THE UNIVERSITY OF CHICAGO

IMPROVED STATISTICAL METHODS FOR GENETIC ASSOCIATION STUDIES

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES
AND THE PRITZKER SCHOOL OF MEDICINE
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

COMMITTEE ON GENETICS, GENOMICS AND SYSTEMS BIOLOGY

BY
YUNQI YANG

CHICAGO, ILLINOIS
DECEMBER 2024

To my parents

Through the lens of statistics, we unravel the mysteries of life.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

Reflecting on my Ph.D. journey, I know I would never regret the decision of pursuing this path at the University of Chicago. The people I've met and the challenges I've been through have shaped me, making me a better version of myself.

First and foremost, I would like to express my deepest gratitude to my advisor, Professor Matthew Stephens. Matthew is an exceptional statistician. He always has good intuitions and his guidance and support have been invaluable to my research. Over these five years, I have gained not only technical skills but also an understanding of how to conduct research. His philosophy of starting from simple ideas and always questioning why things are as they are has greatly influenced my thinking. Matthew's curiosity and high standards have inspired me and shaped my perspective on what constitutes good research.

Then, I'd like to thank Peter Carbonetto for his help and support on my research. Peter is very patient and careful about details. I've learned very useful knowledge about scientific computing from him, which ease my life dealing with data. Then, I want to thank my committee members, Xin He (chair), Jingshu Wang and Haky Im for their insightful comments on my research progress. I also quite enjoying working with Jingshu on the research project in Chapter 4, where we learned together about genetics.

My sincere thanks go to the current and past members of the Stephens lab, especially those who have engaged in discussions and collaborations with me: Karl Tayeb, Yusha Liu, Yuxin Zou and Gao Wang. I am also grateful to the program in Computational Biology (PCB), the Department of Human Genetics, and the Committee on Genetics, Genomics, and Systems Biology. Being a part of these communities has been an incredibly rewarding experience, surrounded by people passionate about science, eager to share ideas, and dedicated to conducting groundbreaking research.

I would like to thank my friends who supported me throughout my Ph.D. journey. This five-year adventure has been more than an academic pursuit; it has been a defining period

of my twenties. There have been joyful and memorable moments, as well as challenges, doubts, and setbacks. I am immensely grateful for the friends who stood by me, offered understanding, and encouraged me, especially during my final year: Hsiang-Yu Tsai, Peiyi Zhang, Wei Lu, Xiaotong Sun, Xinyi Li, Yu Han, and Yusha Liu.

Lastly, I want to express my deepest gratitude to my parents, who have always provided me with the best they could offer. My father, a lifelong learner with a passion for acquiring knowledge, has always reminded me that every accomplishment is just the beginning, and that learning is a lifelong mission. Their approach to education has had a profound impact on me, ultimately guiding me toward the path of pursuing a Ph.D.

Thank you all for being a part of this incredible journey.

ABSTRACT

Genetic association studies have successfully identified numerous genetic variants associated with complex diseases and gene expression levels, providing unprecedented opportunities to discover new biology through downstream analyses. Examples of such analyses include multivariate methods, which can enhance the power to detect signals by borrowing information across similar or correlated conditions; fine-mapping analysis, which aims to identify potentially causal loci among many highly correlated genetic variants; and Mendelian randomization, which estimates the causal effect of one trait on another using genetic variants as instruments.

In this dissertation, we focus on improving methods for downstream analysis, with the goal of enhancing the power of statistical inference. In Chapter 2, we improve the fitting algorithm of a widely used multivariate method, the multivariate adaptive shrinkage (MASH) by Urbut et al. [2019]. In Chapter 3, we develop a new method for fine-mapping time-to-event outcomes, building on the existing "Sum of Single Effects" (SuSiE) fine-mapping approach by Wang et al. [2020]. In Chapter 4, we address the challenges associated with the small sample sizes of within-family genotype data. Specifically, we develop methods to improve the efficiency of estimates derived from within-family data, which also lead to variance reduction in Mendelian randomization.

CHAPTER 1

INTRODUCTION

Genetic association studies, such as genome-wide association studies (GWASs) and molecular quantitative trait locus (molQTL) mapping, have successfully identified numerous genetic variants linked to complex human traits and molecular phenotypes [Uffelmann et al., 2021, Abdellaoui et al., 2023, Visscher et al., 2017, Aguet et al., 2023, Gibson et al., 2015]. The extensive summary statistics generated from these studies offer unprecedented opportunities to uncover new biology. Traditionally, genetic association studies are conducted using a single-locus model under a single condition, where a single-nucleotide polymorphism (SNP) is tested for association with a trait of interest in a cohort of largely unrelated individuals. While this approach has been effective, it also presents challenges and opportunities for improving downstream analysis.

Opportunity 1: Enhancing statistical power beyond single condition analysis

While condition-by-condition analysis is simple and straightforward, it is not optimal for detecting significant signals. The widespread pleiotropy observed in the human genome indicates substantial biological overlap among different traits [Watanabe et al., 2019, Turley et al., 2018, Zou et al., 2023, Turchin and Stephens, 2019]. Additionally, in gene expression QTL (eQTL) analyses across multiple tissues, many eQTLs are found to have effects across multiple tissues or subsets of tissues [GTEx Consortium et al., 2015, Urbut et al., 2019, Barbeira et al., 2020, Natri et al., 2024]. Therefore, a more effective approach for analyzing such data is through multivariate analysis, which leverages information across similar conditions to improve the power of detecting significant units.

One key challenge in the multivariate analysis is to account for heterogeneous effect-sharing patterns—for instance, different eQTLs may act in different subsets of tissues and with different effect sizes. Empirical Bayes (EB) methods [Flutre et al., 2013, Urbut et al., 2019] offer an attractive solution by first estimating a “prior distribution" that captures the

sharing and similarity of effects across conditions, and then using Bayes theorem to combine the prior with observed data, thereby improving effect size estimation.

However, learning the complex sharing patterns from data is both statistically and computationally challenging, as it involves fitting a mixture of multivariate normals prior with unknown covariance matrices. In the multivariate adaptive shrinkage (MASH) proposed by Urbut et al. [2019], a two-stage procedure is employed to address these challenges. In the first stage, sharing patterns are estimated using an algorithm called "Extreme Deconvolution" (ED) by Bovy et al. [2011]. In the second stage, given the covariances estimated in the first stage, the mixture proportions in the prior are estimated by maximizing the likelihood across all data. However, we have noticed several limitations of the ED algorithm: it can be slow to converge, results are sensitive to initialization, and estimated covariance matrices can be quite unstable, particularly when the number of conditions, R, is large relative to the sample size, n.

In Chapter 2, we aim to improve the first step estimation in MASH. We describe new empirical Bayes methods that provide improvements in both speed and accuracy over existing methods. The two key ideas behind the methods are: (1) adaptive regularization to improve accuracy in high-dimensional settings with many conditions; (2) improving the speed of the model fitting algorithms by exploiting analytical results on the properties of the covariance matrices. Additionally, we provide an R package, udr ("Ultimate Deconvolution in R"), available at `https://github.com/stephenslab/udr`, which implements all these methods with a convenient user-friendly interface.

Problem 2: Extremely high correlations among nearby genetic variants due to Linkage Disequilibrium (LD)

Although GWASs have identified many significant variants associated with complex human traits, many of these are not causal variants. Non-causal genetic variants can also be significant in GWAS if they are correlated with nearby causal variants. Fine-mapping is an

analysis tool designed to identify putatively causal variants that contribute to these traits.

Genetic fine-mapping is often framed as a variable selection problem based on multiple regression models, where the outcome is the trait of interest, and the candidate predictor variables are the available genetic variants. Performing variable selection helps identify variants that may causally affect the trait. There are two key challenges specific to genetic fine-mapping as a variable selection problem. First, nearby genetic variants are often highly correlated, with many pairs having correlations greater than 0.99 or even equaling 1, making the inference even harder. Therefore, it is crucial for fine-mapping methods to have uncertainty quantification about which variables to select, and as a result, Bayesian variable selection in regression (BVSR) methods are often preferred in fine-mapping [Sillanpaa and Bhattacharjee, 2005, Servin and Stephens, 2007, Guan and Stephens, 2011, Carbonetto et al., 2012, Hormozdiari et al., 2014, Wallace et al., 2015, Newcombe et al., 2016, Wen et al., 2016].

Second, a genomic region usually contains thousands of SNPs, making the problem high dimensional. Therefore, it is crucial to develop methods that enable fast and scalable computation. Motivated by these demands, Wang et al. [2020] proposed a new fine-mapping approach, the "Sum of Single Effects" (SuSiE) regression. SuSiE reformulated the multiple regression model as the sum of multiple single effect regression models of Servin and Stephens [2007], where each single effect vector contains exactly one non-zero effect variable. This new reformulation not only leads to fast computation of posterior inference, but also provides convenient novel summaries of uncertainty, offering a Credible Set of variables for each selection. These features have made SuSiE widely used for fine-mapping.

Previous method development for fine-mapping has primarily focused on quantitative traits, such as height and body mass index. More recently, with the increasing of electronic health records linked to large biobanks, analyzing time-to-event (TTE) phenotypes—such as disease age of onset and progression—has become more common in genetics, providing

critical insights into the genetics of disease development and progression. A unique challenge in analyzing TTE data is that the time-to-event may not be observed for every individual, a phenomenon known as “censoring". A simple solution is to exclude individuals with missing outcomes and treat the time-to-event phenotype as a quantitative trait. This is not ideal as censored individuals still provide partial information—they have not yet experienced the event by the time they are censored. Fortunately, many statistical models have been developed to account for censoring, with one of the most widely used being the proportional hazards (PH) model by Cox [1972], commonly known as the CoxPH model.

In Chapter 3, we develop a new fine-mapping method for time-to-event phenotypes, where we extend the SuSiE framework to CoxPH model, which we refer to as CoxPH-SuSiE. CoxPH-SuSiE uses the same parameterization for covariates and a similar model-fitting procedure as SuSiE, thereby inheriting SuSiE's advantages. In our simulations, CoxPH-SuSiE outperformed other variable selection methods for TTE data. We also applied CoxPH-SuSiE to fine-map self-reported asthma cases in the UK Biobank.

Problem 3: GWAS effect estimates also capture contributions from the genotypes of an individual's family members.

The primary goal of GWAS is to estimate the direct effect of an individual's genetic variation on his/her own trait. However, the marginal association from a single-locus model also reflects other factors beyond the direct effect of the variant [Veller and Coop, 2024, Davies et al., 2019]. The effect size estimate of a SNP includes contributions from other causal SNPs in linkage disequilibrium (LD), which is the focus of Chapter 3 of this dissertation. In Chapter 4, we focus on another source of bias: the contributions from the genotypes of the individual's family members, and we call these effects the indirect genetic effects. One example is the genetic nurture effect, where parental genotype influences offspring outcomes by shaping the environment provided by the parents [Kong et al., 2018], see also Figure 1.1.

Failing to adjust for indirect genetic effects can lead to biases in Mendelian randomization

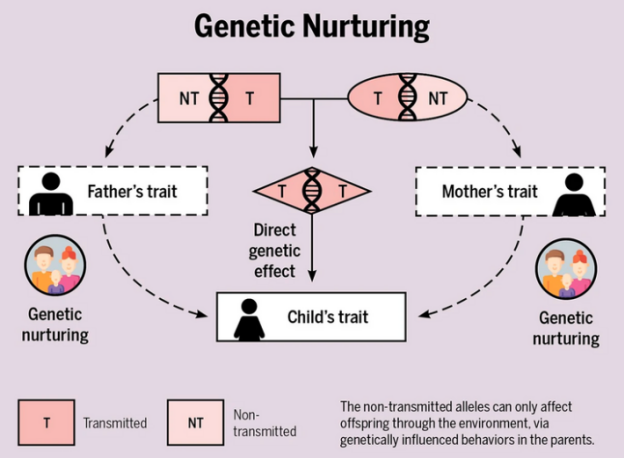

Figure 1.1: An illustration of genetic nurture path from Hart et al. [2021]

(MR). The genetic nurture path in the outcome trait opens a backdoor path between the genetic variants and the outcome, violating the independence assumption of MR [Brumpton et al., 2020, Davies et al., 2019]. Young et al. [2019] suggest that GWAS should be performed using family-based genotype data and models that account for indirect genetic effects. However, family-based genotype data is much less prevalent than population-based genotype data worldwide. As a result, estimates derived from family data have larger standard errors and are less precise.

In Chapter 4, we focus on reducing the variance of estimates derived from family-based genotype data. Following Chen and Chen [2000], we developed a variance calibration approach in the regression context that leverages information from large external datasets, which may not include family genotype data, such as population-based GWAS summary statistics. We use the biased estimates from large population studies to improve the efficiency of family-based estimates.

CHAPTER 2

IMPROVED METHODS FOR EMPIRICAL BAYES MULTIVARIATE MULTIPLE TESTING AND EFFECT SIZE ESTIMATION

Abstract

Estimating the sharing of genetic effects across different conditions is important to many statistical analyses of genomic data. The patterns of sharing arising from these data are often highly heterogeneous. To flexibly model these heterogeneous sharing patterns, Urbut et al. [2019] proposed the multivariate adaptive shrinkage (MASH) method to jointly analyze genetic effects across multiple conditions. However, multivariate analyses using MASH (as well as other multivariate analyses) require good estimates of the sharing patterns, and estimating these patterns efficiently and accurately remains challenging. Here we describe new empirical Bayes methods that provide improvements in speed and accuracy over existing methods. The two key ideas are: (1) adaptive regularization to improve accuracy in settings with many conditions; (2) improving the speed of the model fitting algorithms by exploiting analytical results on covariance estimation. In simulations, we show that the new methods provide better model fits, better out-of-sample performance, and improved power and accuracy in detecting the true underlying signals. In an analysis of eQTLs in 49 human tissues, our new analysis pipeline achieves better model fits and better out-of-sample performance than the existing MASH analysis pipeline. We have implemented the new methods, which we call "Ultimate Deconvolution", in an R package, udr, available on GitHub.

2.1 Introduction

The problem of testing and estimating effect sizes for many units in multiple conditions (or on multiple outcomes) arises frequently in genomics applications. Examples include assessing the effects of many expression quantitative trait loci (eQTLs) in multiple tissues [GTEx Consortium et al., 2015] and assessing the effects of many genetic variants on multiple traits [Zhou and Stephens, 2014, Pickrell et al., 2016, Turchin and Stephens, 2019, Udler et al., 2018, Zou et al., 2024]. The simplest approach to assessing effects in multiple conditions is to analyze each condition separately. However, this fails to exploit sharing or similarity of effects among conditions. For example, a genetic variant that increases expression of a particular gene in the heart may similarly increase expression in other tissues, particularly in tissues that are related to heart. Such sharing of effects could be exploited to improve power and estimation accuracy by borrowing information across conditions. This is, in essence, the motivation for meta-analysis methods [Willer et al., 2010, Han and Eskin, 2011, Wen and Stephens, 2014], and more generally it motivates consideration of multivariate approaches to multiple testing and effect size estimation [Urbut et al., 2019].

While borrowing information across conditions may be a natural idea, getting it to work well in practice requires confronting some thorny issues. One challenge is that the extent to which effects are shared among conditions will vary among data sets; for example, data sets involving very similar conditions may have very high levels of sharing, whereas data sets involving very different conditions may exhibit little to no sharing. Furthermore, some data sets may include some conditions that are very similar and others that are very dissimilar, and these differences may be difficult to specify in advance. Assessing the sharing and similarity of effects among conditions may also be an important goal in itself.

Empirical Bayes (EB) approaches (e.g., Flutre et al. 2013, Urbut et al. 2019) provide an attractive way to confront these challenges. EB methods estimate a prior distribution that captures the sharing or similarity of effects among the conditions, then, using Bayes theorem,

they combine the prior with the observed data to improve the effect estimates. The methods proposed in Urbut et al. [2019] and implemented in the R package mashr assume a mixture of multivariate normal distributions for the prior, which has the twin advantages of being both flexible and computationally convenient. Indeed, mashr has been used to analyze very large data sets involving many conditions [Zou et al., 2024, Lin et al., 2024, Urbut et al., 2021, Barbeira et al., 2020, Araujo et al., 2023, Li et al., 2022, Soliai et al., 2021, Natri et al., 2024, Bonder et al., 2021].

Despite this, these methods still have considerable limitations; in particular, fitting a mixture of multivariate normals prior with unknown covariance matrices raises statistical and computational challenges. Urbut et al. [2019] used a two-stage procedure to deal with (or sidestep) these challenges: in the first stage, the covariance matrices in the prior were estimated by maximum-likelihood on a subset of the data (using the "Extreme Deconvolution" algorithm of Bovy et al. 2011); next, given the covariances estimated in the first stage, the second stage estimated the mixture proportions in the prior by maximzing the likelihood from all the data (using the fast optimization algorithms of Kim et al. 2020). The second stage is a convex optimization problem, and can be solved efficiently and reliably for very large data sets. The first stage, however, presents several challenges, including that the Extreme Deconvolution (ED) algorithm can be slow to converge, the results of running ED are often sensitive to initialization, and the estimated covariance matrices can be quite unstable, particularly when the number of conditions, R, is large relative to sample size, n. These challenges motivated this work, which is a closer examination of these challenges from both a statistical perspective and a computational one. One of the contributions of this study is a new algorithm, which we call "truncated eigenvalue decomposition" (TED). TED often converges much faster than ED (noting that ED applies to some settings where TED does not). We also explore the use of simple regularization schemes that can improve accuracy compared with maximum-likelihood estimation, particularly when the sample size

n is small and the number of conditions R is large (that is, the ratio R/n is large). And we highlight some problems that arise from using low-rank covariance matrices, which was a strategy previously suggested in Urbut et al. [2019] to reduce the number of estimated parameters. We provide an R package, udr ("Ultimate Deconvolution in R"), available at `https://github.com/stephenslab/udr`, which implements all these methods described here within a convenient, user-friendly interface, and that interacts well with our previous R package, mashr [Urbut et al., 2019].

This chapter is structured as follows. Section 2.2 summarizes notation used in the mathematical expressions throughout this chapter. Section 2.3 formally introduces the models, priors and regularization schemes considered, and Section 2.4 describes the model-fitting algorithms and procedures. Section 2.5 evaluates the performance of the different methods and algorithms on data sets simulated under a variety of scenarios. Section 2.6 applies the improved methods to an analysis of a large, multi-tissue eQTL data set from the GTEx Project. Finally, Section 2.7 discusses the promise and limitations of our methods, and future directions.

2.2 Notation used in the mathematical expressions

For the mathematical expressions below, we use bold, capital letters ($\boldsymbol{A}$) to denote matrices; bold, lowercase letters ($\boldsymbol{a}$) to denote column vectors; and plain, lowercase letters (a) to denote scalars. For a matrix $\boldsymbol{A}$, $\|\boldsymbol{A}\|_*$ denotes its nuclear norm, $\boldsymbol{A}^T$ denotes its transpose, $\boldsymbol{A}^{-1}$ denotes its inverse, $\boldsymbol{A}^{-T}$ denotes the inverse of $\boldsymbol{A}^T$, $\mathrm{tr}(\boldsymbol{A})$ denotes the trace, and $|\boldsymbol{A}|$ denotes the matrix determinant. We use $\mathbf{0}$ and $\mathbf{1}$ to denote the vectors whose elements are all zeros and all ones respectively; $\boldsymbol{e}_r$ for the standard basis vector $\boldsymbol{e}_r = (0, \ldots, 0, 1, 0, \ldots, 0)$ with the 1 appearing in the rth position; $\boldsymbol{I}_R$ is the $R \times R$ identity matrix; and $\mathrm{diag}(\boldsymbol{a})$ denotes the diagonal matrix in which the diagonal entries are given by the entries of vector $\boldsymbol{a}$. We use $\mathbb{R}$ for the set of real numbers, $\mathbb{R}^R$ for the set of real-valued vectors of length R,

and $\mathbb{R}^{n\times m}$ for the set of real-valued $n \times m$ matrices. We use $N(\mu, \sigma^2)$ to denote the normal distribution on $\mathbb{R}$ with mean μ and variance σ^2, and $N(\,\cdot\,; \mu, \sigma^2)$ denotes its density. And we use $N_R(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for the multivariate normal distribution on $\mathbb{R}^R$ with mean $\boldsymbol{\mu} \in \mathbb{R}^R$ and $R \times R$ covariance matrix $\boldsymbol{\Sigma}$, and $N_R(\,\cdot\,; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes its density. We use $P_R^+ \subseteq \mathbb{R}^{R\times R}$ to denote the set of real-valued $R \times R$ (symmetric) positive semi-definite matrices. $\mathcal{S}_R \subseteq \mathbb{R}^R$ denotes the R-dimensional simplex.

2.3 The empirical Bayes multivariate normal means model

In this section, we define the "empirical Bayes multivariate normal means" (EBMNM) model. We describe several variations of this model that involve constraints or penalties on the model parameters.

