, A. H., and Bacanu, S.-A. (2013). Dist: direct imputation of summary statistics for unmeasured snps. *Bioinformatics*, 29(22):2925–2927.

Lee, J. D., Sun, D. L., Sun, Y., Taylor, J. E., et al. (2016). Exact post-selection inference, with application to the lasso. *Annals of Statistics*, 44(3):907–927.

Lee, Y., Francesca, L., Pique-Regi, R., and Wen, X. (2018). Bayesian multi-snp genetic association analysis: Control of fdr and use of summary statistics. *bioRxiv*, page 316471.

Lewin, A., Saadi, H., Peters, J. E., Moreno-Moral, A., Lee, J. C., Smith, K. G., Petretto, E., Bottolo, L., and Richardson, S. (2016). Mt-hess: an efficient bayesian approach for simultaneous association detection in omics datasets, with application to eqtl mapping in multiple tissues. *Bioinformatics*, 32(4):523–532.

Lloyd-Jones, L. R., Zeng, J., Sidorenko, J., Yengo, L., Moser, G., Kemper, K. E., Wang, H., Zheng, Z., Magi, R., Esko, T., et al. (2019). Improved polygenic prediction by bayesian multiple regression on summary statistics. *Nature communications*, 10(1):1–11.

Logsdon, B. A., Hoffman, G. E., and Mezey, J. G. (2010). A variational bayes algorithm for fast and accurate multiple locus genome-wide association analysis. *BMC bioinformatics*, 11(1):1–13.

Lozano, J. A., Hormozdiari, F., Joo, J. W. J., Han, B., and Eskin, E. (2017). The multivariate normal distribution framework for analyzing association studies. *bioRxiv*, page 208199.

Mak, T. S. H., Porsch, R. M., Choi, S. W., Zhou, X., and Sham, P. C. (2017). Polygenic scores via penalized regression on summary statistics. *Genetic epidemiology*, 41(6):469–480.

Maller, J. B., McVean, G., Byrnes, J., Vukcevic, D., Palin, K., Su, Z., Howson, J. M., Auton, A., Myers, S., Morris, A., et al. (2012). Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nature genetics*, 44(12):1294.

Mbatchou, J., Barnard, L., Backman, J., Marcketta, A., Kosmicki, J. A., Ziyatdinov, A., Benner, C., O'Dushlaine, C., Barber, M., Boutkov, B., et al. (2021). Computationally efficient whole-genome regression for quantitative and binary traits. Technical report, Nature Publishing Group.

McCarthy, D. J., Chen, Y., and Smyth, G. K. (2012). Differential expression analysis of multifactor rna-seq experiments with respect to biological variation. *Nucleic acids research*, 40(10):4288–4297.

McCullagh, P. and Nelder, J. A. (2019). *Generalized linear models*. Routledge.

Meuwissen, T. H., Hayes, B. J., and Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, 157(4):1819–1829.

Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the american statistical association*, 83(404):1023–1032.

Moser, G., Lee, S. H., Hayes, B. J., Goddard, M. E., Wray, N. R., and Visscher, P. M. (2015). Simultaneous discovery, estimation and prediction analysis of complex traits using a bayesian mixture model. *PLoS Genet*, 11(4):e1004969.

Newcombe, P. J., Conti, D. V., and Richardson, S. (2016). Jam: a scalable bayesian framework for joint analysis of marginal snp effects. *Genetic epidemiology*, 40(3):188–201.

Ott, J. (1999). *Analysis of human genetic linkage*. JHU Press.

Park, Y., Sarkar, A. K., He, L., Davila-Velderrain, J., De Jager, P. L., and Kellis, M. (2017). A bayesian approach to mediation analysis predicts 206 causal target genes in alzheimer's disease. *bioRxiv*, page 219428.

Pasaniuc, B. and Price, A. L. (2017). Dissecting the genetics of complex traits using summary association statistics. *Nature Reviews Genetics*, 18(2):117.

Pasaniuc, B., Zaitlen, N., Shi, H., Bhatia, G., Gusev, A., Pickrell, J., Hirschhorn, J., Strachan, D. P., Patterson, N., and Price, A. L. (2014). Fast and accurate imputation of summary statistics enhances evidence of functional enrichment. *Bioinformatics*, 30(20):2906–2914.

Pickrell, J. K. (2014). Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *The American Journal of Human Genetics*, 94(4):559–573.

Pinheiro, J. C. and Bates, D. M. (1996). Unconstrained parametrizations for variance-covariance matrices. *Statistics and computing*, 6(3):289–296.

Purcell, S. M. and Chang, C. C. (2019). Plink2.0.

Ruffieux, H., Davison, A. C., Hager, J., Inshaw, J., Fairfax, B. P., Richardson, S., Bottolo, L., et al. (2020). A global-local approach for detecting hotspots in multiple-response regression. *Annals of Applied Statistics*, 14(2):905–928.

Rurale, G., Persani, L., and Marelli, F. (2018). Glis3 and thyroid: a pleiotropic candidate gene for congenital hypothyroidism. *Frontiers in endocrinology*, 9:730.

Schaid, D. J., Chen, W., and Larson, N. B. (2018). From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nature Reviews Genetics*, 19(8):491–504.

Scholz, G. M., Cartledge, K., and Hall, N. E. (2001). Identification and characterization of harc, a novel hsp90-associating relative of cdc37. *Journal of Biological Chemistry*, 276(33):30971–30979.

Servin, B. and Stephens, M. (2007). Imputation-based analysis of association studies: candidate regions and quantitative traits. *PLoS genetics*, 3(7):e114.

Shriner, D. (2012). Moving toward system genetics through multiple trait analysis in genome-wide association studies. *Frontiers in genetics*, 3:1.

Sillanpaa, M. J. and Bhattacharjee, M. (2005). Bayesian association-based fine mapping in small chromosomal segments. *Genetics*, 169(1):427–439.

Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology*, 3(1).

Spain, S. L. and Barrett, J. C. (2015). Strategies for fine-mapping complex traits. *Human molecular genetics*, 24(R1):R111–R119.

Stephens, M. (2013). A unified framework for association analysis with multiple related phenotypes. *PloS one*, 8(7):e65245.

Stephens, M. (2017). False discovery rates: a new deal. *Biostatistics*, 18(2):275–294.

Stephens, M. and Balding, D. J. (2009). Bayesian statistical methods for genetic association studies. *Nature Reviews Genetics*, 10(10):681–690.

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., et al. (2015). Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *Plos med*, 12(3):e1001779.

Tang, M., Sun, J., Shimizu, K., and Kadota, K. (2015). Evaluation of methods for differential expression analysis on multi-group rna-seq count data. *BMC bioinformatics*, 16(1):1–14.

Taylor, J. and Tibshirani, R. J. (2015). Statistical learning and selective inference. *Proceedings of the National Academy of Sciences*, 112(25):7629–7634.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(3):273–282.

Trynka, G., Hunt, K. A., Bockett, N. A., Romanos, J., Mistry, V., Szperl, A., Bakker, S. F., Bardella, M. T., Bhaw-Rosun, L., Castillejo, G., et al. (2011). Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. *Nature genetics*, 43(12):1193.

Ulirsch, J. C., Lareau, C. A., Bao, E. L., Ludwig, L. S., Guo, M. H., Benner, C., Satpathy, A. T., Kartha, V. K., Salem, R. M., Hirschhorn, J. N., et al. (2019). Interrogation of human hematopoiesis at single-cell and single-variant resolution. *Nature genetics*, 51(4):683–693.

Urbut, S. M. (2017). *Flexible Statistical Methods for Jointly Modeling Effects*. PhD thesis, The University of Chicago.

Urbut, S. M., Wang, G., Carbonetto, P., and Stephens, M. (2019). Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. *Nature genetics*, 51(1):187–195.

Veyrieras, J.-B., Kudaravalli, S., Kim, S. Y., Dermitzakis, E. T., Gilad, Y., Stephens, M., and Pritchard, J. K. (2008). High-resolution mapping of expression-qtls yields insight into human gene regulation. *PLoS Genet*, 4(10):e1000214.

Vilhjálmsson, B. J., Yang, J., Finucane, H. K., Gusev, A., Lindström, S., Ripke, S., Genovese, G., Loh, P.-R., Bhatia, G., Do, R., et al. (2015). Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *The american journal of human genetics*, 97(4):576–592.

Visscher, P. M., Brown, M. A., McCarthy, M. I., and Yang, J. (2012). Five years of gwas discovery. *The American Journal of Human Genetics*, 90(1):7–24.

Vuckovic, D., Bao, E. L., Akbari, P., Lareau, C. A., Mousas, A., Jiang, T., Chen, M.-H., Raffield, L. M., Tardaguila, M., Huffman, J. E., et al. (2020). The polygenic and monogenic basis of blood traits and diseases. *Cell*, 182(5):1214–1231.

Wakefield, J. (2009). Bayes factors for genome-wide association studies: comparison with p-values. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, 33(1):79–86.

Wallace, C. (2020). Eliciting priors and relaxing the single causal variant assumption in colocalisation analyses. *PLoS genetics*, 16(4):e1008720.

Wallace, C. (2021). A more accurate method for colocalisation analysis allowing for multiple causal variants. *bioRxiv*.

Wallace, C., Cutler, A. J., Pontikos, N., Pekalski, M. L., Burren, O. S., Cooper, J. D., García, A. R., Ferreira, R. C., Guo, H., Walker, N. M., et al. (2015). Dissection of a complex disease susceptibility region using a bayesian stochastic search approach to fine mapping. *PLoS Genet*, 11(6):e1005272.

Wang, G., Sarkar, A., Carbonetto, P., and Stephens, M. (2020). A simple new approach to variable selection in regression, with application to genetic fine mapping. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(5):1273–1300.

Wang, W. and Stephens, M. (2018). Empirical bayes matrix factorization. *arXiv preprint arXiv:1802.06931*.

Wei, Y., Tenzen, T., and Ji, H. (2015). Joint analysis of differential gene expression in multiple studies using correlation motifs. *Biostatistics*, 16(1):31–46.

Weissbrod, O., Hormozdiari, F., Benner, C., Cui, R., Ulirsch, J., Gazal, S., Schoech, A. P., Van De Geijn, B., Reshef, Y., Márquez-Luna, C., et al. (2020). Functionally informed fine-mapping and polygenic localization of complex trait heritability. *Nature Genetics*, 52(12):1355–1363.

Wen, X. et al. (2016a). Molecular qtl discovery incorporating genomic annotations using bayesian false discovery rate control. *Annals of Applied Statistics*, 10(3):1619–1638.

Wen, X., Lee, Y., Luca, F., and Pique-Regi, R. (2016b). Efficient integrative multi-snp association analysis via deterministic approximation of posteriors. *The American Journal of Human Genetics*, 98(6):1114–1129.

Wen, X., Pique-Regi, R., and Luca, F. (2017). Integrating molecular qtl data into genome-wide genetic association analysis: Probabilistic assessment of enrichment and colocalization. *PLoS genetics*, 13(3):e1006646.

Wen, X. and Stephens, M. (2014). Bayesian methods for genetic association analysis with heterogeneous subgroups: from meta-analyses to gene-environment interactions. *The annals of applied statistics*, 8(1):176.

Willer, C. J., Li, Y., and Abecasis, G. R. (2010). Metal: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics*, 26(17):2190–2191.

Yang, J., Ferreira, T., Morris, A. P., Medland, S. E., Madden, P. A., Heath, A. C., Martin, N. G., Montgomery, G. W., Weedon, M. N., Loos, R. J., et al. (2012). Conditional and joint multiple-snp analysis of gwas summary statistics identifies additional variants influencing complex traits. *Nature genetics*, 44(4):369–375.

Zeggini, E., Weedon, M. N., Lindgren, C. M., Frayling, T. M., Elliott, K. S., Lango, H., Timpson, N. J., Perry, J. R., Rayner, N. W., Freathy, R. M., et al. (2007). Replication of genome-wide association signals in uk samples reveals risk loci for type 2 diabetes. *Science*, 316(5829):1336–1341.