The EBMNM model assumes that we observe vectors $\boldsymbol{x}_j \in \mathbb{R}^R$, that are independent, noisy, normally-distributed measurements of underlying true values $\boldsymbol{\theta}_j \in \mathbb{R}^R$:

$$\boldsymbol{x}_j \mid \boldsymbol{\theta}_j \sim N_R(\boldsymbol{\theta}_j, \boldsymbol{V}_j), \quad j = 1, \ldots, n, \tag{2.1}$$

in which the covariances $\boldsymbol{V}_j \in P_R^+$ are assumed to be known and invertible. Our ultimate goal is to perform inference for the unknown means $\boldsymbol{\theta}_j$ from the observed data $\boldsymbol{x}_j$. This model is a natural generalization of the well-studied (univariate) normal means model [Robbins, 1951, Efron and Morris, 1972, Stephens, 2017, Bhadra et al., 2019, Johnstone, 2019, Sun, 2020], and so we call it the "multivariate normal means model". An important special case is when the measurement error distribution is the same for all observations; i.e., $\boldsymbol{V}_j = \boldsymbol{V}$, $j = 1, \ldots, n$. We refer to this as the "homoskedastic" case. Some of our computational methods are designed specifically for the homoskedastic case, which is easier to solve than the "heteroskedastic" case.

The EBMNM model (2.1) further assumes that the unknown means are independent

and identically distributed draws from some distribution, which we refer to as the "prior distribution". While other choices are possible, here we assume the prior distribution is a mixture of zero-mean multivariate normals:

$$p(\boldsymbol{\theta}_j \mid \boldsymbol{\pi}, \mathcal{U}) = \sum_{k=1}^{K} \pi_k N_R(\boldsymbol{\theta}_j; \mathbf{0}, \boldsymbol{U}_k), \tag{2.2}$$

where $\boldsymbol{\pi} \in \mathcal{S}_K$ is the set of mixture proportions, $\boldsymbol{U}_k \in P_R^+$ are covariance matrices, and $\mathcal{U} := \{\boldsymbol{U}_1, \ldots, \boldsymbol{U}_K\}$ denotes the full collection of covariance matrices.

Combining (2.1) and (2.2) yields the marginal distribution

$$p(\boldsymbol{x}_j \mid \boldsymbol{\pi}, \mathcal{U}) = \sum_{k=1}^{K} \pi_k N_R(\boldsymbol{x}_j; \mathbf{0}, \boldsymbol{U}_k + \boldsymbol{V}_j), \tag{2.3}$$

and the marginal log-likelihood,

$$\begin{aligned} \ell(\boldsymbol{\pi}, \mathcal{U}) &:= \sum_{j=1}^{n} \log p(\boldsymbol{x}_j \mid \boldsymbol{\pi}, \mathcal{U}) \\ &= \sum_{j=1}^{n} \log \sum_{k=1}^{K} \pi_k N_R(\boldsymbol{x}_j; \mathbf{0}, \boldsymbol{U}_k + \boldsymbol{V}_j) \end{aligned} \tag{2.4}$$

The EB approach to fitting the model (2.1–2.2) proceeds in two stages:

1. Estimate the prior parameters $\boldsymbol{\pi}, \mathcal{U}$ by maximizing a penalized log-likelihood,

$$(\hat{\boldsymbol{\pi}}, \hat{\mathcal{U}}) := \underset{\boldsymbol{\pi} \in \mathcal{S}_K, \boldsymbol{U}_k \in P_R^{+,k}}{\operatorname{argmax}} \ell(\boldsymbol{\pi}, \mathcal{U}) - \sum_{k=1}^{K} \tilde{\rho}(\boldsymbol{U}_k), \tag{2.5}$$

   where $\tilde{\rho}$ denotes a penalty function introduced to regularize $\boldsymbol{U}_k$, and $P_R^{+,k} \subseteq P_R^+$ allows for constraints on $\boldsymbol{U}_k$ (which may be different for each component of the mixture). The exact penalties and constraints considered are described below. When $P_R^{+,k} = P_R^+$ and $\tilde{\rho}(\boldsymbol{U}_k) = 0$ for all $\boldsymbol{U}_k$, solving (2.5) corresponds to maximum-likelihood estimation of

$\boldsymbol{\pi}, \mathcal{U}$.

2. Compute the posterior distribution for each $\boldsymbol{\theta}_j$ given $\boldsymbol{\pi}, \boldsymbol{U}$ estimated in the first stage:

$$\begin{aligned} p_{\mathsf{post}}(\boldsymbol{\theta}_j) &:= p(\boldsymbol{\theta}_j \mid \boldsymbol{x}_j, \hat{\boldsymbol{\pi}}, \hat{\mathcal{U}}) \\ &\propto p(\boldsymbol{x}_j \mid \boldsymbol{\theta}_j)\, p(\boldsymbol{\theta}_j \mid \hat{\boldsymbol{\pi}}, \hat{\mathcal{U}}). \end{aligned} \tag{2.6}$$

The posterior distributions (2.6) have an analytic form, and are mixtures of multivariate normal distributions (see for example Urbut et al. 2019, Bovy et al. 2011 and Supplementary Section A.1). Since these posterior distributions have a closed form, it is relatively straightforward to compute posterior summaries such as the posterior mean and posterior standard deviation, and measures for significance testing such as the *local false sign rate* (*lfsr*) [Stephens, 2017]. Therefore, we focus on methods for accomplishing the first step, solving (2.5).

2.3.1 A reformulation to ensure scale invariance

Intuitively, one might hope that changing the units of measurement of all the observed $\boldsymbol{x}_j$ would simply result in corresponding changes to the units of the estimated $\boldsymbol{\theta}_j$. This idea can be formalized as requiring that solutions of the EBMNM model should obey a "scale invariance" property. This consideration motivates us to reformulate (2.5). Let $\hat{\boldsymbol{\theta}}_j(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, \boldsymbol{V}_1, \ldots, \boldsymbol{V}_n)$ denote the posterior mean for $\boldsymbol{\theta}_j$ computed by solving the EBMNM problem with data $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, \boldsymbol{V}_1, \ldots, \boldsymbol{V}_n$. We say that the solution is "scale invariant" if, for any $s > 0$, the following holds:

$$\hat{\boldsymbol{\theta}}_j(s\boldsymbol{x}_1, \ldots, s\boldsymbol{x}_n, s^2\boldsymbol{V}_1, \ldots, s^2\boldsymbol{V}_n) = s\hat{\boldsymbol{\theta}}_j(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, \boldsymbol{V}_1, \ldots, \boldsymbol{V}_n). \tag{2.7}$$

That is, multiplying all the observed data points by s (which multiplies the corresponding error variance by s^2) has the effect of multiplying the estimated means by s. (Note: we state scale invariance in terms of the posterior means only for simplicity; the concept is easily generalized to require that the whole posterior distribution for $\boldsymbol{\theta}_j$ scales similarly.)

Without the penalty $\tilde{\rho}$ in (2.5), it is easy to show that the scale invariance property (2.7) holds provided that the constraints satisfy $\boldsymbol{U}_k \in P_R^{+,k} \implies s_k \boldsymbol{U}_k \in P_R^{+,k}, \forall s_k > 0$. With penalty, scale invariance holds provided that $\tilde{\rho}(\boldsymbol{U}) = \tilde{\rho}(s\boldsymbol{U}), \forall s, \boldsymbol{U}$; that is, provided that the penalty function depends only on the "shape" of $\boldsymbol{U}$ and not on its "scale". To ensure scale invariance, we therefore consider penalties of the form

$$\tilde{\rho}(\boldsymbol{U}) = \min_{s>0} \rho(\boldsymbol{U}/s), \tag{2.8}$$

where ρ is a penalty that may depend on both the shape and scale of $\boldsymbol{U}$. To give some intuition, suppose that ρ encourages all the eigenvalues of $\boldsymbol{U}$ to be close to 1. Then $\tilde{\rho}$ will encourage the eigenvalues to be close to each other, without requiring that they specifically be close to 1. Therefore, plugging (2.8) into (2.5), we have

$$(\hat{\boldsymbol{\pi}}, \hat{\mathcal{U}}) := \underset{\boldsymbol{\pi} \in \mathcal{S}_K, \boldsymbol{U}_k \in P_R^{+,k}, \boldsymbol{s} > \boldsymbol{0}}{\operatorname{argmax}} \ell(\boldsymbol{\pi}, \mathcal{U}) - \sum_{k=1}^{K} \rho(\boldsymbol{U}_k / s_k), \tag{2.9}$$

where $\boldsymbol{s} := (s_1, \ldots, s_K)$. We use (2.9) for the remainder of this chapter.

2.3.2 Constraints and penalties

Estimating covariance matrices in high-dimensional settings (*i.e.*, large R) is known to be a challenging problem (e.g., Johnstone and Paul 2018, Fan et al. 2016, Ledoit and Wolf 2022). Even in the simpler setting of independent and identically distributed observations from a single multivariate normal distribution, the maximum-likelihood estimate of the covariance

matrix can be unstable, and so various covariance regularization approaches have been proposed to address this issue [Ledoit and Wolf, 2004, Friedman et al., 2008, Cai and Liu, 2011, Won et al., 2013, Chi and Lange, 2014]. Interestingly, in the context of using EBMNM for significance testing, adding penalties have additional benefits (see the numerical experiments below).

We consider two different penalties that have been previously used for covariance regularization:

1. The "inverse Wishart" (IW) penalty:

$$\rho_\lambda^{\mathrm{IW}}(\boldsymbol{U}) := \frac{\lambda}{2}\{\log|\boldsymbol{U}| + \mathrm{tr}(\boldsymbol{U}^{-1})\} \tag{2.10}$$

$$= \frac{\lambda}{2}\sum_{r=1}^{R}\big(\log e_r + 1/e_r\big). \tag{2.11}$$

2. The "nuclear norm" (NN) penalty:

$$\rho_\lambda^{\mathrm{NN}}(\boldsymbol{U}) := \frac{\lambda}{2}\{0.5\|\boldsymbol{U}\|_* + 0.5\|\boldsymbol{U}^{-1}\|_*\} \tag{2.12}$$

$$= \frac{\lambda}{2}\sum_{r=1}^{R}\big(0.5e_r + 0.5/e_r\big). \tag{2.13}$$

Here, $e_1, \ldots, e_R$ denote the eigenvalues of $\boldsymbol{U}$, and $\lambda > 0$ controls the strength of the penalty. We chose $\lambda = R$ in our simulations, but one could also use cross-validation to select λ.

The IW penalty on $\boldsymbol{U}_k$ corresponds to *maximum a posteriori* (MAP) estimation of $\boldsymbol{U}_k$ under an inverse-Wishart prior, with prior mode $\boldsymbol{I}_R$ and $\lambda - R - 1$ degrees of freedom [Fraley and Raftery, 2007]. This penalty was also mentioned (but not used or evaluated) in Bovy et al. [2011].

The nuclear norm penalty was studied as an alternative to the IW penalty for estimation of covariance matrices in Chi and Lange [2014]. They claimed that "In this paper, we

introduce a novel prior which effects the desired adjustment on the sample eigenvalues. Maximum a posteriori (MAP) estimation under the prior boils down to a simple nonlinear transformation of the sample eigenvalues." To our knowledge, this penalty has not been studied in the EBMNM setting. The nuclear norm penalty in Chi and Lange [2014] includes a hyperparameter $\alpha \in (0, 1)$ that controls the balance between $\|\boldsymbol{U}\|_*$ and $\|\boldsymbol{U}^{-1}\|_*$, but our approach to ensuring scale-invariance has the consequence that changing α is equivalent to changing λ (see Supplementary Section A.7), so we set $\alpha = 0.5$.

As can be seen from (2.11) and (2.13), both penalties depend only on the eigenvalues of $\boldsymbol{U}$, and decompose into additive functions of the R eigenvalues. Both penalties are minimized when $\boldsymbol{U} = \boldsymbol{I}_R$, and more generally encourage $\boldsymbol{U}$ to be well-conditioned by making it closer to the identity matrix (by pushing the eigenvalues closer to 1).

As an alternative to penalized estimation of $\boldsymbol{U}$, we also consider estimating $\boldsymbol{U}$ under different constraints:

1. A scaling constraint, $\boldsymbol{U}_k = c_k \boldsymbol{U}_{0k}$, for some chosen $\boldsymbol{U}_{0k} \in P_R^+$.

2. A rank-1 constraint, $\boldsymbol{U}_k = \boldsymbol{u}_k \boldsymbol{u}_k^T$, for some $\boldsymbol{u}_k \in \mathbb{R}^R$.

The scaling constraint could be useful for situations in which the $\boldsymbol{\theta}_j$ may obey an expected sharing structure, or for sharing structures that are easier to interpret. For example, $\boldsymbol{U}_{0k} = \boldsymbol{I}_R$ captures the situation in which all the effects $\theta_{j1}, \ldots, \theta_{jR}$ are independent, and $\boldsymbol{U}_{0k} = \boldsymbol{1}\boldsymbol{1}^T$ captures the situation in which all the effects $\theta_{j1}, \ldots, \theta_{jR}$ are equal. Such covariances are referred to as "canonical" covariance matrices in Urbut et al. [2019].

The rank-1 constraint also leads to potentially more interpretable covariance matrices, and can be thought of as a form of regularization because low-rank matrices have fewer parameters to be estimated. It may also allow for faster computations. Urbut et al. [2019] in fact make extensive use of the rank-1 constraint. However, our results will show that this constraint can cause problems for significance testing and so may be better avoided in practice.

2.4 Fitting the EBMNM model

We now describe three algorithms we have implemented for fitting variations of the EBMNM model described above: the ED algorithm from Bovy et al. 2011; an algorithm based on methods commonly used in factor analysis (FA); and another based on the truncated eigenvalue decomposition (TED). Each of these algorithms applies to a subset of EBMNM models (Table 2.1). In some situations, only one algorithm can be applied; for example, only ED can handle heteroskedastic variances with no constraints on $\boldsymbol{U}$. However, in other settings all three algorithms can be applied (e.g., homoskedastic errors, no constraints on $\boldsymbol{U}$). Below, we empirically assess the relative merits of the different algorithms in simulations, see also Supplementary Tables A.1 and A.2 for a comparison of the algorithms' computational properties in the different settings.

	constraints on $\boldsymbol{U}$			
	none		rank-1	
algorithm	hom.	het.	hom.	het.
ED	✓	✓		
FA	✓		✓	✓
TED	✓		✓	

Table 2.1: Summary of the EBMNM algorithms and the situations in which they apply. In this table we consider 4 variations of EBMNM: no constraint on $\boldsymbol{U}$ or a rank-1 constraint; and homoskedastic (hom.) errors (all $\boldsymbol{V}_j$ are the same) or heteroskedastic (het.) errors (one or more $\boldsymbol{V}_j$ differ). Abbreviations used in this table are: ED = Extreme Deconvolution), FA = factor analysis, TED = truncated eigenvalue decomposition. A checkmark (✓) indicates that the algorithm (ED, FA, TED) can be applied to the particular variation. Note that fitting $\boldsymbol{U}$ with a scaling constraint involves only a 1-d optimization and is treated separately.

2.4.1 Algorithms for the single-component EBMNM model with no penalty

While these algorithms are implemented for the mixture prior (2.2) with the penalties described above, the algorithms are much easier to describe in the special case of one mixture component ($K = 1$), and without penalty. So initially we focus on this simpler case, and later we extend to the general form with penalties and $K \geq 1$.

With $K = 1$ and no penalty, the prior is $\boldsymbol{\theta}_j \sim N(\mathbf{0}, \boldsymbol{U})$, the model is

$$\boldsymbol{x}_j \mid \boldsymbol{U} \sim N_R(\mathbf{0}, \boldsymbol{U} + \boldsymbol{V}_j), \quad j = 1, \ldots, n. \tag{2.14}$$

and the goal is to compute the maximum-likelihood estimate of $\boldsymbol{U}$:

$$\hat{\boldsymbol{U}} := \underset{\boldsymbol{U} \in P_R^+}{\operatorname{argmax}} \sum_{j=1}^{n} \log N_R(\boldsymbol{x}_j; \mathbf{0}, \boldsymbol{U} + \boldsymbol{V}_j) \tag{2.15}$$

The three algorithms for solving (2.15) are as follows.

Truncated Eigenvalue Decomposition (TED) This algorithm, which to the best of our knowledge is new in this context, exploits the fact that, in the special case where $\boldsymbol{V}_j = \boldsymbol{I}_R$, $j = 1, \ldots, n$, the maximum-likelihood estimate (2.15) can be computed exactly. At first glance, one might try to solve for $\hat{\boldsymbol{U}}$ by setting $\hat{\boldsymbol{U}} + \boldsymbol{I}_R$ to the sample covariance matrix, $\boldsymbol{S} := \sum_{j=1}^{n} \boldsymbol{x}_j \boldsymbol{x}_j^T / n$, then recovering $\hat{\boldsymbol{U}}$ as $\hat{\boldsymbol{U}} = \boldsymbol{S} - \boldsymbol{I}_R$. However, $\boldsymbol{S} - \boldsymbol{I}_R$ is not necessarily a positive semi-definite matrix; that is, it may have one or more eigenvalues that are negative. One could deal this problem by setting the negative eigenvalues to zero, and indeed this approach is correct; that is, letting $(\boldsymbol{S})_+$ be the matrix obtained from $\boldsymbol{S}$ by performing an eigenvalue decomposition of $\boldsymbol{S}$ and truncating its negative eigenvalues to 0, $\hat{\boldsymbol{U}} = (\boldsymbol{S} - \boldsymbol{I}_R)_+$ is the maximum-likelihood estimate of $\boldsymbol{U}$ [Tipping and Bishop, 1999]. This idea can also be used to solve the more general case, $\boldsymbol{V}_j = \boldsymbol{V}$, essentially by transforming the data to $\boldsymbol{R}^{-1}\boldsymbol{x}_j$ where $\boldsymbol{V} = \boldsymbol{R}\boldsymbol{R}^T$, estimating $\boldsymbol{R}^{-1}\boldsymbol{U}\boldsymbol{R}^{-T}$ from this transformed data, then reversing this

transformation, see Supplementary Section A.4.

Extreme Deconvolution (ED) This is an EM algorithm [Dempster et al., 1977], an iterative approach to solving (2.15); the name comes from Bovy et al. [2011]. ED uses the natural "data augmentation" representation of (2.14):

$$\begin{aligned} \boldsymbol{\theta}_j &\sim N_R(\mathbf{0}, \boldsymbol{U}) \\ \boldsymbol{x}_j \mid \boldsymbol{\theta}_j &\sim N_R(\boldsymbol{\theta}_j, \boldsymbol{V}_j). \end{aligned} \tag{2.16}$$

Following the usual EM derivation, the updates can be derived as

$$\boldsymbol{U}^{\text{new}} = \frac{1}{n} \sum_{j=1}^{n} \boldsymbol{B}_j + \boldsymbol{b}_j \boldsymbol{b}_j^T, \tag{2.17}$$

where $\boldsymbol{b}_j$ and $\boldsymbol{B}_j$ are, respectively, the posterior mean and covariance of $\boldsymbol{\theta}_j$ given $\boldsymbol{U}$:

$$\boldsymbol{b}_j := \boldsymbol{U}(\boldsymbol{U} + \boldsymbol{V}_j)^{-1} \boldsymbol{x}_j \tag{2.18}$$

$$\boldsymbol{B}_j := \boldsymbol{U} - \boldsymbol{U}(\boldsymbol{U} + \boldsymbol{V}_j)^{-1} \boldsymbol{U}. \tag{2.19}$$

The update (2.17) is guaranteed to increase (or not decrease) the objective function in (2.15), and repeated application of (2.17–2.19) will converge to a stationary point of the objective.

Factor analysis (FA) This is also an EM algorithm, but based on a different data augmentation than ED; the name comes from its close connection to EM algorithms for factor analysis models [Ghahramani and Hinton, 1996, Rubin and Thayer, 1982, Zhao et al., 2008, Liu and Rubin, 1998, McLachlan and Peel, 2000]. In its simplest form, the FA approach imposes a rank-1 constraint on $\boldsymbol{U}$, $\boldsymbol{U} = \boldsymbol{u}\boldsymbol{u}^T$, where $\boldsymbol{u} \in \mathbb{R}^R$ is to be estimated. The model

(2.14) then admits the following data augmentation representation:

$$\begin{aligned} a_j &\sim N(0,1) \\ \boldsymbol{x}_j \mid \boldsymbol{u}, \boldsymbol{V}_j, a_j &\sim N(a_j \boldsymbol{u}, \boldsymbol{V}_j). \end{aligned} \tag{2.20}$$

The usual EM derivation gives the update

$$\boldsymbol{u}^{\text{new}} = \left(\sum_{j=1}^{n} (\mu_j^2 + \sigma_j^2) \boldsymbol{V}_j^{-1} \right)^{-1} \left(\sum_{j=1}^{n} \mu_j \boldsymbol{V}_j^{-1} \boldsymbol{x}_j \right), \tag{2.21}$$

in which μ_j and σ_j^2 denote, respectively, the posterior mean and posterior covariance of a_j given $\boldsymbol{u}$,

$$\mu_j := \sigma_j^2 \boldsymbol{u}^T \boldsymbol{V}_j^{-1} \boldsymbol{x}_j \tag{2.22}$$

$$\sigma_j^2 := 1/(1 + \boldsymbol{u}^T \boldsymbol{V}_j^{-1} \boldsymbol{u}). \tag{2.23}$$

The update (2.21) is guaranteed to increase (or not decrease) the objective function in (2.15), and repeated application of (2.21–2.23) will converge to a stationary point of the objective. These updates can be extended to higher-rank covariances where the goal is to find the maximum-likelihood estimate subject of $\boldsymbol{U}$ subject to $\boldsymbol{U}$ having rank at most R', where $R' \leq R$. However, when $R' > 1$, the updates have closed-form expressions only for homoskedastic errors, $\boldsymbol{V}_j = \boldsymbol{V}$.