Zhou, W., Nielsen, J. B., Fritsche, L. G., Dey, R., Gabrielsen, M. E., Wolford, B. N., LeFaive, J., VandeHaar, P., Gagliano, S. A., Gifford, A., et al. (2018). Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nature genetics*, 50(9):1335–1341.

Zhou, X., Carbonetto, P., and Stephens, M. (2013). Polygenic modeling with bayesian sparse linear mixed models. *PLoS Genet*, 9(2):e1003264.

Zhu, X. and Stephens, M. (2017). Bayesian large-scale multiple regression with summary statistics from genome-wide association studies. *The annals of applied statistics*, 11(3):1561.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320.

APPENDIX A

SUPPLEMENTARY FOR *SUSIE-SUFF*

A.1 *SuSiE* using sufficient statistics

We can compute the sufficient statistics either directly from column centered individual-level $\{\boldsymbol{y}, \boldsymbol{X}\}$, or derive them using the summary statistics from J simple linear regressions ($\hat{b}_j = \boldsymbol{x}_j^\intercal \boldsymbol{y} / \boldsymbol{x}_j^\intercal \boldsymbol{x}_j$ with standard error, $\hat{s}_j$), sample correlation matrix $\hat{\boldsymbol{R}}_s$ estimated from $\boldsymbol{X}$, variance of $\boldsymbol{y}$ ($\hat{\sigma^2}_y = \frac{\boldsymbol{y}^\intercal \boldsymbol{y}}{N-1}$) and sample size N. Based on the z score for each single SNP, $\hat{z}_j = \hat{b}_j / \hat{s}_j$, the correlation coefficient R_j^2 for the corresponding simple linear regression model is

$$R_j^2 = \frac{\hat{z}_j^2}{\hat{z}_j^2 + N - 2}. \tag{A.1.1}$$

Using correlation coefficient R_j^2, the estimated residual variance from the simple linear regression model is

$$\hat{\sigma^2}_j = \hat{\sigma^2}_y \frac{(N-1)(1-R_j^2)}{N-2} \tag{A.1.2}$$

$$= \hat{\sigma^2}_y \frac{N-1}{\hat{z}_j^2 + N - 2}. \tag{A.1.3}$$

Therefore, we can compute the sufficient statistics as

$$\boldsymbol{X}^\intercal \boldsymbol{X} = \boldsymbol{D}\hat{\boldsymbol{R}}_s\boldsymbol{D}, \tag{A.1.4}$$

$$\boldsymbol{x}_j^\intercal \boldsymbol{y} = \frac{\hat{\sigma^2}_j}{\hat{s}_j}\hat{z}_j, \tag{A.1.5}$$

$$\boldsymbol{y}^\intercal \boldsymbol{y} = \hat{\sigma^2}_y(N-1), \tag{A.1.6}$$

in which $\boldsymbol{D} \coloneqq \text{diag}(\sqrt{\boldsymbol{x}_1^\intercal \boldsymbol{x}_1}, \cdots, \sqrt{\boldsymbol{x}_p^\intercal \boldsymbol{x}_p})$ and $\boldsymbol{x}_j^\intercal \boldsymbol{x}_j = \frac{\hat{\sigma^2}_j}{\hat{s}_j^2}$.

APPENDIX B

SUPPLEMENTARY FOR *SUSIE-RSS*

B.1 Modifying LD matrix with z scores

Without the original individual genotype data, the only information we have about the original samples is the observed marginal z scores. The observed z scores contain some correlation information from the original data. In the case there are two SNPs with the exactly same z scores, the two SNPs must be perfectly correlated in the original genotype data. Therefore, we modify the LD matrix using information in the observed z scores.

We treat the (centered and scaled) genotypes of each individual from the reference panel, $\boldsymbol{x}_i^{\text{ref}}$ (the ith row of $\boldsymbol{X}_{\text{ref}}$), as being independent and identically distributed draws from the "suitable" population with LD matrix $\boldsymbol{R}$, that is

$$\boldsymbol{x}_i^{\text{ref}} \sim N(\boldsymbol{0}, \boldsymbol{R}), \quad i = 1, \cdots n_{\text{ref}}. \tag{B.1.1}$$

Under the null ($\boldsymbol{z} = \boldsymbol{0}$), the model for single-SNP association statistics with population LD matrix $\boldsymbol{R}$ is

$$\hat{\boldsymbol{z}} \sim N(\boldsymbol{0}, \boldsymbol{R}). \tag{B.1.2}$$

Since both (B.1.1) and (B.1.2) have the same correlation matrix $\boldsymbol{R}$, we estimate $\boldsymbol{R}$ using sample correlation with pooled data. We treat the observed z scores, $\hat{\boldsymbol{z}}$, as one

additional observation to the reference panel,

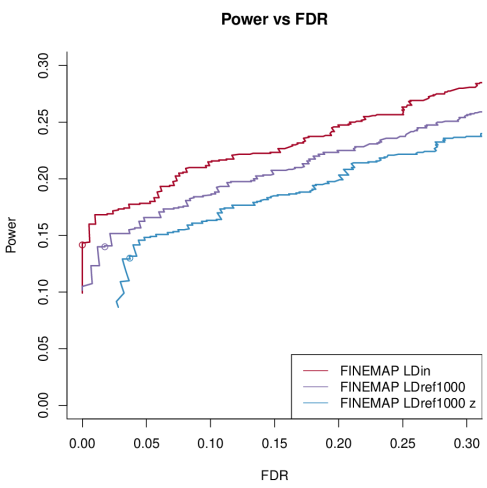
$$\hat{\boldsymbol{R}}' = \frac{1}{n_{\text{ref}}+1}\left(\boldsymbol{X}_{\text{ref}}^{\intercal}\boldsymbol{X}_{\text{ref}} + \hat{\boldsymbol{z}}\hat{\boldsymbol{z}}^{\intercal}\right) \tag{B.1.3}$$

$$= \frac{1}{n_{\text{ref}}+1}\left[\left(\sum_{i=1}^{n_{\text{ref}}} \boldsymbol{x}_i^{\text{ref}}\boldsymbol{x}_i^{\text{ref}\intercal}\right) + \hat{\boldsymbol{z}}\hat{\boldsymbol{z}}^{\intercal}\right]. \tag{B.1.4}$$