The three algorithms have different strengths and weaknesses, and different settings to which they apply (Table 2.1). The TED algorithm has the advantage of directly computing the maximum-likelihood estimate, which seems preferable to an iterative approach. However, TED only applies in the case of homoskedastic errors. The ED approach is more general, applying to both heteroskedastic and homoskedastic errors, although it cannot fit rank-1 matrices. The FA approach is attractive for low-rank matrices, particularly for fitting rank-

1 matrices.

Our claim that ED cannot fit rank-1 covariance matrices deserves discussion, especially since Urbut et al. [2019] used ED to fit such matrices. As pointed out by Urbut et al. [2019], if ED is initialized to a low-rank matrix with rank R', then the updated estimates (2.17) are also rank (at most) R'. Thus, if ED is initalized to a rank-1 matrix, the final estimate is also rank-1. However, the ED estimates are not only low rank, but also span the same subspace as the initial estimates, a property we refer to as "subspace-preserving". Thus, if ED is initialized to $\boldsymbol{U} = \boldsymbol{u}\boldsymbol{u}^T$, the updated estimates will be of the form $a\boldsymbol{u}\boldsymbol{u}^T$ for some scalar a. In other words, the ED update does not change rank-1 matrices, except by a scaling factor, and so the final estimate will simply be proportional to the initial estimate. This flaw motivated us to implement the FA method. However, as our numerical comparisons will illustrate, the rank-1 matrices turn out to have other drawbacks that lead us not to recommend their use anyhow.

2.4.2 Extending the algorithms to a mixture prior

All of the algorithms—TED, ED and FA—can be generalized from the $K = 1$ case to the $K \geq 1$ case using the standard EM approach to dealing with mixtures. The resulting algorithms have a simple common structure, summarized in Algorithm 1. (This algorithm also allows for an inclusion of a penalty, which is treated in the next section. See also Supplementary Section A.2 and A.3 for a derivation of this algorithm.) This EM algorithm involves iterating the following steps: (i) compute the weights, w_{jk} (sometimes called the "responsibilities"), each which represent the conditional probability that mixture component k gave rise to observation j given the current estimates of $\boldsymbol{\pi}$,$\mathcal{U}$,

$$w_{jk} = \frac{\pi_k N_R(\boldsymbol{x}_j; \mathbf{0}, \boldsymbol{U}_k + \boldsymbol{V}_j)}{\sum_{k'=1}^{K} \pi_{k'} N_R(\boldsymbol{x}_j; \mathbf{0}, \boldsymbol{U}_{k'} + \boldsymbol{V}_j)}; \tag{2.24}$$

Algorithm 1: EM for fitting the EBMNM model.

Input: Data vectors $\boldsymbol{x}_j \in \mathbb{R}^R$ and correspoding covariance matrices $\boldsymbol{V}_j \in P_R^+$, $j = 1, \ldots, n$; K, the number of mixture components; initial estimates of the prior covariance matrices $\mathcal{U} = \{\boldsymbol{U}_1, \ldots, \boldsymbol{U}_K\}$, $\boldsymbol{U}_k \in P_R^{+,k}$, $k = 1, \ldots, K$; initial estimates of the scaling parameters $\boldsymbol{s} = \{s_1, \ldots, s_K\} \in \mathbb{R}^K$; initial estimates of the mixture weights $\boldsymbol{\pi} \in \mathcal{S}_K$.

Output: $\mathcal{U}$, $\boldsymbol{\pi}$.

repeat
  for $j \leftarrow 1$ **to** n **do**
    for $k \leftarrow 1$ **to** K **do**
      Update w_{jk} using (2.24).
    end
  end
  for $k \leftarrow 1$ **to** K **do**
    $\pi_k \leftarrow \sum_{j=1}^n w_{jk}/n$
    $\boldsymbol{U}_k \leftarrow \operatorname{argmax}_{\boldsymbol{U} \in P_R^{+,k}} \phi(\boldsymbol{U}; \boldsymbol{w}_k) - \rho(\boldsymbol{U}/s_k)$
      ▷ Note that some algorithms compute this argmax inexactly.
    $s_k \leftarrow \operatorname{argmin}_{s>0} \rho(\boldsymbol{U}_k/s)$
  end
until convergence criterion is met;

(ii) update $\boldsymbol{\pi}$ by averaging the weights (this is the standard EM update for estimating mixture proportions, and is the same for all the algorithms); (iii) update the covariance matrices $\mathcal{U}$ (this is the step where the algorithms differ); and (iv) update the scaling parameters, $\boldsymbol{s}$. (The update of the scaling parameters is the same for all algorithms, and depends on the chosen penalty. For details, see Supplementary Section A.5.)

With this data augmentation, the updates for the covariance matrices $\boldsymbol{U}_k$ have the following form:

$$\boldsymbol{U}^{\text{new}} = \operatorname*{argmax}_{\boldsymbol{U} \in P_R^{+,k}} \phi(\boldsymbol{U}; \boldsymbol{w}_k), \tag{2.25}$$

where

$$\phi(\boldsymbol{U}; \boldsymbol{w}) := \sum_{j=1}^n w_j \log N_R(\boldsymbol{x}_j; \boldsymbol{0}, \boldsymbol{U} + \boldsymbol{V}_j). \tag{2.26}$$

The function ϕ can be viewed as a *weighted* version of the log-likelihood with normal

prior (2.14), and the updates for this weighted problem are very similar to the updates for a normal prior. For example, the TED update, which solves the weighted problem exactly in the case $\boldsymbol{V}_j = \boldsymbol{I}_R$, involves truncating the eigenvalues of $\hat{\boldsymbol{S}} - \boldsymbol{I}_R$ where $\hat{\boldsymbol{S}} := \sum_{j=1}^n w_j \boldsymbol{x}_j \boldsymbol{x}_j^T / (\sum_{j=1}^n w_j)$ is *the weighted sample covariance matrix.* Details of the TED, ED and FA updates for weighted log-likelihoods are given in Supplementary Section A.5.

2.4.3 Modifications to the algorithms to incorporate the penalties

With a penalty, the updates (2.25) are instead

$$\boldsymbol{U}^{\text{new}} = \underset{\boldsymbol{U} \in P_R^{+,k},\, s_k > 0}{\text{argmax}} \quad \phi(\boldsymbol{U}; \boldsymbol{w}_k) - \rho(\boldsymbol{U}/s_k). \tag{2.27}$$

We have adapted the TED approach, in the case $\boldsymbol{V}_j = \boldsymbol{I}_R$, to incorporate either the IW or NN penalty. These extensions replace the simple truncation of the eigenvalues with solving R (1-d) optimization problem for each eigenvalue e_r, see equation (2.11) and (2.13); while these 1-d optimization problems do not have closed form solutions, they are easily solved using off-the-shelf numerical methods. This approach can also be applied, via the data transformation approach, to the general homoskedastic case $\boldsymbol{V}_j = \boldsymbol{V}$; however, this transformation approach implicitly changes the penalty so that it encourages $\boldsymbol{U}_k/s_k$ to be close to $\boldsymbol{V}$ rather than being close to $\boldsymbol{I}_R$. This change in the penalty appears to be necessary to make the TED approach work when $\boldsymbol{V} \neq \boldsymbol{I}_R$. See Supplementary Section A.5 for details.

Incorporating an IW penalty into the ED approach is also straightforward [Bovy et al., 2011] and results in a simple change to the closed-form updates (2.17). The NN penalty does not result in closed-form updates for ED and so we have not implemented it.

Incorporating penalties into the FA updates may be possible but we have not done so.

2.4.4 Significance testing

In EBMNM, inferences are based on the posterior distribution, $p_{\mathsf{post}}(\boldsymbol{\theta}_j) = p(\boldsymbol{\theta}_j \mid \boldsymbol{x}_j, \hat{\mathcal{U}}, \hat{\boldsymbol{\pi}})$, in which $\hat{\mathcal{U}}, \hat{\boldsymbol{\pi}}$ denote the estimates returned by Algorithm 1. To test for significance, we use the local false sign rate (*lfsr*), which has been used in both univariate [Stephens, 2017, Xie and Stephens, 2022] and multivariate [Urbut et al., 2019, Liu et al., 2023, Zou et al., 2024] settings. The *lfsr* is defined as

$$lfsr_{jr} := \min\{p_{\mathsf{post}}(\theta_{jr} \geq 0), p_{\mathsf{post}}(\theta_{jr} \leq 0)\}. \tag{2.28}$$

In particular, a small *lfsr* indicates high confidence in the sign of θ_{jr}. The *lfsr* is robust to modeling assumptions, which is helpful for reducing sensitivity to the choice of prior [Stephens, 2017].

2.5 Numerical comparisons

We ran simulations to (i) compare the performance of the different approaches to updating the covariance matrices $\boldsymbol{U}_k$; (ii) assess the benefits of the penalties and constraints; and (iii) assess the sensitivity of the results to the choice of K, the number of mixture components in the prior.

We used the Dynamic Statistical Comparisons software (`https://github.com/stephenslab/dsc`) to perform the simulations. The code and workflowr website is also available online at `https://github.com/yunqiyang0215/udr-paper`.

2.5.1 Data generation

We simulated all data sets from the EBMNM model (2.1–2.2); that is, for each data set, we simulated the "true" means $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n \in \mathbb{R}^R$ from the mixture prior (2.2) then we simulated observed data, the "noisy" means $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^R$, independently given the true means.

Note that model fitting and inferences were performed using only $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$; the true means $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n$ were only used to evaluate accuracy of the inferences. Further, to evaluate the ability of the model to generalize to other data, we also simulated test sets with true means $\boldsymbol{\theta}_1^{\text{test}}, \ldots, \boldsymbol{\theta}_{n_{\text{test}}}^{\text{test}}$ and observed vectors $\boldsymbol{x}_1^{\text{test}}, \ldots, \boldsymbol{x}_{n_{\text{test}}}^{\text{test}}$. These test sets were simulated in the same way as the training sets.

In all cases, we set the number of mixture components K to 10, with uniform mixture weights, $\pi_1, \ldots, \pi_{10} = 1/10$. We generated the $K = 10$ covariance matrices $\boldsymbol{U}_1, \ldots, \boldsymbol{U}_{10}$ in two different ways, which we refer to as the "hybrid" and "rank-1" scenarios:

1. **Hybrid scenario.** We used 3 canonical covariance matrices, and randomly generated an additional 7 covariance matrices randomly from an inverse-Wishart distribution with scale matrix $\boldsymbol{S} = 5\boldsymbol{I}_R$ and $\nu = R + 2$ degrees of freedom. The 3 canonical matrices were as follows: $\boldsymbol{U}_1 = 5\boldsymbol{e}_1\boldsymbol{e}_1^T$, a matrix of all zeros except for a 5 in the top-left position, which generates "singleton" mean vectors with a single non-zero element, $\boldsymbol{\theta}_j = (\theta_{j1}, 0, \ldots, 0)$; $\boldsymbol{U}_2 = 5\mathbf{1}\mathbf{1}^T$, which generates equal means $\boldsymbol{\theta}_j = (\alpha_j, \ldots, \alpha_j)$ for some scalar α_j; and $\boldsymbol{U}_3 = 5\boldsymbol{I}_R$, which generates mean vectors $\boldsymbol{\theta}_j$ that are independent in each dimension.

2. **Rank-1 scenario.** We used 5 covariance matrices of the form $\boldsymbol{U}_k = 5\boldsymbol{e}_k\boldsymbol{e}_k^T$, $k = 1, \ldots 5$, which generate mean vectors of length R with zeros everywhere except at the kth position. The remaining 5 covariances were random rank-1 matrices of the form $\boldsymbol{U}_k = \boldsymbol{u}_k\boldsymbol{u}_k^T$, $\boldsymbol{u}_k \sim N_R(\mathbf{0}, \boldsymbol{I}_R)$, $k = 6, \ldots, 10$.

In both scenarios, we simulated large, low-dimension data sets ($n = 10{,}000$, $R = 5$), and smaller, high-dimension data sets ($n = 1{,}000$, $R = 50$). We refer to these data sets as "large n/R" and "small n/R", respectively. To allow comparisons between TED, ED and FA, in all cases we simulated data sets with homoskedastic errors; that is, $\boldsymbol{V}_j = \boldsymbol{I}_R$, $j = 1, \ldots, n$ and $\boldsymbol{V}_j^{\text{test}} = \boldsymbol{I}_R$, $j = 1, \ldots, n_{\text{test}}$.

2.5.2 Results

Comparison of convergence

We first focused on comparing the convergence of the different updates (TED, ED and FA). For brevity, we write "TED" as shorthand for "the algorithm with TED updates", and similarly for ED and FA. To compare the updates under the same conditions, we used FA here to fit full-rank covariance matrices, not rank-1 matrices. We ran TED and ED with and without the IW penalty.

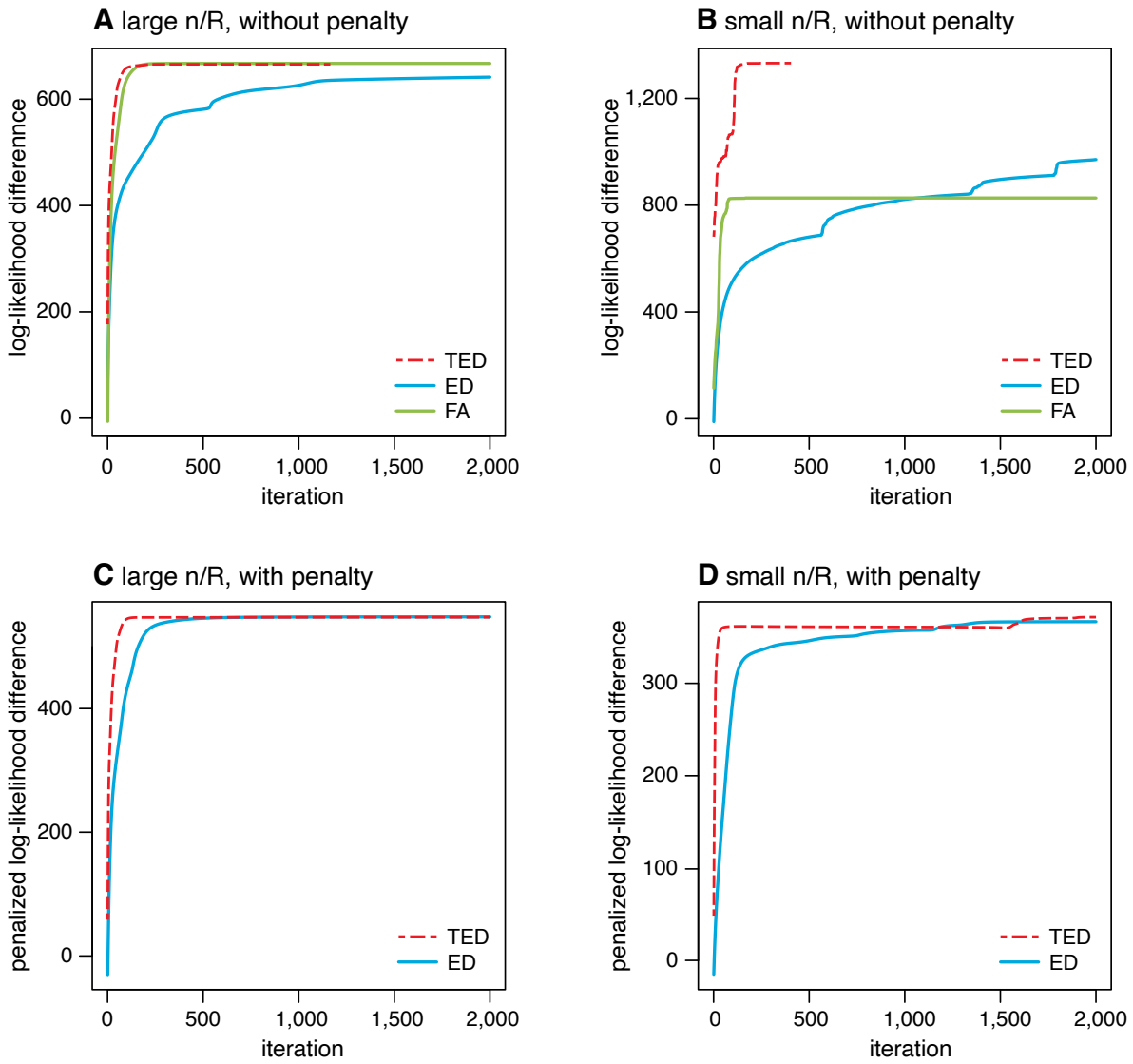

Figure 2.1: Illustrative examples comparing convergence of TED, ED and FA. Each plot shows the algorithms' progress over iterations on a single simulated data set. A and C show the results for the same data set, and similarly for B and D. In all cases, we ran 2,000 iterations, although in some cases, TED updates stopped early because the updates converged to a stationary point of the objective (the likelihood or the penalized likelihood). For examples A and B, no penalty was used; for examples C and D, the inverse Wishart (IW) penalty was used with penalty strength $\lambda = R$. Log-likelihood and penalized log-likelihood differences are plotted with respect to the (penalized) log-likelihood near the initial estimate. An initial round of 20 ED iterations common to all the runs (the "warm start") is not shown in these plots.

We ran all methods on 100 "large n/R" data sets and 100 "small n/R" data sets simulated

in the hybrid scenario, setting $K = 10$ to match the simulated truth. To reduce the likelihood that the different updates converge to different local solutions, we performed a prefitting stage in which we ran 20 iterations of ED from a random initial starting point (specifically, the initialization was $\pi_k = 1/10$, $s_k = 1$, with randomly generated $\boldsymbol{U}_k$, $k = 1, \ldots, 10$). We call this a "warm start". We then ran each algorithm for at most 2,000 iterations after this warm start. Figure 2.1 shows illustrative results for two data sets (one large n/R and one small n/R), and Figure 2.2 summarizes the results across all simulations.

The results in Figure 2.1 illustrate typical behavior. Among the unpenalized updates, TED and FA converged much more quickly than ED. For the penalized updates, TED still converged more quickly than ED, but the difference is less striking than without the penalty. In the small n/R example the three methods appear to have converged to different solutions (despite the use of a warm start). This is not unexpected due to the non-convexity of the objective function, and illustrates an important general point to keep in mind: improvements in the quality of solution obtained may be due to either faster convergence to the solution, convergence to a local solution with higher objective, or both.

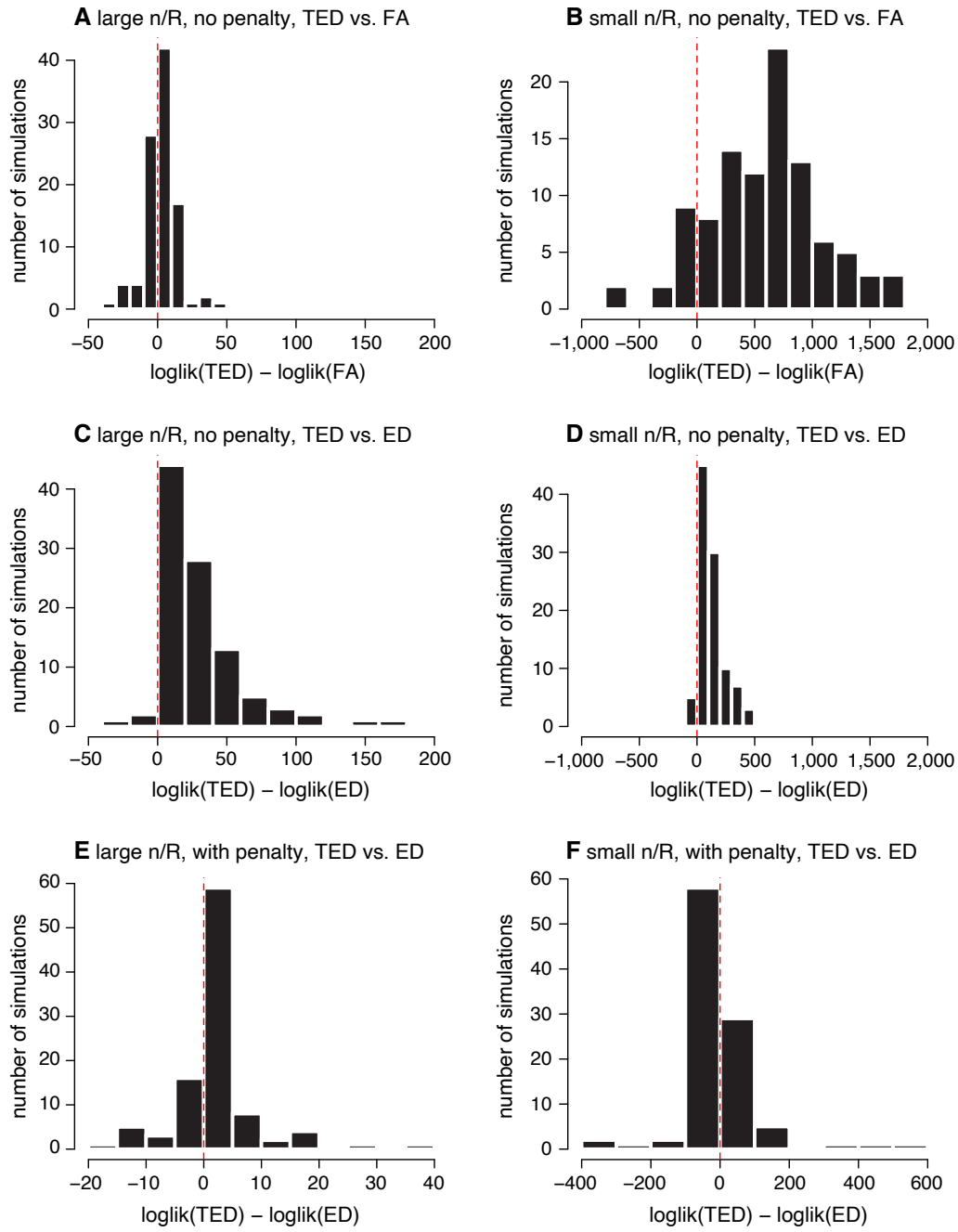

Figure 2.2: Comparison of convergence of TED, ED and FA. Each plot summarizes the results from 100 simulations. In each simulation, log-likelihood achieved after (at most) 2,000 iterations was recorded. Panels E and F show differences in the penalized log-likelihoods (IW penalty, $\lambda = R$).