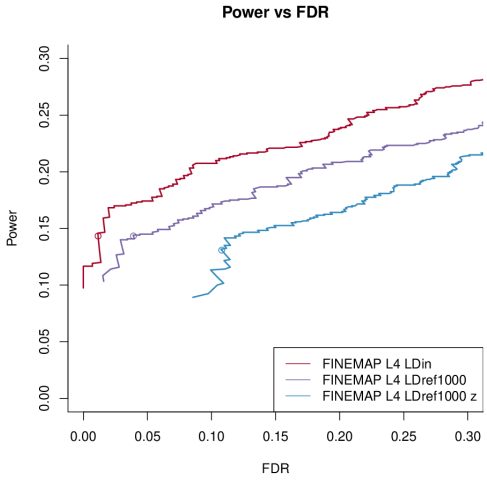
We convert $\hat{\boldsymbol{R}}'$ to correlation matrix.

In our simulation, we observe this modification makes things worse (Figure B.1). The possible explanation is the observed z scores are not under the null. The strong signals in the observed z scores destroy the LD estimates.

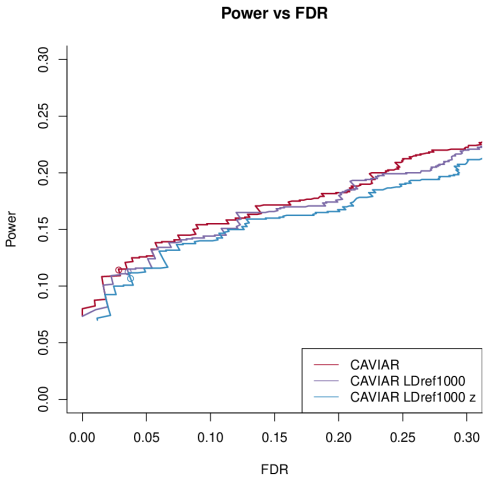
(a) FINEMAP with oracle maximum number of effects



(b) FINEMAP with 4 maximum effects



(c) CAVIAR with oracle maximum number of effects



(d) *SuSiE-RSS* with 10 maximum effects

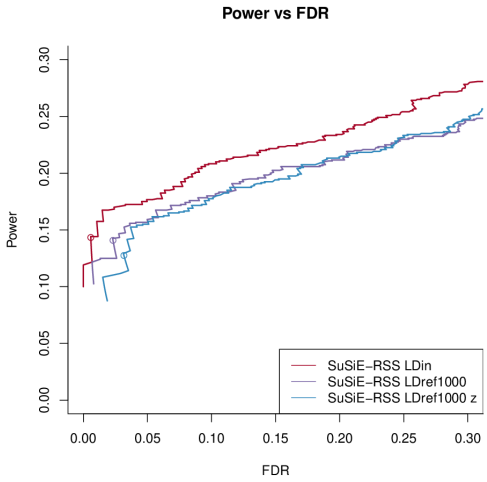


Figure B.1: **Evaluation of posterior inclusion probabilities (PIPs) using LD from reference panel with correction from z scores.**

APPENDIX C

SUPPLEMENTARY FOR *MVSUSIE-RSS*

C.1 Details of posterior computations for the *BMR* model with a mixture prior

The basic result for the *BMR* model is given in the main text (Proposition 4). We consider the *BMR* model with an intercept and missing values in the following subsections.

Under the *BMR* model, the likelihood for $\boldsymbol{b}$ is

$$L(\boldsymbol{b}; \boldsymbol{x}, \boldsymbol{Y}, \boldsymbol{V}) \coloneqq p(\boldsymbol{Y}|\boldsymbol{x}, \boldsymbol{b}, \boldsymbol{V}) \tag{C.1.1}$$

$$= |2\pi\boldsymbol{V}|^{-N/2} \exp\big\{ -\tfrac{1}{2}\mathrm{tr}\Big[\boldsymbol{V}^{-1}(\boldsymbol{Y} - \boldsymbol{x}\boldsymbol{b}^\intercal)^\intercal(\boldsymbol{Y} - \boldsymbol{x}\boldsymbol{b}^\intercal)\Big]\big\} \tag{C.1.2}$$

$$= |2\pi\boldsymbol{V}|^{-N/2} \exp\big\{ -\tfrac{1}{2}[\mathrm{tr}(\boldsymbol{V}^{-1}\boldsymbol{Y}^\intercal\boldsymbol{Y}) + (\boldsymbol{b} - \hat{\boldsymbol{b}})^\intercal\boldsymbol{S}^{-1}(\boldsymbol{b} - \hat{\boldsymbol{b}}) - \hat{\boldsymbol{b}}^\intercal\boldsymbol{S}^{-1}\hat{\boldsymbol{b}}]\big\}. \tag{C.1.3}$$

The terms involving $\boldsymbol{b}$ are multivariate normal density up to a constant of proportionality, $L(\boldsymbol{b}; \boldsymbol{x}, \boldsymbol{Y}, \boldsymbol{V}) \propto N_R(\boldsymbol{b}; \hat{\boldsymbol{b}}, \boldsymbol{S})$.

C.1.1 Bayesian simple multivariate regression with an intercept

Here we extend the simple multivariate regression (4.1.10) to include an intercept. We show that including the intercept in the model is equivalent to "centering" $\boldsymbol{x}$ and the columns of $\boldsymbol{Y}$ so that they all have means of zero. This equivalence is achieved in two ways: from a point estimation perspective, centering $\boldsymbol{x}$ and the columns of

$\boldsymbol{Y}$ is equivalent to computing a maximum-likelihood estimate of the intercept; from a Bayesian perspective, centering $\boldsymbol{x}$ and columns of $\boldsymbol{Y}$ is equivalent to integrating out the intercept with respect to an improper, uniform prior. These results are summarized in a proposition.