The results in Figure 2.2 confirm that some of the patterns observed in the illustrative example are true more generally. In the unpenalized case (Panels A–D), TED often arrived

at a better solution than FA or ED, although in the large n/R setting FA was comparable to TED (Panel A). In the penalized case, the TED and ED solutions were much more similar for both large and small n/R (Panels E, F).

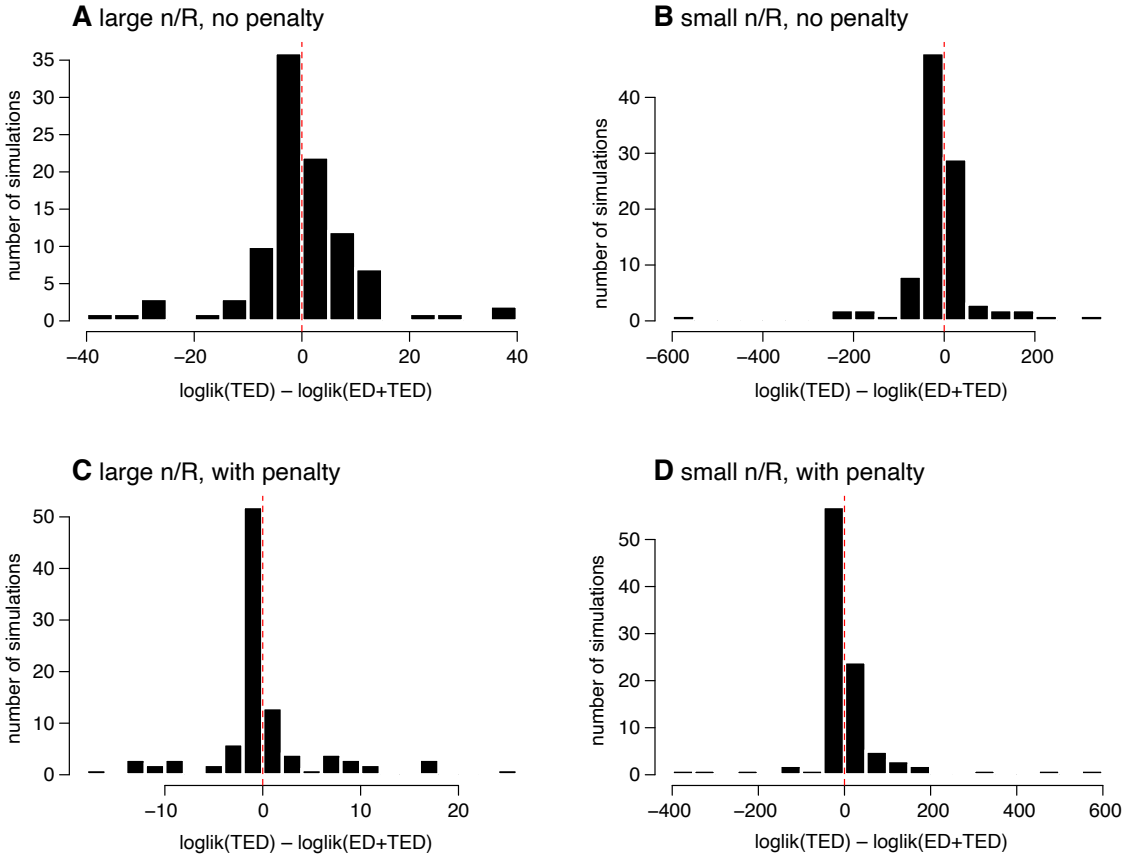

Figure 2.3: Results demonstrating the ability of TED to "rescue" ED. The log-likelihood or penalized log-likelihood obtained by performing at most 2,000 TED updates is compared against performing 2,000 iterations of ED, followed by another round of (at most) 2,000 iterations of TED updates ("ED+TED"). Each plot summarizes the results from 100 simulations, the same simulations as in Figure 2.2.

The improved performance of (unpenalized) TED vs. ED could be due either to faster convergence (e.g., Figure 2.1A) or due to convergence to a better local optimum (e.g., Figure 2.1B). To investigate this, we assessed whether TED can "rescue" ED by running TED initialized to the ED solution ("ED+TED"). If ED converges to a poorer local optimum, then TED will not rescue it, and ED+TED will be similar to ED; on the other hand, if ED

is simply slow to converge, then ED+TED will be similar to TED. The results in Figure 2.3 show that TED usually rescues ED; the ED+TED estimates were consistently very similar to the TED estimates, regardless of whether a penalty was used or not. This suggests that the improved performance of TED is generally due to faster convergence.

To assess how additional iterations improve ED's performance, we reran ED for 100,000 iterations instead of only 2,000. Since these runs take a long time, we examined only a few simulated data sets randomly selected from the large n/R and small n/R scenarios (Supplementary Figure A.1). Even after 100,000 iterations, ED fell measurably short of 1,000 updates of TED (average difference of 40.6 log-likelihood units).

In summary, our experiments confirm that, for homoskedastic errors (which is the setting where all three methods apply), TED exhibited the best performance. This is not unexpected since TED solves the subproblem (eq. 2.25 or 2.27) exactly whereas ED and FA do not. Our experiments also show that including a penalty can help improve convergence behavior, especially for ED. Thus, including the penalty has computational benefits in addition to statistical benefits demonstrated below.

Comparison of the penalties and constraints

Next we evaluated the benefits of different penalties and constraints for estimation and significance testing of the underlying means $\boldsymbol{\theta}_j$. We focused on TED and ED with and without penalties, and on fitting rank-1 matrices using TED. (Since TED solves the subproblem exactly, rather than iteratively, TED should be better than FA in this setting with homoskedastic variances. The main benefit of FA is that it can fit rank-1 matrices with heteroskedastic variances.)

We compared methods using the following evaluation measures:

- **Plots of power vs. false sign rate (FSR).** These plots are similar to the more commonly-used plots of "power vs. false discovery rate," but improve robustness and

generality by requiring "true discoveries" to have the correct sign. The better methods are those that achieve higher power at a given FSR. See the Supplementary Section A.8 for definitions.

- **Empirical False Sign Rate (FSR).** We report the empirical FSR among tests that were significant at $lfsr < 0.05$. A well-behaved method should have a small empirical FSR, certainly smaller than 0.05. We consider an FSR exceeding 0.05 to be indicative of a poorly behaved method.

- **Accuracy of predictive distribution.** To assess generalizability of the estimates of $\mathcal{U}$ and $\boldsymbol{\pi}$, we compared the marginal predictive density (2.3) in test samples, $p(\boldsymbol{x}_j^{\text{test}} \mid \hat{\mathcal{U}}, \hat{\boldsymbol{\pi}}, \boldsymbol{V}_j^{\text{test}})$, against the "ground-truth" marginal predictive density $p(\boldsymbol{x}_j^{\text{test}} \mid \mathcal{U}^{\text{true}}, \boldsymbol{\pi}^{\text{true}}, \boldsymbol{V}_j^{\text{test}})$, where $\mathcal{U}^{\text{true}}, \boldsymbol{\pi}^{\text{true}}$ denote the parameters used to simulate the data. We summarized the relative accuracy in the predictions as

$$\frac{1}{n_{\text{test}}} \sum_{j=1}^{n_{\text{test}}} \log \left\{ \frac{p(\boldsymbol{x}_j^{\text{test}} \mid \mathcal{U}^{\text{true}}, \boldsymbol{\pi}^{\text{true}})}{p(\boldsymbol{x}_j^{\text{test}} \mid \hat{\mathcal{U}}, \hat{\boldsymbol{\pi}})} \right\}. \tag{2.29}$$

  This measure can be interpreted as (an approximation of) the Kullback-Leibler (K-L) divergence from the true predictive distribution to the estimated predictive distribution. Smaller K-L divergences are better.

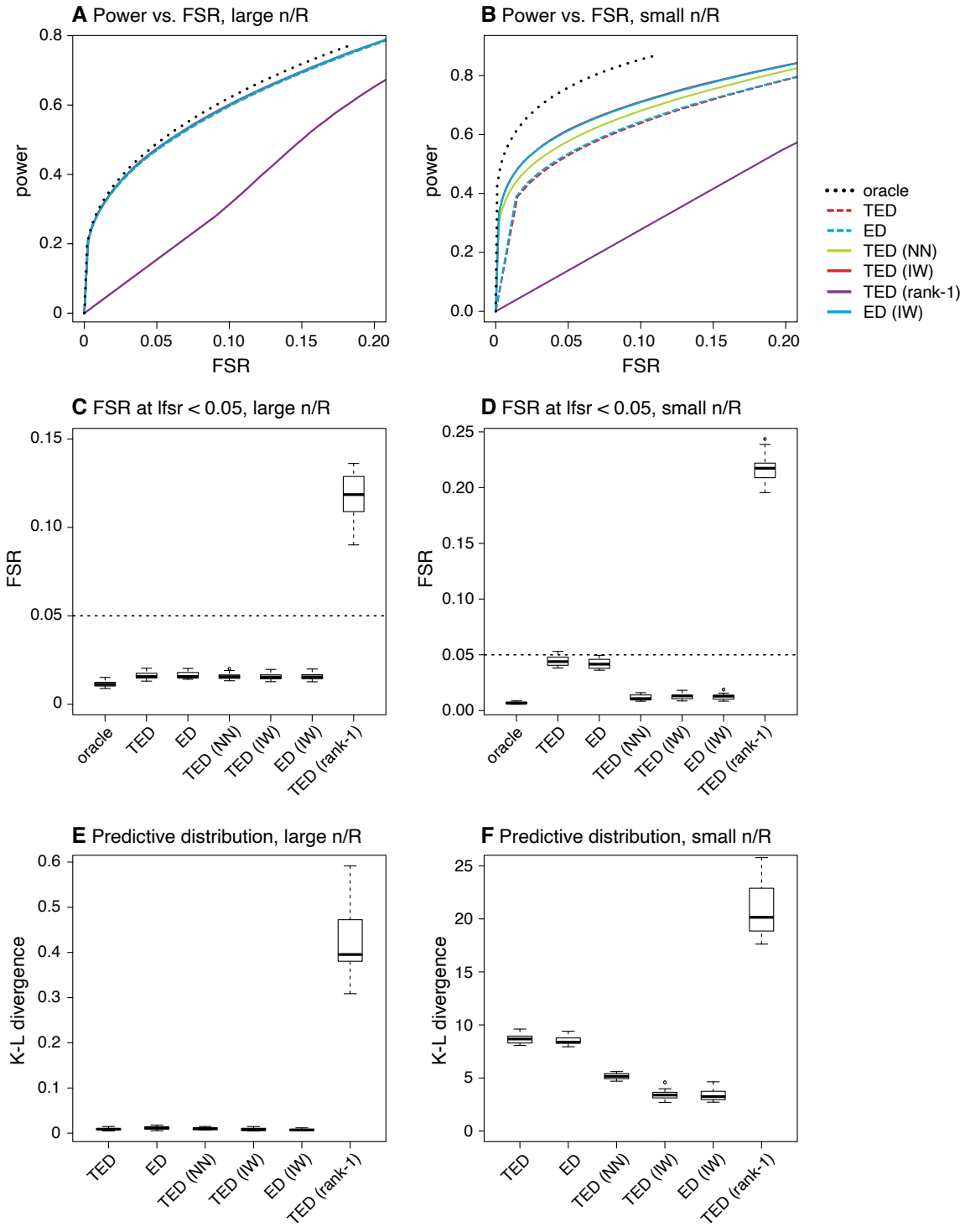

Figure 2.4: Comparison of penalties, constraints and updates (ED vs. TED) in the "hybrid" simulated data sets. For the IW and NN penalties, the penalty strength was set to $\lambda = R$. In C and D, the target FSR is shown as a dotted horizontal line at 0.05. In A, most of the methods are not visible because the lines overlap at the top near the oracle result. Note that the oracle model always achieves a K-L divergence of zero.

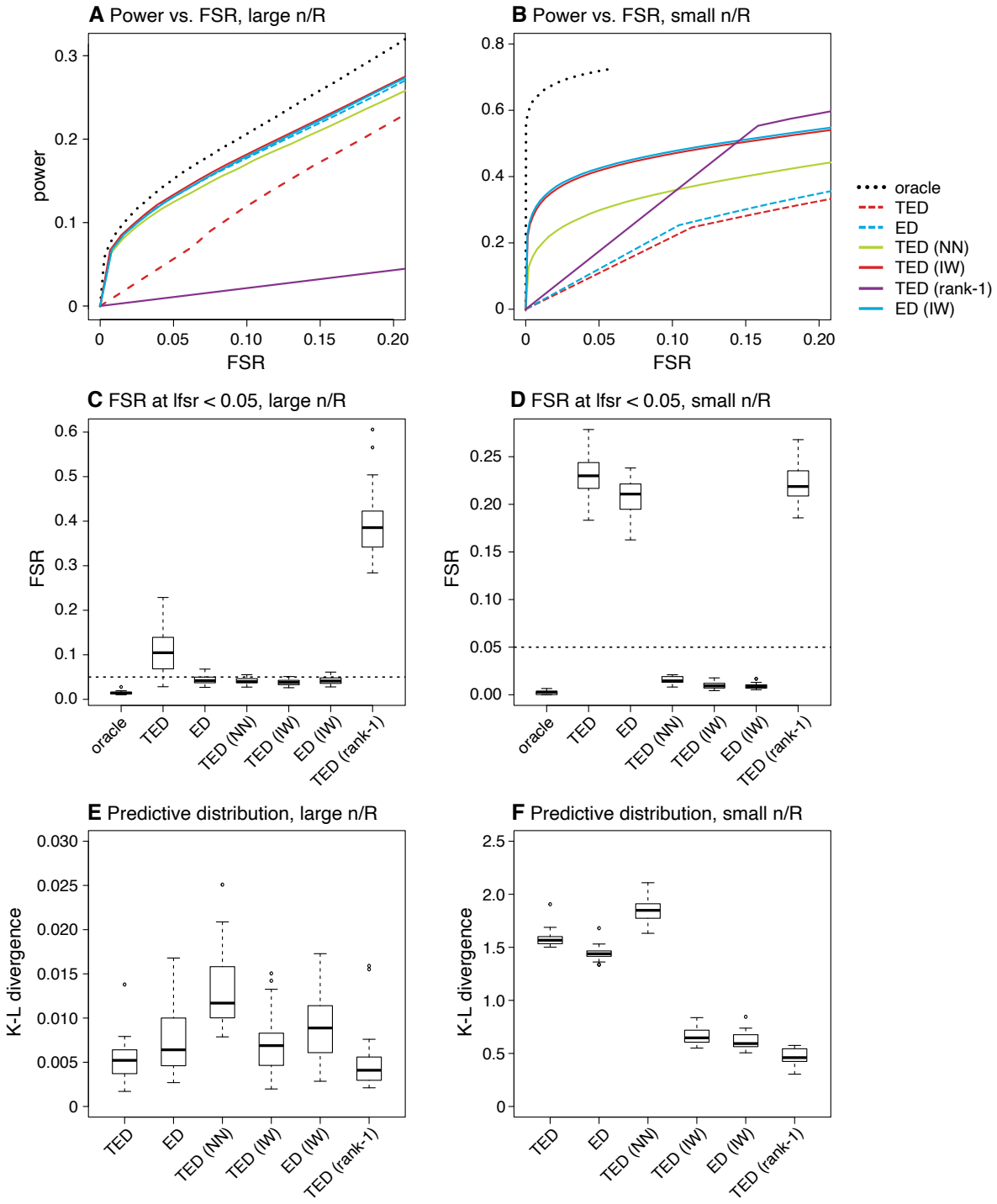

Figure 2.5: Comparison of penalties, constraints and updates (ED vs. TED) in the "rank-1" simulated data sets. For the IW and NN penalties, the penalty strength was set to $\lambda = R$. In C and D, the target FSR is shown as a dotted horizontal line at 0.05. Note that the oracle model always achieves a K-L divergence of zero.

We simulated 20 data sets with large n/R and another 20 data sets with small n/R under both the hybrid and rank-1 scenarios (80 data sets in total). In all cases, model fitting was performed as above, again with $K = 10$. The results are summarized in Figures 2.4 and 2.5. In all comparisons, we included results from the "oracle" EBMNM model—that is, the model used to simulate the data—as a point of reference.

Results for the hybrid setting are shown in Figure 2.4. The results show a clear benefit of using a penalty in the small n/R setting: both IW and NN penalties improved the power vs. FSR and the accuracy of predictive distributions. For large n/R data sets, the penalties do not provide a clear benefit, but also do not hurt performance. In both cases TED and ED perform similarly, suggesting that the poor convergence of ED observed in previous comparisons may have less impact on performance than might have been expected. The rank-1 constraints performed very poorly in all tasks, which is perhaps unsurprising since the true covariances were (mostly) not rank-1.

Results for the rank-1 scenario are shown in Figure 2.5. In this case, imposing rank-1 constraints on the covariance matrices improved predictive performance—which makes sense because the true covariances were indeed rank-1—but produced worse performance in other metrics. In particular, the *lfsr* values from the rank-1 constraint are very poorly calibrated. This is because, as noted in Liu et al. [2023], the rank-1 constraint leads to *lfsr* values that do not differ across conditions; see the Discussion (Section 2.7) for more on this. Penalized estimation of the covariance matrices (using either a IW or NN penalty) consistently achieved the best power at a given FSR in both the large n/R and small n/R settings. Interestingly, (unpenalized) TED performed much worse than (unpenalized) ED in the power vs. FSR for large n/R. We speculate that this was due to the slow convergence of ED providing sort of "implicit regularization". However, explicit regularization via a penalty seems preferable to implicit regularization via a poorly converging algorithm, and overall penalized estimation was the winning (or equally winning) strategy across a variety of settings.

Robustness to mis-specifying the number of mixture components

In the above experiments, we fit all models with a value of K that matched the model used to simulate the data. In practice, however, K is unknown, and so we must also consider situations in which K is mis-specified. Intuitively, one might expect that overstating K may lead to overfitting and worse performance; Urbut et al. [2019] argued however that the use of the mixture components centered at zero in the prior (2.2) makes it robust to overfitting because the mean-zero constraint limits their flexibility. Here we assess this claim.

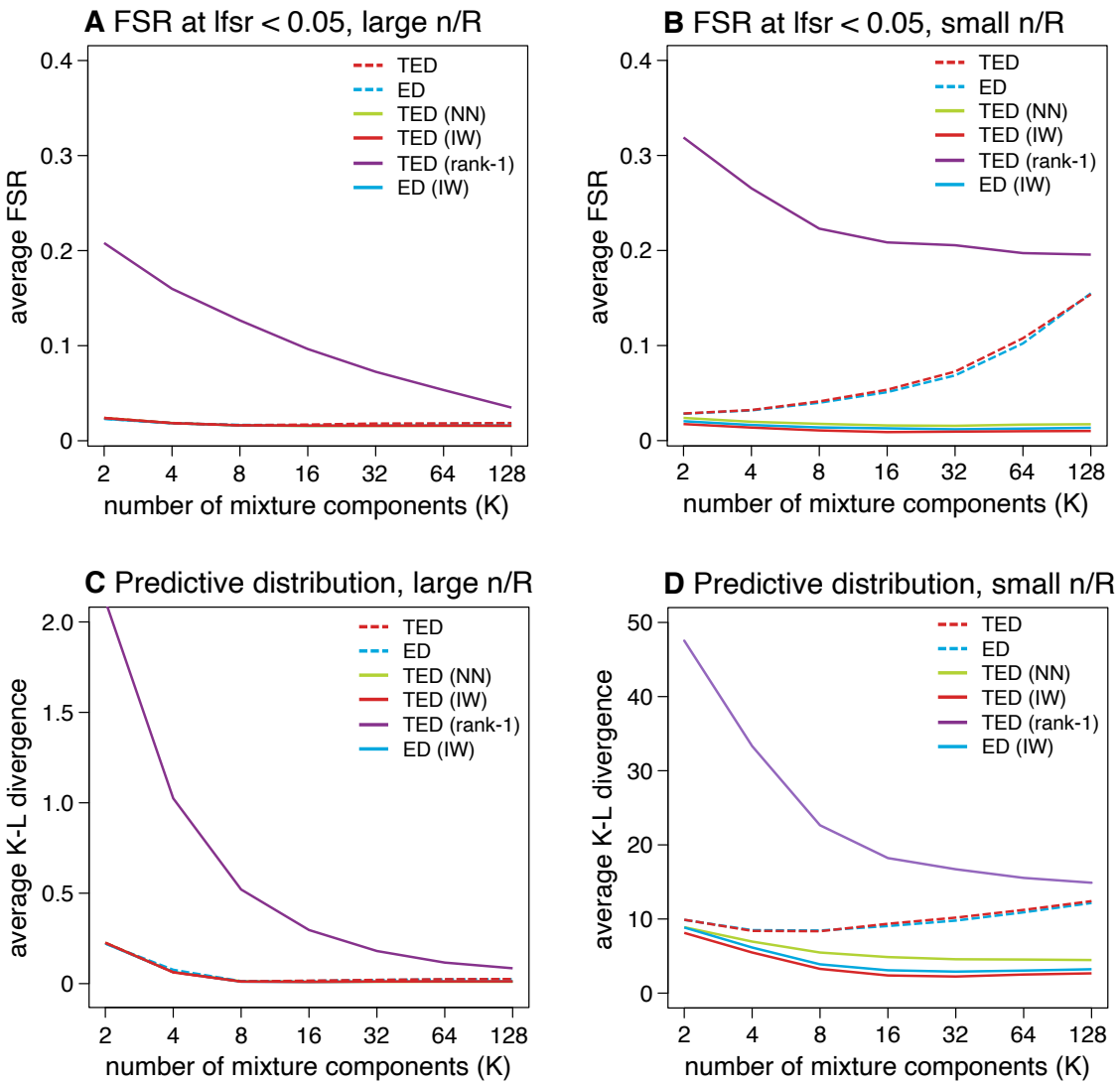

Figure 2.6: Assessment of robustness to mis-specifying K in "hybrid" simulated data sets. All results are averages over 20 data sets, each simulated with $K = 10$ mixture components. In A and C, most of the methods are not visible because they overlap with the "TED-IW" result.

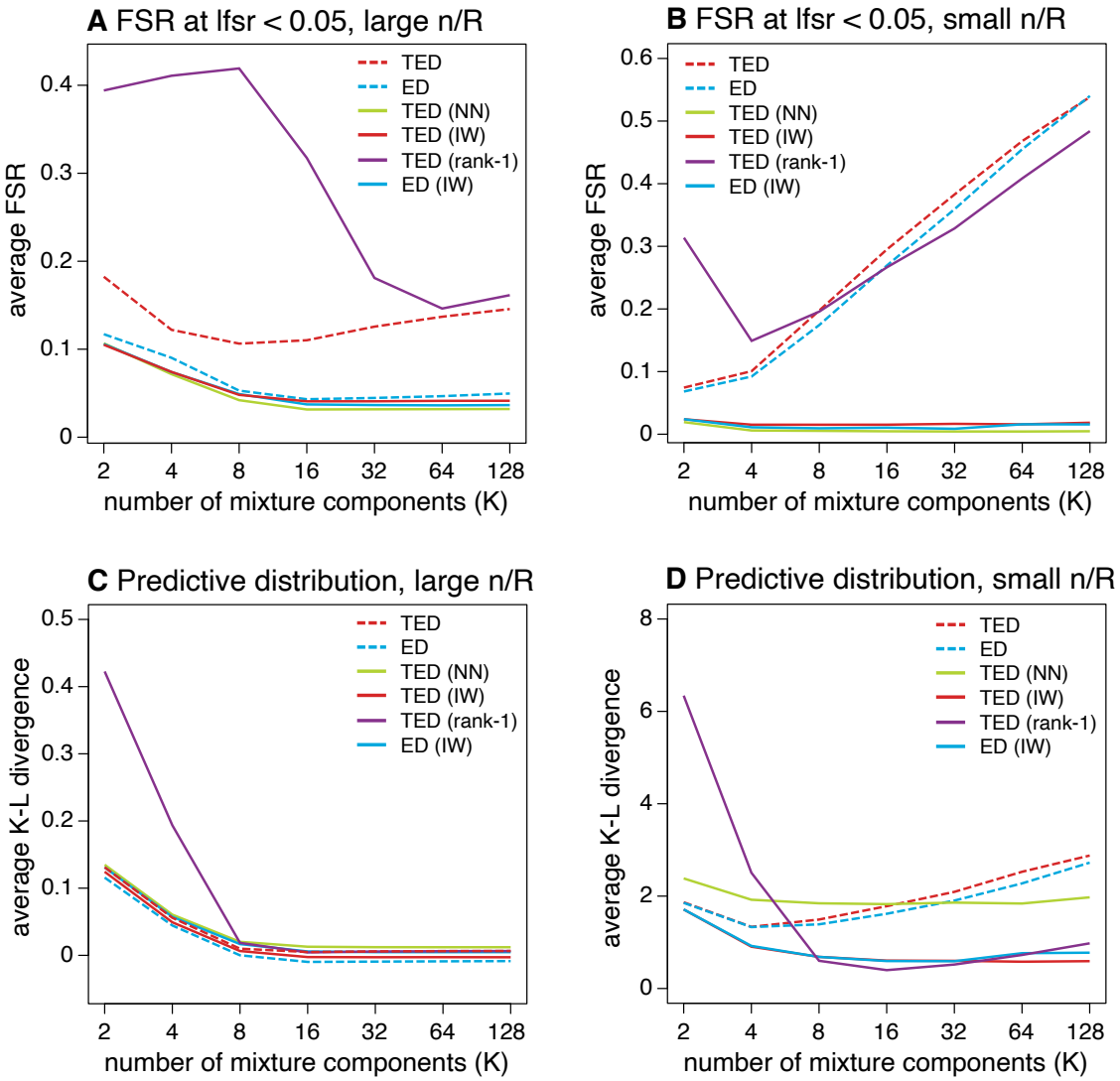

Figure 2.7: Assessment of robustness to mis-specifying K in "rank-1" simulated data sets. All the results shown in the plots are averages over the 20 data sets. All data sets were simulated with $K = 10$ mixture components.

In these experiments, we analyzed the same 80 data sets as in the previous section (simulated with $K = 10$). We fit models with different penalties, constraints and algorithms, with K varying from 2 to 128. We compared results in both the accuracy of the predictive distribution (K-L divergence) and the average FSR at an *lfsr* threshold of 0.05. Results are shown in Figures 2.6 and 2.7. For large n/R, all model fits except those with the rank-1 constraints were robust to overstating K, with similarly good performance even at $K = 128$.

This is generally consistent with the claim in Urbut et al. [2019] that the mixture prior should be robust to overstating K. However, for small n/R the story is quite different: all of the unpenalized algorithms eventually showed a decline in performance when K was too large, presumably due to "overfitting". In comparison, all the penalized methods were more robust to overstating K; the performance did not substantially decline as K increased.

The improved robustness of the penalized methods could be achieved in at least two different ways: they could be using a smaller number of components by estimating some of the mixture weights π_k to be very small; or by estimating some of the components k to have very similar covariances $\boldsymbol{U}_k$ (or both). To investigate these explanations, we looked at the models with $K = 100$ components and recorded the number of "important" components, defined as components k with $\pi_k > 0.01$. We found that the penalized methods tended to produce much fewer "important" components than the unpenalized methods (Supplementary Figure A.4). Essentially, the penalties have the effect of shrinking each $\boldsymbol{U}_k$ toward the identity matrix, so a component is assigned a small weight whenever the identity matrix does not match the truth.

In summary, our results support the use of penalized methods with a large value of K as a simple and robust way to achieve good performance in different settings.

2.6 Analysis of genetic effects on gene expression in 49 human tissues

To illustrate our methods on real data, we used the EBMNM model to analyze the effects of genetic variants on gene expression ("*cis*-eQTLs") in multiple tissues. The use of the EBMNM model for this purpose was first demonstrated in Urbut et al. [2019]. They used a two-stage procedure to fit the EBMNM models: (i) fit the EBMNM model to a subset of "strong" eQTLs to estimate the prior covariance matrices; and (ii) fit a (modified) EBMNM model to all eQTLs—or a random subset of eQTLs—using the covariance matrices from (i).

The model from (ii) was then used to perform inferences and to test for eQTLs in each tissue. Our new methods are relevant to (i), and so we focus on stage (i) here. For (i), Urbut et al. [2019] used the ED algorithm without a penalty. Recognizing that the results are sensitive to initialization, they described a detailed initialization procedure.[1] We coded an initialization procedure similar to this, which we refer to as the "specialized initialization". We note that the specialized initialization adds substantially to both the complexity and computation time of the overall fitting procedure. Our goal was to assess the benefits of different EBMNM analysis pipelines which consisted of combinations of: TED vs. ED updates; specialized initialization vs. a simple random initialization; and penalty vs. no penalty (maximum-likelihood). All these combinations resulted in 8 different analyses of multi-tissue cis-eQTL data.

We analyzed z-scores from tests for association between gene expression in dozen human tissues and genotypes at thousands of genetic variants. The z-scores came from running the "Matrix eQTL" software [Shabalin, 2012] on genotype and gene expression data in release 8 of the Genotype-Tissue Expression (GTEx) Project [GTEx Consortium, 2020]. Following Urbut et al. [2019], we selected the genetic variants with the largest z-score (in magnitude) across tissues for each gene. After data filtering steps, we ended up with a data set of z-scores for $n = 15{,}636$ genes and $R = 49$ tissues. We fit the EBMNM model to these data, with all the $\boldsymbol{V}_j$ set to a common correlation matrix; that is, $\boldsymbol{V}_j = \boldsymbol{C}$, where $\boldsymbol{C}$ is a correlation matrix of non-genetic effects on expression. This correlation matrix $\boldsymbol{C}$ was estimated from the association test z-scores following the approach described in Urbut et al. [2019]. In all runs, we set $K = 40$ to match the number of covariance matrices produced by our specialized initialization. And, in all cases, we initialized the mixture weights to $\pi_k = 1/40$ and the scaling factors (when needed) to $s_k = 1$.

1. An updated version of the initialization procedure of Urbut et al. [2019] is described in the "flash_mash" vignette included in the mashr R package. See also `https://stephenslab.github.io/mashr/articles/flash_mash.html`.

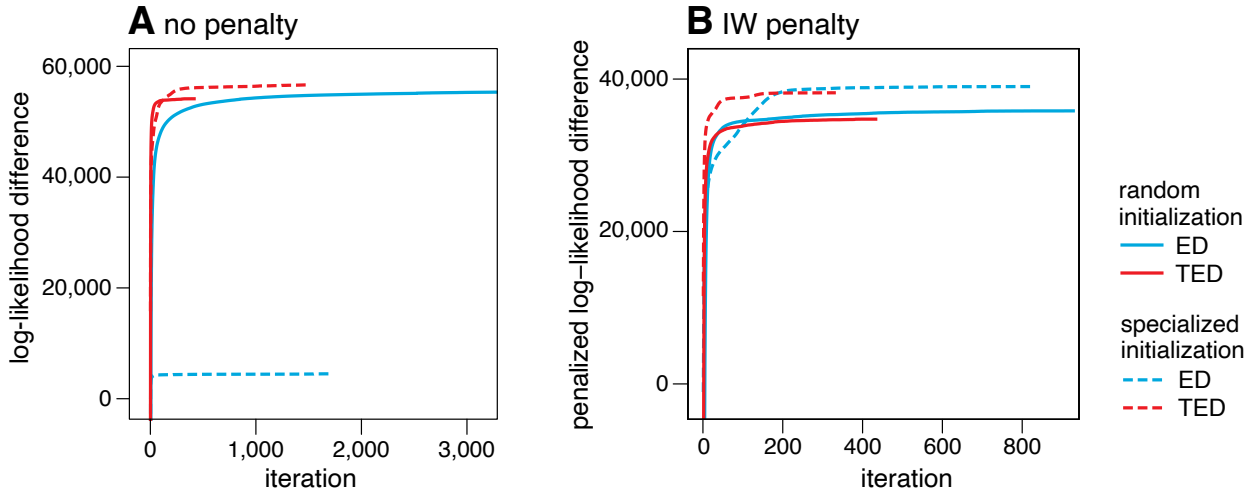

Figure 2.8: Plots showing improvement in model fit over time for the GTEx data, using different initialization schemes, different prior covariance matrix updates, and penalty vs. no penalty (maximum-likelihood). Log-likelihood differences and penalized log-likelihood differences are with respect to the (penalized) log-likelihood near the initial estimate. All models were fit with $K = 40$ mixture components. In B, the inverse wishart (IW) penalty was used with penalty parameter $\lambda = R$. In all cases, the model fitting was halted when the difference in the log-likelihood between two successive updates was less than 0.01, or when 5,000 updates were performed, whichever came first.

Figure 2.8 shows the improvement of the model fits over time in the 8 different analyses. In the analyses without a penalty (A), the TED updates with a specialized initialization produced the best fit, while the ED and TED updates with random initialization were somewhat worse (e.g., TED with random initialization produced a fit that was 2,497.78 log-likelihood units worse, or 0.16 log-likelihood units per gene). Strikingly, the ED updates with specialized initialization resulted in a much worse fit. We attribute this to the fact that the specialized initialization includes many rank-1 matrices, and the "subspace preserving" property of the ED updates means that these matrices are fixed at their initialization (they changed only by a scaling factor), which substantially limits their ability to adapt to the data. In the penalized case, consistent with our simulation results, there was less difference between ED vs. TED. In both cases, the specialized initialization improved the fit relative to random initialization (e.g., TED increased the penalized log-likelihood by 3,176, or about

0.2 per gene). Note that adding a penalty function makes the subspace preserving property of the ED updates irrelevant by forcing the matrices to be full rank.

To compare the quality of the fits obtained by each method, we computed log-likelihoods on held-out ("test set") data, using a 5-fold cross-validation (CV) design. Following the usual CV setup, in each CV fold 80% of the genes were in the training set, and the remaining 20% were in the test set. Then we fit an EBMNM model to the training set following each of the 8 approaches described above, and measured the quality of the fit by computing the log-likelihood in the test set. We also recorded the number of iterations. Within a given fold, the number of components K was the same across all the analyses, and was set depending on the number of covariance matrices produced by the specialized initialization. (K was at least 32 and at most 39.)

Consistent with the simulations, the inclusion of a penalty consistently improved the test set log-likelihood (Table 2.2). With a penalty, the specialized initialization also improved the test set log-likelihood compared with a random initialization. Of the 8 approaches tried, the one used by Urbut et al. [2019]—ED with no penalty and specialized initialization—resulted in the worst test-set likelihood.

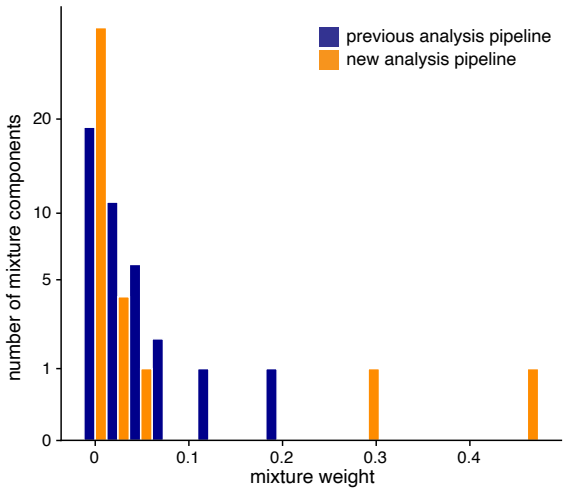

Figure 2.9: Comparison of the prior mixture weights $\boldsymbol{\pi}$ from the previous pipeline vs. the new pipeline. The histogram shows the distributions of the $K40$ prior mixture weights resulting from both pipelines.

initialization	algorithm	penalty	mean relative log-likelihood	average number of iterations
specialized	ED	none	0.00	1,101
specialized	ED	IW	1.21	1,083
specialized	TED	none	0.88	1,054
specialized	TED	IW	1.19	412
random	ED	none	0.25	5,000
random	ED	IW	0.86	1,377
random	TED	none	0.20	450
random	TED	IW	0.94	584

Table 2.2: Cross-validation results on the GTEx data. The "mean relative log-likelihood" column gives the increase in the test-set log-likelihood over the worst log-likelihood among the 8 approaches compared, divided by total number of genes in each test set. The "average number of iterations" column gives the number of iterations performed until the stopping criterion is met (log-likelihood between two successive updates less than 0.01, up to a maximum of 5,000 iterations), averaged over the 5 CV folds.

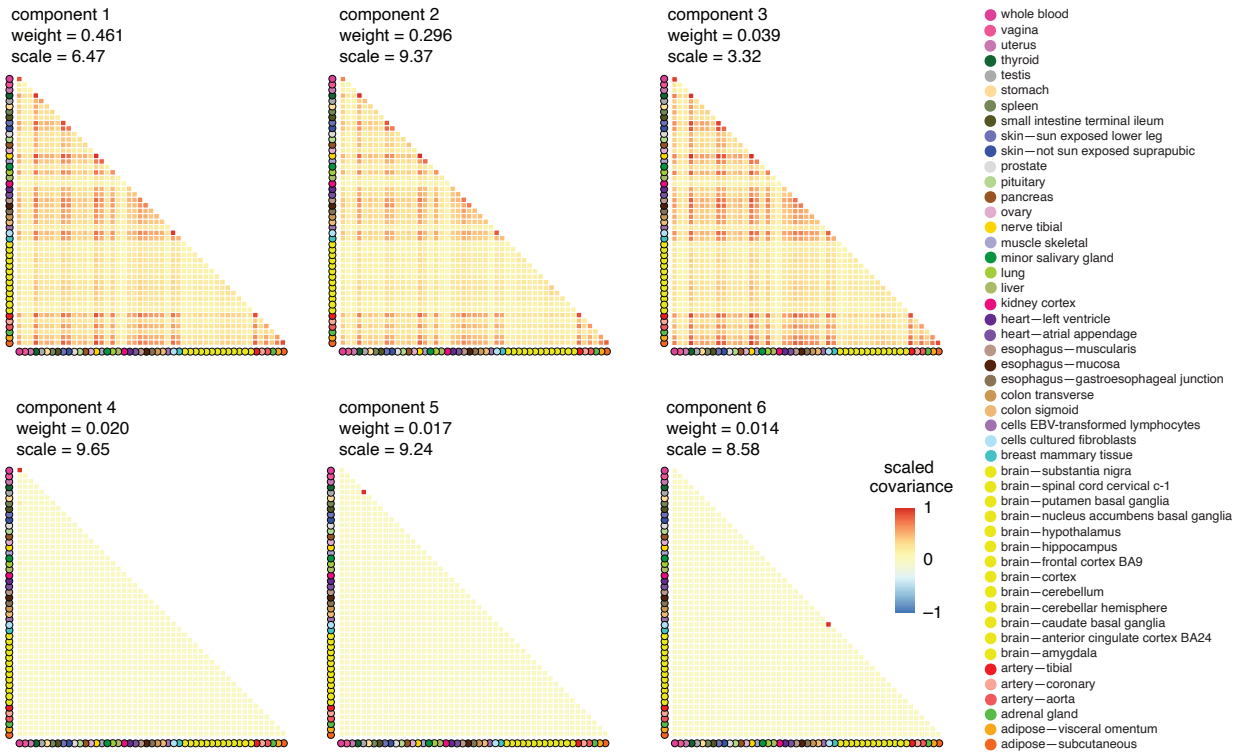

Figure 2.10: Top effect-sharing patterns in the EBMNM model fit to the GTEx data using the previous pipeline. The "top" patterns are the mixture components with the largest weights. Each heatmap shows the 49×49 the scaled covariance matrix $\boldsymbol{U}_k/\sigma_k^2$, where σ_k^2 is the largest diagonal element of $\boldsymbol{U}_k$, so that all elements of the scaled covariance lie between -1 and 1. (The vast majority of the covariances are positive; negative correlations are unexpected.) The scaled covariance matrices are arranged in decreasing order by mixture weight. The "scale" above each heatmap is σ_k. Note that the top three covariance matrices capture broadly similar effect-sharing patterns, but different effect scales.