The simple multivariate regression model with intercept is

$$\boldsymbol{Y} \sim \mathrm{MN}_{N\times R}(\mathbf{1}\boldsymbol{b}_0^\mathsf{T} + \boldsymbol{x}\boldsymbol{b}^\mathsf{T}, \boldsymbol{I}, \boldsymbol{V}), \tag{C.1.4}$$

in which $\mathbf{1} \coloneqq \mathbf{1}_N$ is a vector of ones of length N, and $\boldsymbol{b}_0 \in \mathbb{R}^R$ is the (unknown) intercept. The likelihood of $\boldsymbol{b}_0, \boldsymbol{b}$ is

$$\begin{aligned} &L(\boldsymbol{b}_0, \boldsymbol{b}; \boldsymbol{x}, \boldsymbol{Y}, \boldsymbol{V}) \coloneqq p(\boldsymbol{Y}|\boldsymbol{x}, \boldsymbol{b}_0, \boldsymbol{b}, \boldsymbol{V}) \\ &= (2\pi)^{-NR/2}|\boldsymbol{V}|^{-N/2} \exp\Big\{ -\tfrac{1}{2}\mathrm{tr}\left[\boldsymbol{V}^{-1}(\boldsymbol{Y} - \mathbf{1}\boldsymbol{b}_0^\mathsf{T} - \boldsymbol{x}\boldsymbol{b}^\mathsf{T})^\mathsf{T}(\boldsymbol{Y} - \mathbf{1}\boldsymbol{b}_0^\mathsf{T} - \boldsymbol{x}\boldsymbol{b}^\mathsf{T})\right]\Big\}. \end{aligned} \tag{C.1.5}$$

Proposition 6 (Simple multivariate regression with an intercept)**.** *Consider the simple multivariate regression* (C.1.4)*. The least-squares estimate of* $\boldsymbol{b}_0$*—that is, the* $\boldsymbol{b}_0$ *maximizing the likelihood* $L(\boldsymbol{b}_0, \boldsymbol{b}; \boldsymbol{x}, \boldsymbol{Y}, \boldsymbol{V})$*—and its variance-covariance matrix* $\boldsymbol{S}_0$*, are*

$$\hat{\boldsymbol{b}}_0 = \bar{\boldsymbol{y}} - \bar{x}\boldsymbol{b} \tag{C.1.6}$$

$$\boldsymbol{S}_0 = \tfrac{1}{N}\boldsymbol{V}, \tag{C.1.7}$$

in which $\bar{x} = \frac{1}{N}\boldsymbol{x}^\mathsf{T}\mathbf{1} = \frac{1}{N}\sum_{i=1}^N x_i$ *is the sample mean of* $\boldsymbol{x}$*, and* $\bar{\boldsymbol{y}} = \frac{1}{N}\boldsymbol{Y}^\mathsf{T}\mathbf{1}$ *is the*

vector containing the column means of $\boldsymbol{Y}$*.*

The profile likelihood for $\boldsymbol{b}$ *is*

$$\begin{aligned} L^*(\boldsymbol{b}; \boldsymbol{x}, \boldsymbol{Y}, \boldsymbol{V}) &\coloneqq \max_{\boldsymbol{b}_0} L(\boldsymbol{b}_0, \boldsymbol{b}; \boldsymbol{x}, \boldsymbol{Y}, \boldsymbol{V}) \\ &= L(\boldsymbol{b}; \tilde{\boldsymbol{x}}, \tilde{\boldsymbol{Y}}, \boldsymbol{V}), \end{aligned} \tag{C.1.8}$$

in which $\tilde{\boldsymbol{x}} \coloneqq \boldsymbol{x} - \bar{x}\mathbf{1}$ *and* $\tilde{\boldsymbol{Y}} \coloneqq \boldsymbol{Y} - \mathbf{1}\bar{\boldsymbol{y}}^\intercal$ *are the "centered"* $\boldsymbol{x}$ *and* $\boldsymbol{Y}$*. In other words, the profile likelihood for simple multivariate regression with an intercept is the same as the likelihood for multivariate regression* without *an intercept if we first center* $\boldsymbol{x}$ *and* $\boldsymbol{Y}$*. So centering* $\boldsymbol{x}$ *and* $\boldsymbol{Y}$ *is equivalent to including an intercept that is estimated by maximum-likelihood.*

Next, consider Bayesian calculations for $\boldsymbol{b}_0$ *with a multivariate normal prior,* $\boldsymbol{b}_0 \sim N_R(0, \boldsymbol{U}_0)$*, in which* $\boldsymbol{U}_0$ *is a positive semi-definite covariance matrix. The posterior for* $\boldsymbol{b}_0$ *given* $\boldsymbol{b}$ *is*

$$\boldsymbol{b}_0 | \boldsymbol{b}, \boldsymbol{x}, \boldsymbol{Y}, \boldsymbol{V}, \boldsymbol{U}, \boldsymbol{U}_0 \sim N_R(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_0), \tag{C.1.9}$$

where

$$\boldsymbol{\Sigma}_0 \coloneqq \boldsymbol{U}_0(\boldsymbol{I} + \boldsymbol{S}_0^{-1}\boldsymbol{U}_0)^{-1} \tag{C.1.10}$$

$$\boldsymbol{\mu}_0 \coloneqq \boldsymbol{\Sigma}_0 \boldsymbol{S}_0^{-1} \hat{\boldsymbol{b}}_0. \tag{C.1.11}$$