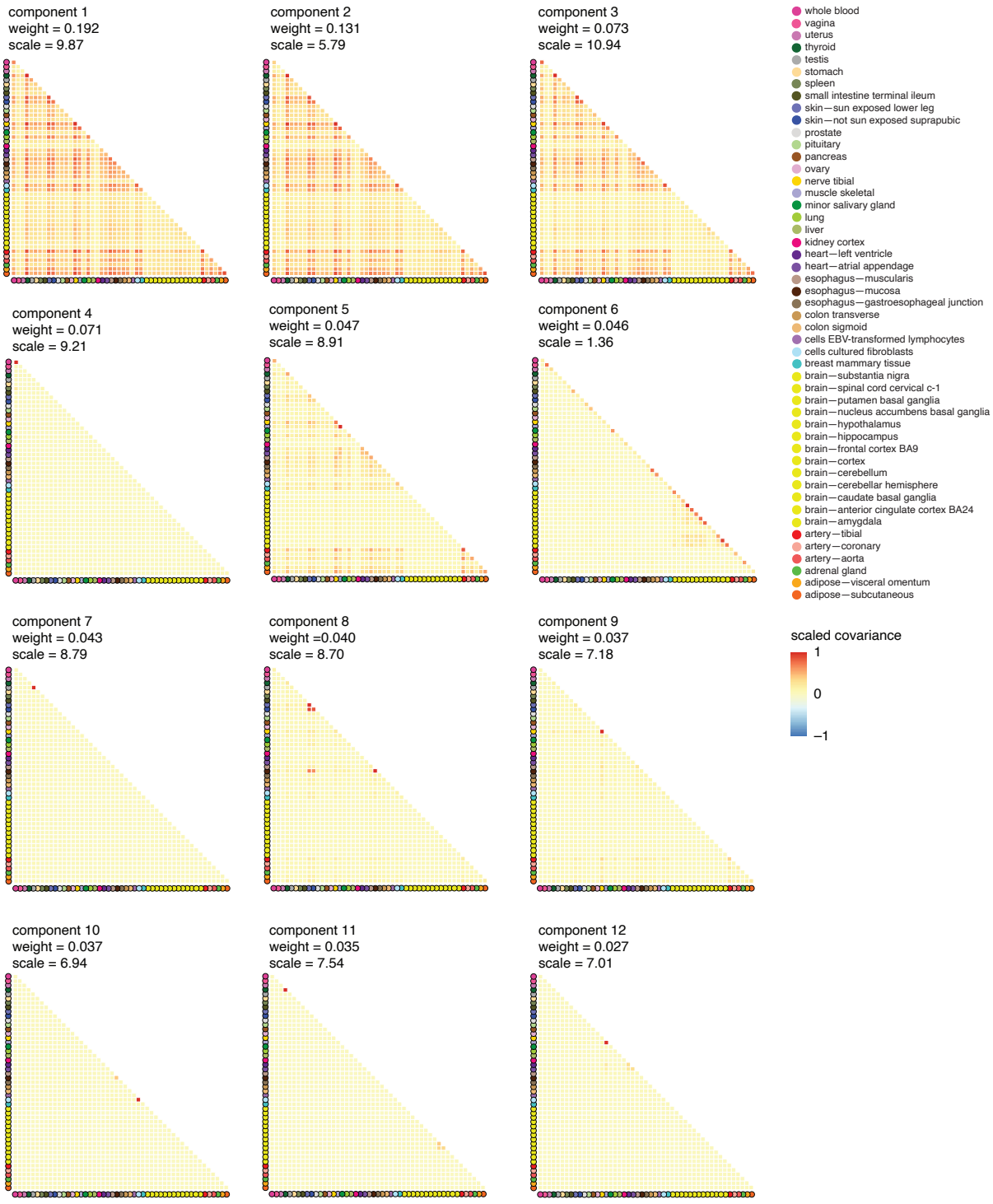

Figure 2.11: Top effect-sharing patterns in the GTEx data generated by the new analysis pipeline. See the caption to Fig. 2.10 for more details.

Although the analyses with a specialized initialization resulted in better fits than the

analyses with a random initialization, the specialized initialization also had substantial computational overhead; running the initialization procedures on these data took more than 1 hour (by comparison, running a single iteration of the EBMNM algorithm typically took about 1 second). Therefore, on balance, one might prefer to dispense with the specialized initialization. Based on these results and considerations, we subsequently examined in more detail the analysis with ED, no penalty and specialized initialization—which was the approach used in Urbut et al. [2019]—and the analysis with TED, IW penalty and random initialization. For brevity, we refer to these analyses as the "previous pipeline" and "new pipeline", respectively.

A notable outcome of the new pipeline is that it produced a prior with weights that were more evenly distributed across the mixture components (Figure 2.9). For example, the new pipeline produced 22 covariances with weights greater than 1%, whereas the previous pipeline produced only 10 covariances with weights more than 1%. Also, the top 16 covariances accounted for 85% of the total weight in the new pipeline, whereas only 6 covariances were needed to equal 85% total weight in the previous pipeline. Inspecting the individual covariances generated from the two analysis pipelines, there are many strong similarities in the estimated sharing patterns (Figures 2.10 and 2.11). But the new analysis pipeline learned a greater variety of tissue-specific patterns (e.g., whole blood, testis, thyroid) and tissue-sharing patterns, many of which appear to reflect underlying tissue biology. For example, sharing pattern 6 (see Figure 2.11) captures sharing of brain-specific effects (including the pituitary gland, which is found at the base of the brain near the hypothalamus). Sharing pattern 8 may reflect the fact that skin and the mucosa layer of the esophagus wall both contain squamous epithelial cells. Additionally, we looked at more closely at the 8 genes with very strong posterior weights (>98%) on component 6, the component capturing eQTL sharing among brain tissues: *HOMER1*, *JPT1*, *SRSF2*, *ABCA1*, *GPATCH8*, *SYNGAP1*, *SPI1* and *LGMN*. Several of these have been linked to neurological and neuropsychiatric

conditions, including major depression, schizophrenia, attention dementia and Alzheimer's disease. For example, *SYNGAP1* is linked to neuronal functions and psychiatric diseases based on results from the GWAS Catalog [Sollis et al., 2022] and from functional studies [Jeyabalan and Clement, 2016, Llamosas et al., 2020]. In summary, the improvements to the EBMNM analysis pipeline should result in the discovery of a greater variety of cross-tissue genetic effects on gene expression.

2.7 Discussion

The growing interest in deciphering shared underlying biological mechanisms has led to a surge in multivariate analyses in genomics; among the many recent examples are multi-trait analyses [Wu et al., 2020, Luo et al., 2020] and multi-ancestry analyses to improve polygenic risk scores [Zhang et al., 2023]. The EBMNM approach described here and in Urbut et al. [2019] provide a versatile and robust multivariate approach to multivariate analysis. We have implemented and compared several algorithms for this problem. These algorithms not only enhance accuracy but also provide a level of flexibility not achieved by other methods.

One of our important findings is that using low-rank covariance matrices in this setting, as was done in Urbut et al. [2019], is not recommended. In particular, while low-rank matrices may be relatively easy to interpret, they lead to poorly calibrated *lfsr* values (e.g., Figure 2.5). For intuition, consider fitting a EBMNM model with $K = 1$ and a rank-1 covariance matrix, $\boldsymbol{U}_1 = \boldsymbol{u}\boldsymbol{u}^T$. (We credit Dongyue Xie for this example.) Under this model, the mean is $\boldsymbol{\theta}_j = \boldsymbol{u}a_j$ for some a_j, and therefore, given a $\boldsymbol{u}$, the signs of all the elements of $\boldsymbol{\theta}_j$ are fully determined by a_j. As a result, all elements of $\boldsymbol{\theta}_j$ will have the same *lfsr*, and so this model cannot capture situations where one is confident in the sign of some elements of $\boldsymbol{\theta}_j$ but not others. This can cause problems *even if the model is correct*, that is, when the true covariances are rank-1, as in Figure 2.5. There are some possible ways to address these issues, say, by imposing sparsity on estimates of $\boldsymbol{u}$, and this could be an area for future work.

Even when the pitfalls of low-rank matrices are avoided, it is still the case that EB methods tend to understate uncertainty compared to "fully Bayesian" methods (e.g., Morris 1983, Wang and Titterington 2005). As a result, one should expect that the estimated *lfsr* values may be anti-conservative; that is, the *lfsr* values are smaller than they should be. Indeed, we saw this anti-conservative behavior across many of our simulations and methods (Supplementary Figures A.2 and A.3). For this reason estimated *lfsr* rates should be treated with caution, and it would be prudent to use more stringent significance thresholds than are actually desired (e.g., an *lfsr* threshold of 0.01 rather than 0.05). In the special case where $\boldsymbol{V} = \boldsymbol{I}$, it should be possible to improve calibration of significance tests by using ideas from Lei and Fithian [2018]. Improving calibration in the general case of dependent multivariate tests seems to be an important area for future research.

CHAPTER 3

COXPH-SUSIE: BAYESIAN VARIABLE SELECTION METHOD FOR SURVIVAL DATA

Abstract

Genome-wide survival analysis of time-to-event (TTE) phenotypes, such as disease onset and progression, have led to the discovery of novel genetic loci that are missed by traditional case-control approaches [Bi et al., 2020]. However, fine-mapping to identify causal variants remains challenging due to high linkage disequilibrium (LD) among genetic variants. To address this, we introduce CoxPH-SuSiE, a novel Bayesian variable selection in regression method that extends the "Sum of Single Effects" (SuSiE) regression [Wang et al., 2020] to the Cox proportional hazards (CoxPH) model.

CoxPH-SuSiE retains the computational efficiency of SuSiE and provides credible sets (CSs) of causal variants. We benchmarked CoxPH-SuSiE against other Bayesian variable selection methods for survival models using simulated data and demonstrated its superior performance in identifying causal variants under complex LD structures. We further applied CoxPH-SuSiE to analyze self-reports of asthma in the UK Biobank. Our results demonstrate that CoxPH-SuSiE offers a robust and efficient solution for fine-mapping genetic variants in TTE phenotypes, providing valuable insights into the genetics of disease progression.

3.1 Introduction

With the increasing availability of biobanks and electronic health records, analyzing time-to-event (TTE) phenotypes—such as disease age of onset, progression and lifespan—has become more common in genetics. TTE data provides critical insights into the genetics of disease development and progression, thereby enhancing our understanding of disease etiology and

guiding intervention planning. Research has shown that modeling TTE phenotypes using survival models can be more powerful than modeling binary disease occurrence status using logistic models in cohort studies, particularly for common events [Green and Symons, 1983, Callas et al., 1998, Staley et al., 2017]. Additionally, genome-wide survival association studies (GWAS) have identified several significant loci in Essential hypertension, Osteoarthrosis, Asthma, Cataract, Coronary atherosclerosis and Type 2 diabetes that are not significant based on case-control status [Bi et al., 2020].

Despite the detection of new loci in GWAS of TTE phenotypes, narrowing down to potential causal variants is crucial to understand the genetic causes of diseases. This step is usually achieved through fine-mapping. Fine-mapping is a challenging problem due to the existence of strong and complex correlation patterns ("linkage disequilibrium", or LD) among nearby genetic variants. Studies often include many pairs of genetic variants with sample correlations larger than 0.99, or even equaling 1. The most successful fine-mapping approaches treat it as a variable selection problem based on regression models, where genetic variants are the candidate predictors [Sillanpaa and Bhattacharjee, 2005].

Bayesian variable selection in regression (BVSR) is an appealing approach to fine-mapping as it can provide uncertainty quantification of which variables to select. And several BVSR methods have been developed for survival models to handle high-dimensional genomics data with numerous covariates, often in the thousands. Newcombe et al. [2017] presented a sparse Bayesian Weibull regression method using a normal prior for non-zero effect variables and implemented a reversible jump MCMC algorithm. Other methods focus on the Cox proportional hazards (CoxPH) regression model [Cox, 1972], the most widely used survival model. Nikooienejad et al. [2020] proposed a Bayesian method for variable selection within a CoxPH regression model, utilizing a nonlocal prior for non-zero coefficients and implemented a stochastic search algorithm for computation. Later, Komodromos et al. [2022] introduced a sparse variational Bayes approach based on CoxPH model with a Laplace prior for non-zero

effect variables. They used mean-field approximation and implemented a coordinate-ascent algorithm to solve the model. However, these authors only applied their methods to datasets with moderate covariate correlations (around 0.6 to 0.8), leaving it unclear how those methods would perform in the fine-mapping setting with many pairs of correlations nearly 1.

A widely-used fine-mapping method for quantitative traits is the "Sum of Single Effects" (SuSiE) regression by Wang et al. [2020]. SuSiE presented a new formulation of BVSR, resulting in a simple and fast model fitting procedure. It employs a normal prior for non-zero coefficients and utilizes a coordinate ascent algorithm for optimization. Additionally, it offers a straightforward way to calculate "Credible Sets" (CSs) of putative causal variants, where each CS is a subset of variants that includes at least one causal variant with a specified probability. Given SuSiE's success in fine-mapping, we extend its framework to CoxPH model, which we refer to as CoxPH-SuSiE. CoxPH-SuSiE uses the same parameterization for covariates and a similar model fitting procedure as SuSiE. Therefore, CoxPH-SuSiE inherits the advantages of SuSiE.

In this chapter, we describe how CoxPH-SuSiE works in detail. We also compare CoxPH-SuSiE with other existing survival BVSR methods in the context of fine-mapping. Section 3.2.3 and 3.3 provide brief background for SuSiE and survival analysis. Section 3.4-3.6 describe how CoxPH-SuSiE works in details. Section 3.7 compares the performance of CoxPH-SuSiE with other existing BVSR methods for survival data. Section 3.8 applies CoxPH-SuSiE to real data, the self-reports of asthma in the UK Biobank. Section 3.9 discusses the advantages and limitations of CoxPH-SuSiE, and possible future directions.

3.2 Background: Sum of Single Effect Regression (SuSiE)

3.2.1 Bayesian simple linear regression

We start from the prerequisite, Bayesian simple linear regression model:

$$\boldsymbol{y} = \boldsymbol{x}b + \boldsymbol{e}, \tag{3.1}$$

$$\boldsymbol{e} \sim N_n(0, \sigma^2 \boldsymbol{I}_n), \tag{3.2}$$

$$b \sim N_1(0, \sigma_0^2). \tag{3.3}$$

Here, $\boldsymbol{x}$, $\boldsymbol{y}$ and $\boldsymbol{e}$ are n-vectors containing values of the explanatory variable, the response and the error. b is a scalar regression coefficient. σ^2 is the variance of the error term and σ_0^2 is the prior variance of b. Given σ^2 and σ_0^2, the posterior computations for this model are fairly simple and can be written in the form of the usual least-squares estimate of b, $\hat{b} = (\boldsymbol{x}^T\boldsymbol{x})^{-1}\boldsymbol{x}^T\boldsymbol{y}$, its variance $s^2 := (\boldsymbol{x}^T\boldsymbol{x})^{-1}\sigma^2$, and the corresponding z score, $z := \hat{b}/s$. The posterior distribution for b is:

$$b|\boldsymbol{x}, \boldsymbol{y}, \sigma^2, \sigma_0^2 \sim N_1(\mu_1, \sigma_1^2), \tag{3.4}$$

where

$$\sigma_1^2(\boldsymbol{x}; \sigma^2, \sigma_0^2) := \frac{1}{1/s^2 + 1/\sigma_0^2} \tag{3.5}$$

$$\mu_1(\boldsymbol{x}, \boldsymbol{y}, \sigma^2, \sigma_0^2) := (\sigma_1^2/s^2)\hat{b}. \tag{3.6}$$

In the Bayesian paradigm, we use Bayes Factor (BF) to measure how strong the evidence

is. The BF for comparing this model with the null model ($b = 0$):

$$\mathrm{BF}(\boldsymbol{x}, \boldsymbol{y}|\sigma_0^2, \sigma^2) = \frac{p(\boldsymbol{y}|\boldsymbol{x}, \sigma^2, \sigma_0^2)}{p(\boldsymbol{y}|\boldsymbol{x}, \sigma^2, b = 0)} = \frac{\int p(\boldsymbol{y}|\boldsymbol{x}, \sigma_0^2, \sigma^2, b)p(b)db}{p(\boldsymbol{y}|\sigma^2, b = 0)} \tag{3.7}$$

$$= \sqrt{\frac{s^2}{\sigma_0^2 + s^2}} \exp\Big(\frac{z^2}{2} \times \frac{\sigma_0^2}{\sigma_0^2 + s^2}\Big). \tag{3.8}$$

Note that under Bayesian simple linear regression, the BF expression is an exact one.

3.2.2 Single effect regression (SER)

The "single-effect regression" (SER) model is the building block of SuSiE, which is defined as a multiple regression model with exactly one non-zero effect variable among all. The SER model is also trivial to fit and we will see later that fitting the SER model helps solving the more complicated SuSiE model. The SER model is as follows:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{e}, \tag{3.9}$$

$$\boldsymbol{e} \sim N_n(0, \sigma^2 \boldsymbol{I}_n), \tag{3.10}$$

$$\boldsymbol{b} = b\boldsymbol{\gamma}, \tag{3.11}$$

$$\boldsymbol{\gamma} \sim \mathrm{Mult}(1, \boldsymbol{\pi}), \tag{3.12}$$

$$b \sim N_1(0, \sigma_0^2). \tag{3.13}$$

Again, $\boldsymbol{y}$ is the n-vector of response data. $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)$ is an $n \times p$ matrix storing n observations of p explanatory variables and $\boldsymbol{b}$ is the p-vector of regression coefficients. $\boldsymbol{e}$ is the independent error terms and $\boldsymbol{\gamma} \in \{0, 1\}^p$ is a p-vector of indicator variables. The scalar b represents the "single effect". $\mathrm{Mult}(m, \boldsymbol{\pi})$ denotes the multinomial distribution on class counts that is obtained when m samples are drawn with class probabilities $\boldsymbol{\pi}$. σ^2, σ_0^2 and $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_p)$ are hyperparameters, representing the residual variance, the prior variance of the non-zero effect and prior inclusion probabilities, where π_j indicates the prior

probability that variable j has non-zero effect.

The inference goal here is to find the posterior distribution of $\boldsymbol{b}$, more specifically, $b|\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}$, which were derived in Wang et al. [2020]. Given σ_0^2 and σ^2, the posterior distribution of $b|\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}$ under SER model (3.9)-(3.13) are:

$$\boldsymbol{\gamma}|\boldsymbol{X}, \boldsymbol{y}, \sigma^2, \sigma_0^2 \sim \text{Mult}(1, \boldsymbol{\alpha}), \tag{3.14}$$

$$b|\boldsymbol{X}, \boldsymbol{y}, \sigma^2, \sigma_0^2, \gamma_j = 1 \sim N_1(\mu_{1j}, \sigma_{1j}^2), \tag{3.15}$$

where $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_p)$ is the vector of posterior inclusion probabilities (PIPs), and μ_{1j}, σ_{1j}^2 are the posterior mean and variance of simple linear regression in (3.5)-(3.6) given $\boldsymbol{\gamma}_j = 1$:

$$\boldsymbol{\alpha}_j = P(\boldsymbol{\gamma}_j = 1|\boldsymbol{X}, \boldsymbol{y}, \sigma_0^2, \sigma^2) = \frac{\pi_j \text{BF}_j}{\sum_{j'=1}^{p} \pi_{j'} \text{BF}'_j} \tag{3.16}$$

$$\mu_{1j} = \mu_1(\boldsymbol{x}_j, \boldsymbol{y}; \sigma^2, \sigma_0^2) \tag{3.17}$$

$$\sigma_{1j}^2 = \sigma_1(\boldsymbol{x}_j; \sigma^2, \sigma_0^2). \tag{3.18}$$

Wang et al. [2020] also defined a function that returns the posterior distribution of $\boldsymbol{b}$ under the SER model:

$$\text{SER}(\boldsymbol{X}, \boldsymbol{y}; \sigma^2, \sigma_0^2) := (\boldsymbol{\alpha}, \boldsymbol{\mu}_1, \boldsymbol{\sigma}_1^2), \tag{3.19}$$

where $\boldsymbol{\mu}_1 := (\mu_{11}, \cdots, \mu_{1p})$ and $\boldsymbol{\sigma}_1^2 := (\sigma_{11}^2, \cdots, \sigma_{1p}^2)$.

3.2.3 Sum of single effect regression

The SER model, while straightforward to fit, has quite narrow applicability because of the assumption that there is exactly one nonzero effect. The sum of single effect regression (SuSiE) model can be viewed as a natural extension to SER model, which expresses the

overall effect vector $\boldsymbol{b}$ as a summation of multiple single effect vectors, $\boldsymbol{b} = \sum_{l=1}^{L} \boldsymbol{b}_l$. The SuSiE model is defined as follows:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{e}, \tag{3.20}$$

$$\boldsymbol{e} \sim N_n(0, \sigma^2 \boldsymbol{I}_n), \tag{3.21}$$

$$\boldsymbol{b} = \sum_{l=1}^{L} \boldsymbol{b}_l \tag{3.22}$$

$$\boldsymbol{b}_l = b_l \boldsymbol{\gamma}_l, \tag{3.23}$$

$$\boldsymbol{\gamma}_l \sim \text{Mult}(1, \boldsymbol{\pi}), \tag{3.24}$$

$$b_l \sim N_1(0, \sigma_{0l}^2). \tag{3.25}$$

Here, L is the number of single effect vectors. SuSiE allows the prior variance $\boldsymbol{\sigma}_{0l}^2 = (\sigma_{01}^2, \cdots, \sigma_{0L}^2)$ to be different for each single effect vector. In the special case of $L = 1$, the SuSiE model simplifies to the SER model.

Wang et al. [2020] introduced the Iterative Bayesian Stepwise Selection (IBSS) algorithm to fit the SuSiE model; see Algorithm 2. The intuition is that, given $\boldsymbol{b}_{l' \neq l}$, solving the l^{th} sub-problem corresponds to fitting a SER model. Thus, the IBSS algorithm iteratively updates the posterior distribution for each $\boldsymbol{b}_l$ under the SER model, given the current estimates of the other $\boldsymbol{b}_{l'}, l' \neq l$. The hyper-parameters σ^2 and $\boldsymbol{\sigma}_0^2$ are estimated using an empirical Bayes approach.

Algorithm 2: Iterative Bayesian stepwise selection (IBSS)

Data: $\boldsymbol{X}, \boldsymbol{y}$

Require: Number of effects, L; hyperparameters $\sigma^2, \boldsymbol{\sigma}_0^2$

. **Require:** A function $\text{SER}(\boldsymbol{X}, \boldsymbol{y}, \sigma^2, \sigma_0^2) \to (\boldsymbol{\alpha}, \boldsymbol{\mu}_1, \boldsymbol{\sigma}_1)$ that computes the posterior distribution for $\boldsymbol{b}_l$ under the SER model;

Initialization: posterior means $\bar{\boldsymbol{b}}_l = \boldsymbol{0}$, for $l = 1, \ldots, \boldsymbol{L}$;

repeat

  for $l \in 1, \ldots, \boldsymbol{L}$ **do**

    $\boldsymbol{\theta}_l \leftarrow \boldsymbol{X} \sum_{l' \neq l} \bar{\boldsymbol{b}}_{l'}$

    $(\boldsymbol{\alpha}_l, \boldsymbol{\mu}_{1l}, \boldsymbol{\sigma}_{1l}) \leftarrow \text{SER}(\boldsymbol{X}, \boldsymbol{y} - \boldsymbol{\theta}_l; \sigma_{0l}^2, \sigma^2)$

    $\bar{\boldsymbol{b}}_l \leftarrow \boldsymbol{\alpha}_l \circ \boldsymbol{\mu}_{1l}$ "$\circ$" denotes element-wise multiplication.

  end

until *convergence criterion satisfied*;

Return: $\boldsymbol{\alpha}_1, \boldsymbol{\mu}_{11}, \boldsymbol{\sigma}_{11}, \ldots, \boldsymbol{\alpha}_L, \boldsymbol{\mu}_{1L}, \boldsymbol{\sigma}_{1L}$

Wang et al. [2020] showed that IBSS algorithm is a coordinate ascent algorithm for optimizing a variational approximation (VA) to the posterior distribution for $\boldsymbol{b}_1, \cdots, \boldsymbol{b}_L$ under the SuSiE model. It finds an approximation $q(\boldsymbol{b}_1, \cdots, \boldsymbol{b}_L)$ to the true posterior distribution $p_{\text{post}} = p(\boldsymbol{b}_1, \cdots, \boldsymbol{b}_L | \boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\sigma}_0^2, \sigma^2)$ by minimizing the Kullback-Leibler (KL) divergence from q to p_{post}. The approximation q is designed to factor out as L independent components:

$$q(\boldsymbol{b}_1, \cdots, \boldsymbol{b}_L) = \prod_{l=1}^{L} q_l(\boldsymbol{b}_l). \tag{3.26}$$

Therefore, the L single effect vectors are independent a posteriori. Notably, SuSiE does not require each q_l factorizes over the p elements of $\boldsymbol{b}_l$, so that q_l can capture strong dependencies among the elements of $\boldsymbol{b}_l$ under the assumption that only one element of $\boldsymbol{b}_l$ is non-zero.

SuSiE provides convenient novel summaries of uncertainty in variable selection through Credible Sets, where a level-ρ Credible Set (CS) of variables is defined as follows:

Definition 1. *A level-ρ Credible Set (CS) is a subset of variables that has probability $\geq \rho$ of containing at least one effect variable.*

Given the posterior inclusion probabilities $\boldsymbol{\alpha}$, it's straightforward to construct a CS. First, sort the variants in descending order based on α_j. Then, add variants to the CS until their cumulative probability exceeds ρ. In SuSiE, the L single effect vectors will yield L CSs. SuSiE further prunes the CSs based on "purity", which is defined as the smallest absolute correlation among all pairs of variables within the CS. The rationale is that CSs containing many uncorrelated variables lack inferential value and are therefore disregarded in practice.

3.3 Background: Survival analysis

Survival analysis models time-to-event data, where the event of interest may be death, disease recurrence, or failure. In survival analysis, people are usually interested in estimating the probability of occurrence of the event over time, or understanding the relationship between predictors or risk factors and the time-to-event outcome.

An example setting where such data arises is in prospective cohort studies, where a group of individuals are followed over a certain period of time to determine whether they develop a specific disease. A unique challenge of survival data is that the time-to-event may not be observed for every individual. This phenomenon is called "censoring" and can occur for several reasons: some individuals may not develop the disease by the end of the study, they may be lost to follow-up, or they may drop out of the study. Despite the absence of event times for some individuals, censoring still contains partial information. Specifically, it indicates that these individuals had not developed the outcome by the time of censoring. Therefore, survival analysis incorporates censoring information into the modeling process Clark et al. [2003].

Conventionally, people use T_i to denote the time to event for individual i, and C_i for corresponding censoring time. We either observe the event time or the censoring time,

whichever occurs first, i.e. we observe $Y_i = \min(T_i, C_i)$. We use an indicator variable δ_i to distinguish which occurs first, $\delta_i = \mathbf{1}(T_i \leq C_i)$. To characterize the distribution of survival time T, people make use of the following four functions, which are inter-related:

1. Probability density function:

$$f(t) = \lim_{\Delta \to 0} \frac{1}{\Delta} P(t \leq T < t + \Delta) \tag{3.27}$$

$$= \lambda(t) S(t) = \lambda(t) \exp\left\{ - \int_0^t \lambda(s) ds \right\} \tag{3.28}$$

2. Survival function:

$$S(t) = P(T > t) = 1 - P(T \leq t) \tag{3.29}$$

$$= 1 - F(t) = 1 - \int_0^t f(s) ds, \tag{3.30}$$

   where $F(t)$ is the cumulative distribution function (CDF).

3. Hazard function: the instantaneous rate at time t, given that the event has not occurred prior to time t.

$$\lambda(t) = \lim_{\Delta \to 0} \frac{1}{\Delta} P(t \leq T < t + \Delta | T \geq t) \tag{3.31}$$

$$= \frac{f(t)}{S(t)} = - \frac{\partial}{\partial t} \log S(t) \tag{3.32}$$

4. Cumulative hazard function:

$$\Lambda(t) = \int_0^t \lambda(s) ds = - \log S(t). \tag{3.33}$$

3.3.1 Cox proportional hazards regression

In survival analysis, the relationship between time to event and covariates is usually modeled through the hazard function:

$$\lambda(t|\boldsymbol{x}_i) = \lim_{\Delta \to 0} \frac{1}{\Delta} P(t \leq T < t + \Delta | T \geq t, \boldsymbol{x}_i), \tag{3.34}$$

where $\boldsymbol{x}_i$ is the vector of covariates thought to influence the survival for individual i. The survival regression models can be viewed as consisting of two components: the baseline hazard function, which describes how the risk of event changes over time at baseline levels of covariates; and the effect parameters, which characterize how the hazard varies in response to explanatory covariates, for instance, race, sex and certain treatments.

Cox [1972] proposed a semi-parametric proportional hazards model, commonly known as the Cox Proportional Hazards (CoxPH) model, which allows for an arbitrary baseline hazard function. The proportional hazards (PH) assumption states that covariates are multiplicatively related to the hazard. Specifically, let $\boldsymbol{x}_i$ be a vector of covariate values for individual i, the hazard function for CoxPH model is:

$$\lambda(t|\boldsymbol{x}_i) = \lambda_0(t) \exp\{\boldsymbol{b}^T \boldsymbol{x}_i\}, \tag{3.35}$$

where $\lambda_0(t)$ is the baseline hazard. $\boldsymbol{b}$ is the vector of covariate coefficients, which represents log hazard ratios, i.e. $\exp\{b_j\}$ can be interpreted as comparing the hazard of two individuals whose covariate values are the same except $\boldsymbol{x}_{ij} = l + 1$ and $\boldsymbol{x}_{i'j} = l$,

$$\frac{\lambda(t|\boldsymbol{x}_i)}{\lambda(t|\boldsymbol{x}_{i'})} = \exp\{b_j\}. \tag{3.36}$$

Also note that the CoxPH model (3.44) does not have an intercept term, as the intercept does not depend on individuals' covariates, therefore, it gets absorbed into the baseline hazard function $\lambda_0(t)$. This formulation implies that the proportionality remains constant over time, while the baseline hazard $\lambda_0(t)$ may change. $\lambda_0(t)$ is not parameterized, meaning it does not require a specific distributional assumption for survival time. This makes CoxPH model a very flexible and widely used tool in survival analysis.

Partial likelihood

The primary interest of CoxPH model is to estimate and make inference on the covariate coefficients $\boldsymbol{b}$. The standard maximum likelihood approach won't work since the distribution of survival time T is arbitrary in CoxPH model. Cox [1972] proposed to condition on the occurrence of events and the corresponding event times. This is because intervals without failures provide no information about $\boldsymbol{b}$, as the baseline hazard component $\lambda_0(t)$ could potentially be zero during these intervals.

Suppose the observed data is $\{(y_i, \delta_i, \boldsymbol{x}_i) : i = 1, \cdots, n\}$, containing K events. Let $t_{(1)} < t_{(2)} < \cdots < t_{(K)}$ denote the ordered event times across the n observations and we assume there is no tie in event times for now. Let $\mathcal{R}_k$ be the risk set for the k^{th} event, which contains all individuals who were at risk to experience the event, that is, individuals with $y_i \geq t_{(k)}$. Let $\boldsymbol{x}_{(k)}$ denote the covariate vector for the individual who experienced the event at $t_{(k)}$. For failure time $t_{(k)}$, conditionally on the risk set $\mathcal{R}_k$, the probability that the failure is on the individual as observed is

$$\frac{\lambda_0(t_{(k)}) \exp\{\boldsymbol{b}^T \boldsymbol{x}_{(k)}\}}{\sum_{i \in \mathcal{R}_k} \lambda_0(t_{(k)}) \exp\{\boldsymbol{b}^T \boldsymbol{x}_i\}} = \frac{\exp\{\boldsymbol{b}^T \boldsymbol{x}_{(k)}\}}{\sum_{i \in \mathcal{R}_k} \exp\{\boldsymbol{b}^T \boldsymbol{x}_i\}}. \tag{3.37}$$

The partial likelihood aggregates the conditional probabilities across K risk sets:

$$L_p(\boldsymbol{b}) = \prod_{k=1}^{K} \frac{\exp\{\boldsymbol{b}^T \boldsymbol{x}_{(k)}\}}{\sum_{i \in \mathcal{R}_k} \exp\{\boldsymbol{b}^T \boldsymbol{x}_i\}}, \tag{3.38}$$

and the log-partial likelihood is:

$$l_p(\boldsymbol{b}) = \sum_{k=1}^{K} \left\{ \boldsymbol{b}^T \boldsymbol{x}_{(k)} - \log\left(\sum_{i \in \mathcal{R}_k} \exp\{\boldsymbol{b}^T \boldsymbol{x}_i\} \right) \right\}. \tag{3.39}$$

The estimation and inference on $\boldsymbol{b}$ is based on $L_p(\boldsymbol{b})$ by maximizing partial likelihood. Standard errors are obtained via the partial likelihood observed information matrix:

$$\mathcal{I} = -\frac{\partial^2}{\partial \boldsymbol{b} \partial \boldsymbol{b}^T} l_p(\boldsymbol{b}). \tag{3.40}$$

A Bayesian justification of Cox's partial likelihood

Under the fully Bayesian semiparametric paradigm, priors are specified on both cumulative baseline hazard $\Lambda_0(t)$ and regression parameters $\boldsymbol{b}$, denoting as $p(\Lambda_0)$ and $p(\boldsymbol{b})$. These two priors are generally assumed independent. The marginal likelihood function for $\boldsymbol{b}$ is:

$$L(\boldsymbol{b}) = \int L(\boldsymbol{b}, \Lambda_0 | D) p(\Lambda_0) d\Lambda_0, \tag{3.41}$$

where $D = \{(y_i, \delta_i, \boldsymbol{x}_i) : i = 1, \cdots, n\}$ denotes the observed data and $L(\boldsymbol{b}, \Lambda_0 | D)$ denotes the full likelihood under (3.34). When a gamma process prior is specified for $\Lambda_0(t)$, such that

$$\Lambda_0(t) \sim \mathcal{G}(c\Lambda^*(t), c), \tag{3.42}$$

where c and Λ^* are the hyperparameters of the gamma process, Kalbfleisch [1978] first showed that as $c \to 0$, and to a first-order approximation, the marginal likelihood function for $\boldsymbol{b}$ is proportional to $L_p(\boldsymbol{b})$. Then, Bayesian inference about $\boldsymbol{b}$ in this case is carried out based on the approximate posterior density $p_{PL}(\boldsymbol{b}|D)$,

$$p_{PL}(\boldsymbol{b}|D) \propto L_p(\boldsymbol{b}|D) \times p(\boldsymbol{b}). \tag{3.43}$$

3.4 Bayesian CoxPH model with one covariate

In this section, we describe how to fit a Bayesian CoxPH model with a single covariate, which is the foundation for CoxPH single effect regression (SER) and CoxPH-SuSiE. We consider the following single covariate model with data from n individuals:

$$\lambda_i(t) = \lambda_0(t) \exp\{bx_i\}, i = 1, \cdots, n \tag{3.44}$$

$$b \sim N(0, \sigma_0^2), \tag{3.45}$$

where $\lambda_0(t)$ is the baseline hazard at time t and b is the coefficient for the single covariate x. This model assumes the prior distribution for b is a centered normal distribution with variance σ_0^2.

From Section 3.2.3, we can see the key elements to fit SER are: the posterior distribution of b and the Bayes factor. Of course there are other ways of doing inference for model (3.44)-(3.45), we choose a way which is most similar to SuSiE where we focus on getting the posterior distribution of b and the BF. Since both of these have no closed forms, we apply a quadratic approximation to the log-partial likelihood $l_p(b)$ at maximum partial likelihood

estimate (MPLE), denoting as $\hat{b}$. By Taylor expansion,

$$l_p(b) \approx l_p(\hat{b}) + l_p'(\hat{b})(b-\hat{b}) + \frac{l_p''(\hat{b})}{2}(b-\hat{b})^2 \tag{3.46}$$

$$= l_p(\hat{b}) + \frac{l_p''(\hat{b})}{2}(b-\hat{b})^2. \tag{3.47}$$

Then exponentiating both left-hand side and right-hand side,

$$L_p(b) \approx \hat{L}_p(b) = \exp\{l_p(\hat{b})\} \exp\{\frac{l_p''(\hat{b})}{2}(b-\hat{b})^2\}, \tag{3.48}$$

where $\hat{L}_p()$ denotes the approximate partial likelihood function and the second term on the right-hand side is a Gaussian kernel with variance equal to $-1/l_p''(\hat{b})$. In CoxPH model, $-1/l_p''(\hat{b})$ is the standard error for MPLE, denoted as s^2.

3.4.1 Posterior distribution of b

We apply the approximation for $L_p()$ in (3.48) and obtain an approximate posterior distribution of b, which is Gaussian.

$$b|\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\delta}, \sigma_0^2 \overset{\text{approx.}}{\sim} N(\mu_1, \sigma_1^2), \tag{3.49}$$

where

$$\sigma_1^2(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\delta}; \sigma_0^2) = \frac{1}{1/s^2 + 1/\sigma_0^2} \tag{3.50}$$

$$\mu_1(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\delta}; \sigma_0^2) = (\sigma_1^2/s^2)\hat{b}. \tag{3.51}$$

$\hat{b}$ is MPLE and s^2 is the corresponding standard error, which equals $-1/l_p''(\hat{b})$.

3.4.2 Bayes factor computation

The Bayes factor (BF) for comparing model (3.44)-(3.45) versus the null model ($b = 0$) is:

$$\text{BF} = \frac{P(D|H_1)}{P(D|H_0)} = \frac{\int\int p(\boldsymbol{y}, \boldsymbol{\delta}|\boldsymbol{x}, b, \Lambda_0)p(b)p(\Lambda_0)dbd\Lambda_0}{\int p(\boldsymbol{y}, \boldsymbol{\delta}|\boldsymbol{x}, b = 0, \Lambda_0)p(\Lambda_0)d\Lambda_0} = \frac{\int \big\{ \int p(\boldsymbol{y}, \boldsymbol{\delta}|\boldsymbol{x}, b, \Lambda_0)p(\Lambda_0)d\Lambda_0 \big\}p(b)db}{\int p(\boldsymbol{y}, \boldsymbol{\delta}|\boldsymbol{x}, b = 0, \Lambda_0)p(\Lambda_0)d\Lambda_0}. \tag{3.52}$$

Using results from Kalbfleisch 1978, Sinha et al. 2003, described in Section 3.3.1, when use a very diffuse gamma process prior on $\Lambda_0(t)$, the BF can be approximated as a function of the partial likelihood which does not depend on Λ_0:

$$\text{BF} \cong \frac{\int L_p(\boldsymbol{y}, \boldsymbol{\delta}|\boldsymbol{x}, b)p(b)db}{L_p(\boldsymbol{y}, \boldsymbol{\delta}|\boldsymbol{x}, b = 0)}. \tag{3.53}$$

The BF in (3.53) has no closed-form expression so we used approximate BFs.

Approximate Bayes factors

We considered three approximate BFs for this problem; these approximations were developed in consultation with Karl Tayeb. The first approximate BF is proposed by Wakefield [2009],

$$\text{BF}^{\text{W}} := \sqrt{\frac{s^2}{\sigma_0^2 + s^2}} \exp\Big(\frac{z^2}{2} \times \frac{\sigma_0^2}{\sigma_0^2 + s^2}\Big), \tag{3.54}$$

where z and s denote the z-score and standard error under CoxPH regression model of (3.44).

Then, we introduce another BF based on the quadratic approximation in (3.48), which we call Laplace BF:

$$\text{BF}^{\text{L}} = \frac{\int \hat{L}_p(b)p(b)db}{L_p(b = 0)} = \sqrt{\frac{s^2}{\sigma_0^2 + s^2}} \exp\big\{\frac{z^2}{2}\frac{\sigma_0^2}{\sigma_0^2 + s^2}\big\} \exp\{-\hat{b}^2/2s^2\}\frac{L_p(\hat{b})}{L_p(0)}, \tag{3.55}$$

where z and s are the same as in (3.54) and $\hat{b}$ denotes the MPLE under model (3.44). The

detailed calculation is available in Appendix B.1.

Gauss-Hermite quadrature

Here we introduce a numerical integration method, Gauss-Hermite quadrature and apply it to BF computation. The equation (9) in Naylor and Smith [1982] gives the following approximation formula:

$$\int_{-\infty}^{\infty} f(t)\phi(t;\mu,\sigma^2)dt \approx \sum_{i=1}^{n} w_i/\sqrt{\pi} f(t_i), \tag{3.56}$$

where $\phi(\cdot;\mu,\sigma^2)$ denotes the density of a Gaussian distribution with mean μ and variance σ^2 ; $t_i = \mu + \sqrt{2}\sigma x_i$; x_i are the zeros of the n^{th} order Hermite polynomial and w_i are corresponding weights; see Davis and Rabinowitz [2007].

This formula can be used to approximate an integral $I = \int_{-\infty}^{\infty} g(t)dt$ by writing it as

$$\int_{-\infty}^{\infty} g(t)dt = \int_{-\infty}^{\infty} h(t)\phi(t;\mu,\sigma^2)dt, \tag{3.57}$$

where $h(t) := g(t)/\phi(t;\mu,\sigma^2)$. Here μ, σ^2 are arbitrary, and so should be chosen to make the approximation as accurate as possible. To achieve this, Liu and Pierce [1994] suggests that μ, σ^2 should be chosen so that $g(x) \approx c\phi(x;\mu,\sigma^2)$ for some constant c.

The numerator of the BF is an integral I with $g(b) = L_p(b)\phi(b;0,\sigma_0^2)$ (see equation (3.53)), which is proportional to the posterior distribution of b. We therefore select μ and σ^2 to be the approximate posterior mean and variance (3.50)-(3.51), and apply the approximation formula (3.56) with $n = 32$. Increasing the value of n can increase the accuracy of the approximation to the BF.

3.5 CoxPH single effect regression

In previous section, we've introduced Bayesian CoxPH model with single covariate. In this section, we describe CoxPH single effect regression (SER), which is the building block of CoxPH-SuSiE. Similar to SER model in (3.9)-(3.13), CoxPH-SER also assumes exactly one of the p explanatory variables has a non-zero coefficient. Specifically, we consider the following model:

$$\lambda_i(t) = \lambda_0(t) \exp\{\boldsymbol{b}^T \boldsymbol{x}_i + o_i\}, i = 1, \cdots, n \tag{3.58}$$

$$\boldsymbol{b} = b\boldsymbol{\gamma} \tag{3.59}$$

$$\boldsymbol{\gamma} \sim \text{Mult}(1, \boldsymbol{\pi}) \tag{3.60}$$

$$b \sim N(0, \sigma_0^2), \tag{3.61}$$

where $\lambda_i(t)$ is the hazard for individual i and $\lambda_0(t)$ is the baseline hazard. $\boldsymbol{x}_i$ stores the values of p explanatory variables of individual i. $\boldsymbol{b}$ is the effect size vector and b is the scalar value for the 'single effect'. $\boldsymbol{\gamma} \in \{0, 1\}^p$ is the same as in the SER model of (3.9)-(3.13), denoting a p-vector of indicator variables with exactly one non-zero entry. The prior distributions of $\boldsymbol{\gamma}$ and b are also the same as the SER model (3.9)-(3.13), where $\boldsymbol{\pi} = (\pi_1, \cdots, \pi_p)$ represents prior inclusion probabilities. Additionally, we introduce an offset o_i for each individual i, which is a fixed constant value added to the linear predictors. In the CoxPH-SER model, the offset o_i is set to 0 for $i = 1, \cdots, n$ (therefore, ignored in the derivation). The role of the offset in solving CoxPH-SuSiE will be explained later in Section 3.6.