The marginal likelihood obtained by averaging over the intercept is

$$\begin{aligned} L^\star(\boldsymbol{b}; \boldsymbol{x}, \boldsymbol{Y}, \boldsymbol{V}, \boldsymbol{U}_0) &\coloneqq \textstyle\int L(\boldsymbol{b}_0, \boldsymbol{b}; \boldsymbol{x}, \boldsymbol{Y}, \boldsymbol{V}) p(\boldsymbol{b}_0 | \boldsymbol{U}_0) d\boldsymbol{b}_0 \\ &= (2\pi)^{-NR/2} |\boldsymbol{V}|^{-N/2} |\boldsymbol{\Sigma}_1|^{1/2} |\boldsymbol{U}_0|^{-1/2} \\ &\quad \times \exp\left\{ \tfrac{1}{2} \mu_0^\intercal \boldsymbol{\Sigma}_0^{-1} \mu_0 - \tfrac{1}{2} \mathrm{tr}\big[\boldsymbol{V}^{-1} (\boldsymbol{Y} - \boldsymbol{x}\boldsymbol{b}^\intercal)^\intercal (\boldsymbol{Y} - \boldsymbol{x}\boldsymbol{b}^\intercal) \big] \right\}. \end{aligned} \tag{C.1.12}$$

In the special case of an (improper) uniform prior on $\boldsymbol{b}_0$*, defined as* $\boldsymbol{b}_0 \sim N_R(0, \boldsymbol{U}_0)$ *with* $\boldsymbol{U}_0^{-1} \to 0$*, the posterior mean reduces to the least-squares estimate,* $\boldsymbol{\mu}_0 = \hat{\boldsymbol{b}}_0$*, the posterior covariance becomes* $\boldsymbol{\Sigma}_0 = \boldsymbol{S}_0$*, and the marginal likelihood simplifies to*

$$\begin{aligned} L^\star(\boldsymbol{b}; \boldsymbol{x}, \boldsymbol{Y}, \boldsymbol{V}, \boldsymbol{U}_0) &= (2\pi)^{-NR/2} |\boldsymbol{V}|^{-N/2} \frac{|\boldsymbol{S}_0|^{1/2}}{|\boldsymbol{U}_0|^{1/2}} \\ &\quad \times \exp\left\{ \tfrac{1}{2} \hat{\boldsymbol{b}}_0^\intercal \boldsymbol{S}_0^{-1} \hat{\boldsymbol{b}}_0 - \tfrac{1}{2} \mathrm{tr}\big[\boldsymbol{V}^{-1} (\boldsymbol{Y} - \boldsymbol{x}\boldsymbol{b}^\intercal)^\intercal (\boldsymbol{Y} - \boldsymbol{x}\boldsymbol{b}^\intercal) \big] \right\} \\ &= \frac{|\boldsymbol{S}_0|^{1/2}}{|\boldsymbol{U}_0|^{1/2}} \times L(\boldsymbol{b}; \tilde{\boldsymbol{x}}, \tilde{\boldsymbol{Y}}, \boldsymbol{V}). \end{aligned} \tag{C.1.13}$$

In other words, the marginal likelihood for multivariate regression with an intercept, when we use an improper uniform prior for the intercept, is the same—that is, up to a constant of proportionality—as the likelihood for multivariate regression without *an intercept when we first center* $\boldsymbol{x}$ *and* $\boldsymbol{Y}$*.*

C.1.2 Bayesian simple multivariate regression with missing data

Here we extend the Bayesian computations for the simple regression model (4.1.10) to allow for missing observations in the response, $\boldsymbol{Y}$. The basic properties derived for the full-data setting are mostly preserved in the missing-data setting; for example, the posterior distribution of $\boldsymbol{b}$ is mixture of multivariate normals when its prior is mixture of multivariate normals. The main complication introduced by missing data is that the least-squares estimate $\hat{\boldsymbol{b}}$ and its variance-covariance matrix $\boldsymbol{S}$ no longer have the simple form given in (4.1.12) and (4.1.13).

To formulate the model in the missing data setting, we define $\boldsymbol{\psi}_i = \{\psi_{i1}, \ldots, \psi_{iR}\}$ such that $\psi_{ir} = 1$ if condition r in sample i is observed, and $\psi_{ir} = 0$ if it is missing. We assume here that at least one condition is observed in each sample; that is, $|\boldsymbol{\psi}_i| \geqslant 1$ for all i. Next, defining $\Psi_i \coloneqq \text{diag}(\psi_{i1}, \ldots, \psi_{iR})$, $\boldsymbol{V}_i \coloneqq \Psi_i \boldsymbol{V} \Psi_i$ and $\boldsymbol{y}_i$ to be the ith row of $\boldsymbol{Y}$, the simple multivariate regression for sample i is written as

$$\Psi_i \boldsymbol{y}_i \sim N_R(x_i \Psi_i \boldsymbol{b}, \boldsymbol{V}_i). \tag{C.1.14}$$

By these definitions, $\Psi_i \boldsymbol{y}_i$ is $\boldsymbol{y}_i$ in which all missing values are replaced with zeros and, similarly, the $R \times R$ matrix $\boldsymbol{V}_i$ is obtained by filling all rows and columns of $\boldsymbol{V}$ with zeros when the rows and columns correspond to missing values in $\boldsymbol{y}_i$. Equation (C.1.14) defines a probability distribution in R dimensions and it has valid density in $1 \leqslant |\boldsymbol{\psi}_i| \leqslant R$ dimensions.

For the next expressions, we define the inverse of $\boldsymbol{V}_i$, denoted $\boldsymbol{V}_i^{-1}$, as the inverse of the submatrix taken from all rows and columns corresponding to observed

values, then reinserting this back to $\boldsymbol{V}_i$. Likewise, we define the determinant of $\boldsymbol{V}_i$, denoted $|\boldsymbol{V}_i|$, as the determinant of the submatrix taken from all rows and columns corresponding to observed values.

The likelihood for $\boldsymbol{b}$ is

$$\begin{aligned} L(\boldsymbol{b}; \boldsymbol{x}, \boldsymbol{Y}, \boldsymbol{V}, \boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_N) &\coloneqq \prod_{i=1}^{N} p(\boldsymbol{y}_i | \boldsymbol{x}, \boldsymbol{b}, \boldsymbol{V}, \boldsymbol{\psi}_i) \\ &= \prod_{i=1}^{N} (2\pi)^{-|\boldsymbol{\psi}_i|/2} |\boldsymbol{V}_i|^{-1/2} \times \exp\left\{-\tfrac{1}{2} \textstyle\sum_{i=1}^{N} (\boldsymbol{y}_i - x_i \boldsymbol{b})^\intercal \boldsymbol{V}_i^{-1} (\boldsymbol{y}_i - x_i \boldsymbol{b})\right\}. \end{aligned} \tag{C.1.15}$$

The least-squares estimate of $\boldsymbol{b}$, and its variance-covariance matrix $\boldsymbol{S}$, are

$$\hat{\boldsymbol{b}} = \boldsymbol{S} \textstyle\sum_{i=1}^{N} x_i \boldsymbol{V}_i^{-1} \boldsymbol{y}_i, \tag{C.1.16}$$

$$\boldsymbol{S} = \left(\textstyle\sum_{i=1}^{N} x_i^2 \boldsymbol{V}_i^{-1} \right)^{-1}. \tag{C.1.17}$$

Using these quantities, the likelihood (C.1.15) can be rewritten as

$$\begin{aligned} &L(\boldsymbol{b}; \boldsymbol{x}, \boldsymbol{Y}, \boldsymbol{V}, \boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_N) \\ &= \prod_{i=1}^{N} (2\pi)^{-|\boldsymbol{\psi}_i|/2} |\boldsymbol{V}_i|^{-1/2} \\ &\quad \times \exp\{-\tfrac{1}{2}[(\hat{\boldsymbol{b}} - \boldsymbol{b})^\intercal \boldsymbol{S}^{-1} (\hat{\boldsymbol{b}} - \boldsymbol{b}) - \hat{\boldsymbol{b}}^\intercal \boldsymbol{S}^{-1} \hat{\boldsymbol{b}} + \textstyle\sum_{i=1}^{N} \boldsymbol{y}_i^\intercal \boldsymbol{V}_i^{-1} \boldsymbol{y}_i]\}. \end{aligned} \tag{C.1.18}$$

While this expression is not simpler than the one above, if we ignore terms that do not involve $\boldsymbol{b}$, the likelihood is proportional to a multivariate normal density, $L(\boldsymbol{b}; \boldsymbol{x}, \boldsymbol{Y}, \boldsymbol{V}, \boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_N) \propto N(\boldsymbol{b}; \hat{\boldsymbol{b}}, \boldsymbol{S})$.

With these results, we can now apply the Bayes factor and posterior calculations derived in Propositions 4 to the missing data setting by making use of the formulae for the least-squares estimate $\hat{\boldsymbol{b}}$ and the variance-covariance matrix $\boldsymbol{S}$ given in (C.1.16, C.1.17).

Note that, when all $\boldsymbol{Y}$ are observed—that is, $|\boldsymbol{\psi}_i| = R$ for all $i = 1, \ldots, N$—then $\boldsymbol{V}_i = \boldsymbol{V}$ for all i, and therefore all expressions here reduce to those given in Section 4.1.1.

C.1.3 Bayesian simple multivariate regression with intercept and missing data

With missing data, we adopt a strategy to the previous section to include an intercept. However, the combination of the missing data and taking care of an intercept does introduce some extra complications, and some of the expressions derived above cannot be reused as easily.

With the notation defined in the previous section, the simple multivariate regression model with intercept is

$$\Psi_i \boldsymbol{y}_i \sim N_R(\Psi_i(\boldsymbol{b}_0 + x_i \boldsymbol{b}), \boldsymbol{V}_i). \tag{C.1.19}$$

The likelihood for $\boldsymbol{b}_0$, $\boldsymbol{b}$ is

$$L(\boldsymbol{b}_0, \boldsymbol{b}; \boldsymbol{x}, \boldsymbol{Y}, \boldsymbol{V}, \psi_1, \ldots, \psi_N) \coloneqq \prod_{i=1}^{N} p(\boldsymbol{y}_i | \boldsymbol{x}, \boldsymbol{b}_0, \boldsymbol{b}, \boldsymbol{V}, \psi_i)$$
$$= \prod_{i=1}^{N} (2\pi)^{-|\psi_i|/2} |\boldsymbol{V}_i|^{-1/2} \exp\{-\tfrac{1}{2} \textstyle\sum_{i=1}^{N} (\boldsymbol{y}_i - \boldsymbol{b}_0 - x_i \boldsymbol{b})^\intercal \boldsymbol{V}_i^{-1} (\boldsymbol{y}_i - \boldsymbol{b}_0 - x_i \boldsymbol{b})\}. \tag{C.1.20}$$

The least-squares estimate of $\boldsymbol{b}_0$, and its variance covariance matrix $\boldsymbol{S}_0$, are

$$\hat{\boldsymbol{b}}_0 = \bar{\boldsymbol{y}} - \bar{\boldsymbol{X}} \boldsymbol{b} \tag{C.1.21}$$

$$\boldsymbol{S}_0 = \frac{1}{N} \bar{\boldsymbol{V}}, \tag{C.1.22}$$

in which we define

$$\bar{\boldsymbol{V}}^{-1} \coloneqq \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{V}_i^{-1} \tag{C.1.23}$$

$$\bar{\boldsymbol{y}} \coloneqq \frac{1}{N} \bar{\boldsymbol{V}} \sum_{i=1}^{N} \boldsymbol{V}_i^{-1} \boldsymbol{y}_i \tag{C.1.24}$$

$$\bar{\boldsymbol{X}} \coloneqq \frac{1}{N} \bar{\boldsymbol{V}} \sum_{i=1}^{N} \boldsymbol{V}_i^{-1} x_i. \tag{C.1.25}$$

Note that in the special case when all $\boldsymbol{Y}$ are observed, that is when $\boldsymbol{V}_i^{-1} = \boldsymbol{V}^{-1}$ for all $i = 1, \ldots, N$, these definitions simplify greatly, $\bar{\boldsymbol{V}} = \boldsymbol{V}, \bar{\boldsymbol{y}} = \sum_{i=1}^{N} \boldsymbol{y}_i / N$ and $\bar{\boldsymbol{X}} = \sum_{i=1}^{N} x_i \boldsymbol{I}_R / N$, which are effectively means.

Similar to the full-data case, the profile likelihood—that is, the likelihood ob-

tained by first maximizing the likelihood with respect to the intercept—also has an analytic form, but unlike the full-data case, we cannot reuse the likelihood for the model without an intercept.

$$L^{\star}(\boldsymbol{b}; \boldsymbol{x}, \boldsymbol{Y}, \boldsymbol{V}, \boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_N) \coloneqq \max_{\boldsymbol{b}_0} L(\boldsymbol{b}_0, \boldsymbol{b}; \boldsymbol{x}, \boldsymbol{Y}, \boldsymbol{V}, \boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_N)$$
$$= \prod_{i=1}^{N} (2\pi)^{-|\boldsymbol{\psi}_i|/2} |\boldsymbol{V}_i|^{-1/2} \exp\{-\tfrac{1}{2} \textstyle\sum_{i=1}^{N} (\tilde{\boldsymbol{y}}_i - \tilde{\boldsymbol{X}}_i \boldsymbol{b})^{\intercal} \boldsymbol{V}_i^{-1} (\tilde{\boldsymbol{y}}_i - \tilde{\boldsymbol{X}}_i \boldsymbol{b})\}, \qquad \text{(C.1.26)}$$

in which $\tilde{\boldsymbol{y}}_i$ and $\tilde{\boldsymbol{X}}_i$ are the "centered" $\boldsymbol{y}_i$ and x_i:

$$\tilde{\boldsymbol{Y}} \coloneqq \boldsymbol{Y} - \mathbf{1}\bar{\boldsymbol{y}}^T, \qquad \text{(C.1.27)}$$

$$\tilde{\boldsymbol{X}}_i \coloneqq x_i \boldsymbol{I}_R - \bar{\boldsymbol{X}}. \qquad \text{(C.1.28)}$$

The least-squares estimate of $\boldsymbol{b}$, and its variance-covariance matrix $\boldsymbol{S}$, are

$$\hat{\boldsymbol{b}} = \boldsymbol{S} \textstyle\sum_{i=1}^{N} \tilde{\boldsymbol{X}}_i^{\intercal} \boldsymbol{V}_i^{-1} \tilde{\boldsymbol{y}}_i, \qquad \text{(C.1.29)}$$

$$\boldsymbol{S} = \big(\textstyle\sum_{i=1}^{N} \tilde{\boldsymbol{X}}_i^{\intercal} \boldsymbol{V}_i^{-1} \tilde{\boldsymbol{X}}_i \big)^{-1}. \qquad \text{(C.1.30)}$$

(The centered quantities $\tilde{\boldsymbol{y}}_i$ and $\tilde{\boldsymbol{X}}_i$ are only introduced to provide intuition; in practice, we limit computational effort in computing $\hat{\boldsymbol{b}}$ and $\boldsymbol{S}$ by expanding out $\tilde{\boldsymbol{y}}_i$ and $\tilde{\boldsymbol{X}}_i$ in the expressions above.) Using these quantities, the profile likelihood can be rewritten as (C.1.18), so that it is proportional to the multivariate normal density $N(\boldsymbol{b}; \hat{\boldsymbol{b}}, \boldsymbol{S})$, in which $\boldsymbol{b}$ and $\boldsymbol{S}$ are given by (C.1.29) and (C.1.30), and $\boldsymbol{Y}$ is replaced with $\tilde{\boldsymbol{Y}}$ from (C.1.27). The Bayesian computations now follow straightforwardly

using Proposition 4, in which least-squares estimate $\hat{\boldsymbol{b}}$ and the variance-covariance matrix $\boldsymbol{S}$ are given by (C.1.29) and (C.1.30).

To simplify the computation somewhat, we consider the following slight approximations:

$$\bar{y}_r \approx \sum_{i=1}^N \psi_{ir} y_{ir} / \sum_{i=1}^N \psi_{ir}, \tag{C.1.31}$$

$$\bar{x}_{rr'} \approx \begin{cases} \sum_{i=1}^N \psi_{ir} x_i / \sum_{i=1}^N \psi_{ir} & \text{if } r = r', \\ 0 & \text{otherwise.} \end{cases} \tag{C.1.32}$$

This approximation will be exact—that is, (C.1.31) and (C.1.32) will recover (C.1.24) and (C.1.25)—when either (1) all y_{ir}'s are observed, (2) $\boldsymbol{V}$ is diagonal, or (3) the missingness patterns do not overlap, *i.e.*, for all pairs of samples (i, i'), either $\boldsymbol{\psi}_i = \boldsymbol{\psi}_{i'}$ or $\boldsymbol{\psi}_i^\intercal \boldsymbol{\psi}_{i'} = 0$.

C.2 Details of Multivariate single-effect regression with a mixture prior

In the presence of intercept or missing values, we define multivariate single-effect regression similarly as (4.1.6) - (4.1.9), with the only change in the likelihood (4.1.6). To include an intercept in the *MSER* model, the (4.1.6) becomes

$$\boldsymbol{Y} \sim \mathrm{MN}_{N\times R}(\mathbf{1}\boldsymbol{b}_0^\intercal + \boldsymbol{X}\boldsymbol{B}, \boldsymbol{I}_N, \boldsymbol{V}). \tag{C.2.1}$$

In the presence of missing values, the (4.1.6) becomes

$$\Psi_i \boldsymbol{y}_i \sim N_R(\Psi_i \boldsymbol{B}^\intercal \boldsymbol{x}_i, \boldsymbol{V}_i), i = 1, \cdots, N. \tag{C.2.2}$$

With both intercept and missing values, the (4.1.6) becomes

$$\Psi_i \boldsymbol{y}_i \sim N_R(\Psi_i(\boldsymbol{b}_0 + \boldsymbol{B}^\intercal \boldsymbol{x}_i), \boldsymbol{V}_i), i = 1, \cdots, N. \tag{C.2.3}$$

The posterior inferences on $\boldsymbol{\gamma}$ and $\boldsymbol{B}$ are straightforward using Proposition 5 directly, in which the Bayes Factors and posterior first and second moments are from the simple regression model as described in Section C.1.1, C.1.2 and C.1.3.