3.5.1 Posteriors under CoxPH-SER model

The posterior distribution of $\boldsymbol{\gamma}$ is:

$$\boldsymbol{\gamma}|\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\delta} \sim \text{Mult}(1, \boldsymbol{\alpha}) \tag{3.62}$$

$$\boldsymbol{\alpha}_j = P(\boldsymbol{\gamma}_j = 1|\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\delta}) = \frac{\pi_j \text{BF}_j}{\sum_{j'} \pi_{j'} \text{BF}_{j'}}, \tag{3.63}$$

where $\boldsymbol{X}$ is the $n \times p$ matrix of values for p covariates across n individuals, and BF_j is the Bayes factor comparing the model with j^{th} covariate only versus the null model. The posterior distribution of b doesn't have a closed-form, but we can use the normal approximation in Section 3.4.1. Therefore,

$$b|\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\delta}, \boldsymbol{\gamma}_j = 1, \sigma_0^2 \overset{\text{approx.}}{\sim} N(\mu_{1j}, \sigma_{1j}^2), \tag{3.64}$$

where μ_{1j} and σ_{1j}^2 are the approximate posterior mean and variance from model (3.44) -(3.45) with $\boldsymbol{x}_j$ being the single covariate.

Now that we have all the pieces required for solving CoxPH-SER model, we define a CoxPH-SER module for later convenience:

$$\text{CoxPH-SER}(\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\delta}, \boldsymbol{o}; \sigma_0^2) := (\boldsymbol{\alpha}, \boldsymbol{\mu}_1, \boldsymbol{\sigma}_1^2), \tag{3.65}$$

where the input data is $(\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\delta}, \boldsymbol{o})$ and the output is $(\boldsymbol{\alpha}, \boldsymbol{\mu}_1, \boldsymbol{\sigma}_1^2)$.

3.5.2 Prior variance estimation

In earlier sections, σ_0^2 is treated as a fixed parameter. In fact, it can be estimated using maximum likelihood, which is essentially an empirical Bayes approach. The marginal likelihood

for σ_0^2 under CoxPH-SER is:

$$L(\sigma_0^2) := p(\boldsymbol{y}|\boldsymbol{X}, \sigma_0^2). \tag{3.66}$$

We use an expectation-maximization (EM) algorithm to solve the problem iteratively. The key idea is to augment the data by $(\boldsymbol{\gamma}, b)$. The E-step computes the expected complete data log-likelihood, which uses the posterior distribution of $\boldsymbol{\gamma}$ and the approximate posterior of $b|\boldsymbol{\gamma}$ (Section 3.5.1) in the calculation. The M-step maximizes the expected complete data log-likelihood with respect to σ_0^2, and results in the following update:

$$\sigma_0^2 \leftarrow \sum_{j=1}^{p} \alpha_j(\mu_{1j}^2 + \sigma_{1j}^2), \tag{3.67}$$

where μ_{1j} and σ_{1j}^2 denote the approximate posterior mean and variance of $b|\boldsymbol{\gamma}_j = 1$. The full derivation for the EM algorithm is available in Appendix B.2.

3.6 CoxPH SuSiE

In this section, we introduce the CoxPH-SuSiE model and the algorithm for solving it. The CoxPH-SuSiE model is as follows:

$$\lambda_i(t) = \lambda_0(t) \exp\{\boldsymbol{x}_i^T \boldsymbol{b}\}, i = 1, \cdots, n \tag{3.68}$$

$$\boldsymbol{b} = \sum_{l=1}^{L} \boldsymbol{b}_l \tag{3.69}$$

$$\boldsymbol{b}_l = \boldsymbol{\gamma}_l b_l \tag{3.70}$$

$$\boldsymbol{\gamma}_l \sim \text{Mult}(1, \boldsymbol{\pi}) \tag{3.71}$$

$$b_l \sim N_1(0, \sigma_{0l}^2), \tag{3.72}$$

where $\lambda_i(t)$ denotes the hazard for individual i and $\lambda_0(t)$ denotes the baseline hazard. $\boldsymbol{b}$ is the overall effect vector, which is the sum of L single effect vectors, denoted as $\boldsymbol{b}_l$ for the l^{th} one. The priors for $\boldsymbol{\gamma}_l$ and $b_l, l = 1, \cdots, L$ are the same as SuSiE model.

To fit CoxPH-SuSiE, we implemented an algorithm similar to the IBSS algorithm used in SuSiE, which we refer to as Generalized Iterative Bayesian Stepwise Selection (GIBSS). Details can be found in Algorithm 3. For each l, the offset $\boldsymbol{o}_l$ is first computed given $\bar{\boldsymbol{b}}_{l' \neq l}$. Recall that in SuSiE, when we solve the l^{th} sub-problem, we keep $l' \neq l$ fixed and subtract $\boldsymbol{X} \sum_{l' \neq l} \bar{\boldsymbol{b}}_{l'}$ from $\boldsymbol{y}$ (see Algorithm 2). This subtraction is no longer possible in CoxPH model because the hazard is not a linear combination of the predictors. Therefore, instead of subtraction, we use offset $\boldsymbol{o}_l := \boldsymbol{X} \sum_{l' \neq l} \bar{\boldsymbol{b}}_{l'}$ for the fixed part. Then we apply CoxPH-SER function to compute the posterior inclusion probabilities $\boldsymbol{\alpha}_l$ and the approximate posterior distribution of $\boldsymbol{b}_l$. The prior variance σ_{0l}^2 is estimated using (3.67).

The IBSS algorithm to fit linear SuSiE is a variational approximation algorithm (see Section 3.2.3), while the GIBSS algorithm is a heuristic procedure; we do not have a clear understanding of the exact objective function it optimizes, and there is no guarantee of convergence. However, GIBSS has the advantages of being modular and simple, and it has shown good empirical performance, see Section 3.7 for simulation results.

Algorithm 3: Generalized Iterative Bayesian stepwise selection (GIBSS)

Data: $\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\delta}$

Require: Number of effects, $\boldsymbol{L}$

Require: A function CoxPH-SER$(\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\delta}, \boldsymbol{o}; \sigma_0^2) \rightarrow (\boldsymbol{\alpha}, \boldsymbol{\mu}_1, \boldsymbol{\sigma}_1)$ that computes the approximate posterior distributions under CoxPH-SER;

Initialization: posterior means $\bar{\boldsymbol{b}}_l = \boldsymbol{0}, \sigma_{0l}^2 = 1$ for $l = 1, \ldots, \boldsymbol{L}$; $\boldsymbol{o} = \boldsymbol{0}$; **repeat**

  for $l \in 1, \ldots, \boldsymbol{L}$ **do**

    $\boldsymbol{o}_l \leftarrow \boldsymbol{X} \sum_{l' \neq l} \bar{\boldsymbol{b}}_{l'}$

    $(\boldsymbol{\alpha}_l, \boldsymbol{\mu}_{1l}, \boldsymbol{\sigma}_{1l}) \leftarrow$ CoxPH-SER$(\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\delta}, \boldsymbol{o}_l; \sigma_{0l}^2)$

    $\sigma_{0l}^2 \leftarrow \sum_{j=1}^{p} \alpha_{jl}(\mu_{1jl}^2 + \sigma_{1jl}^2)$

    $\bar{\boldsymbol{b}}_l = \boldsymbol{\alpha}_l \circ \boldsymbol{\mu}_l$ "$\circ$" denotes element-wise multiplication.

  end

until *convergence criterion satisfied*;

Return: $\boldsymbol{\alpha}_1, \boldsymbol{\mu}_{11}, \boldsymbol{\sigma}_{11}, \ldots, \boldsymbol{\alpha}_L, \boldsymbol{\mu}_{1L}, \boldsymbol{\sigma}_{1L}$

3.7 Simulation

First, we assess the behaviour of different Bayes factors (BFs) as the accuracy of BF can have direct impact on CoxPH-SuSiE results. Then, we conduct simulation using real genotype data to assess the performance of our method and compare with other methods.

3.7.1 Data generation procedure

We define the following function to generate data:

$$(\boldsymbol{y}, \boldsymbol{\delta}, \boldsymbol{X}) \leftarrow \text{simsurv}(\boldsymbol{X}, m, \sigma_0^2, r), \tag{3.73}$$

where $\boldsymbol{y}$ and $\boldsymbol{\delta}$ are both size n vectors, storing the observed times and survival status for n individuals. $\boldsymbol{X}$ is a genotype matrix of size $n \times p$. We use the real genotype data from UK

biobank or GTEx Consortium for $\boldsymbol{X}$. m is the number of non-zero effects, σ_0^2 is the prior variance of non-zero effects and r is the censoring level. simsurv() takes the following steps to generate data:

1. Generate a vector $\boldsymbol{b}$ of size $p+1$, where the first element is 1 and the rest are all 0s. Generate m non-zero effect sizes from $N(0, \sigma_0^2)$ and place them to $\boldsymbol{b}$ at the m non-zero indexes, which are sampled from $\{2, 3, \cdots, p+1\}$. Then, $\boldsymbol{b}$ represents the true effect size vector and the first element in $\boldsymbol{b}$ represents the intercept.

2. Compute survival rate $\lambda_i^s = (1, \boldsymbol{x}_i)^T \boldsymbol{b}$ for each individual i, where $\boldsymbol{x}_i$ is a vector of size p denoting individual i's genotype.

3. Compute censor rate λ^c given censor level r and survival rate $\lambda_i^s, i = 1, \cdots, n$.

$$\frac{\lambda^c}{\lambda^c + \bar{\lambda}^s} = r, \ \lambda^c = \frac{r\bar{\lambda}^s}{1-r}, \tag{3.74}$$

   where $\bar{\lambda}^s = \sum_{i=1}^n \lambda_i^s / n$ is the mean of survival rates across n individuals. Additional discussion about this step is available in Supplementary B.3.

4. Simulate survival time T_i and censor time C_i for each individual i from exponential distribution:

$$T_i \sim \exp(\lambda_i^s) \tag{3.75}$$

$$C_i \sim \exp(\lambda^c), \tag{3.76}$$

   where λ_i^s and λ^c are the rate parameters for the exponential distributions.

5. Determine observed time y_i and survival status δ_i for each individual i, $y_i = \min(T_i, C_i)$. $\delta_i = 1$ if $T_i \leq C_i$ otherwise 0.

This data simulation scheme satisfies the proportional hazard assumption.

3.7.2 Simulation for Bayes factor comparison

We compare the performance of different Bayes factors (BFs): Laplace BF (BF^{L}), Wakefield BF (BF^{W}) and Gauss-Hermite quadrature BF (BF^{GH}) under CoxPH model with one single covariate.

To generate simulation data, we follow the procedure described in Section 3.7.1. We set the sample size n to 500000, which is similar to the sample size of UK biobank. The genotype x_i is generated using Binomial distribution with a pre-specified minor allele frequency (MAF). We vary the effect size of the single effect variable, $(0.01, 0.1)$. We choose these effect sizes as they are similar to the effect size estimates from GWAS. We also experiment with different censoring level $r = (0, 0.2, 0.4, 0.6, 0.8, 0.99)$ and different MAF $= (0.001, 0.01, 0.1)$. For each scenario, 50 replicates are conducted. For computing BF^{GH}, we used the `gauss.quad.prob()` function from a R package `statmod` [Smyth et al., 2017], and the number of nodes were set to 32. This should result in an accurate BF estimate, at least more accurate than the other two approximate BFs.

Figure 3.1 and Figure 3.2 summarize the comparison of different approximate BFs. We view BF^{GH} as the gold standard as it is the most accurate one and plot the other two BFs against BF^{GH}. All comparisons are on the $\log_{10}$ scale. When the true effect size is tiny ($b = 0.01$), both BF^{W} and BF^{L} are accurate when censoring level is not too high. When the censoring level is extremely high ($r = 0.99$) and the minor allele frequency is very low (MAF $= 0.001$), we can see BF^{W} diverges from BF^{GH}. When the true effect size is larger ($b = 0.1$), BF^{W} over estimates the true BF in general (except when censoring level is 0.99 and MAF is 0.001), and the difference between BF^{W} and BF^{GH} gets larger as the true BF gets larger. Across all scenarios, BF^{L} is accurate. Therefore, we use BF^{L} in CoxPH-SuSiE. Even though BF^{GH} can be more accurate, it is computationally intensive, taking an average of 78.07 seconds when the sample size is $n = 500000$.

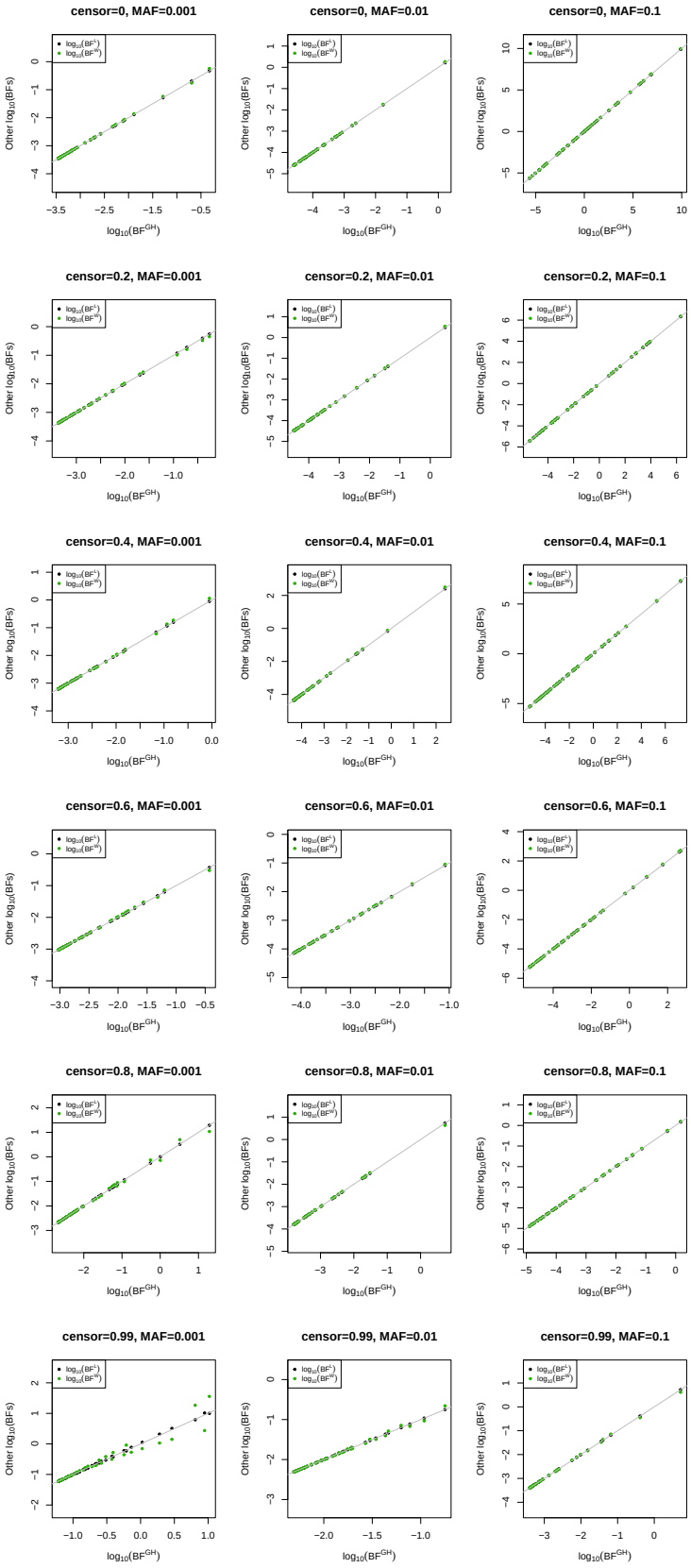

Figure 3.1: Scatter plots of Bayes Factors (BFs) on $\log_{10}$ scale under different censoring levels and minor allele frequencies. The effect size of the single variable in CoxPH model is 0.01. The grey solid line represents $y = x$.

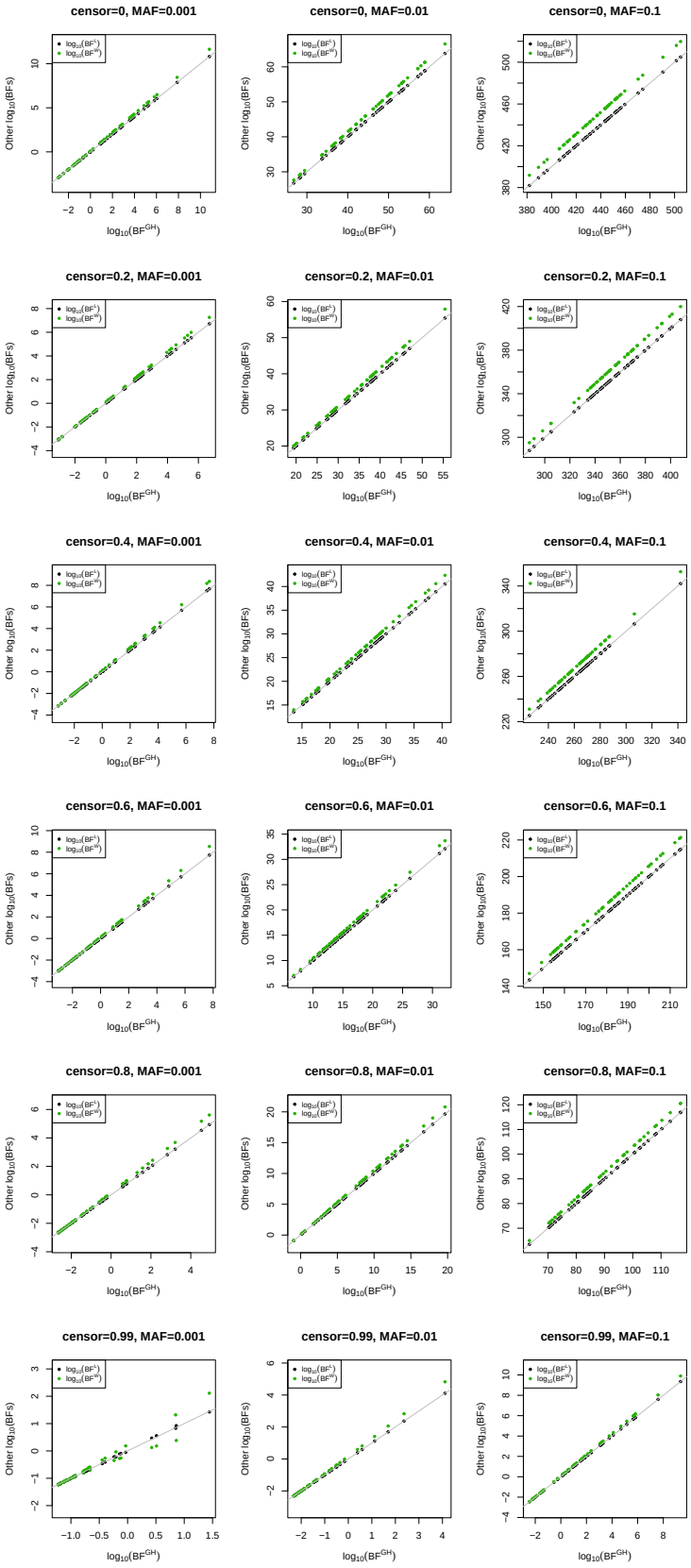

Figure 3.2: Scatter plots of Bayes Factors (BFs) on $\log_{10}$ scale under different censoring levels and minor allele frequencies. The effect size of the single variable in CoxPH model is 0.1. The grey solid line represents $y = x$.

Simulation for methods comparison

We compare our method, CoxPH-SuSiE, with the following methods: BVSNLP of Nikooienejad et al. [2020], survival.svb of Komodromos et al. [2022] and R2BGLiMS of Newcombe et al. [2017]. R2BGLiMS uses a Weibull regression model, which is less flexible than CoxPH regression, but this shouldn't cause a problem since our survival data is simulated under an even simpler distribution, the exponential distribution, which results in a constant baseline hazard under proportional hazards models. This can be viewed as a special case for Weibull regression and CoxPH regression. Additionally, we compare CoxPH-SuSiE with a heuristic approach, which involves first obtaining GWAS summary statistics using CoxPH regression, followed by fine-mapping with SuSiE.RSS developed by Zou et al. [2022]. SuSiE.RSS is based on linear SuSiE, and its derivation does not extend to nonlinear regression, making this approach heuristic.

To compare different methods, we consider two simulation settings: one with a moderate sample size and relatively large effect sizes, and another with a large sample size and smaller effect sizes. In the former setting, referred to as the GTEx simulation, we use real genotype data from GTEx Consortium et al. [2015], with a total sample size of $n = 574$. For each simulation replicate, we randomly select a region on gene ENSG00000132855. In the latter setting, which we call UKB simulation, we use real genotype data from the UK Biobank Bycroft et al. [2018a], sub-sampling $n = 50000$ individuals. For each simulation replicate, we randomly select a region on Chromosome 3. The data generation parameters for both simulations are summarized in Table 3.1. In each simulation setting, we vary the number of effect variables from 0 to 3 and censoring level r. In GTEx simulation, $r = (0, 0.2, 0.4, 0.6, 0.8)$. We ran 20 replicates for each scenario. In UKB simulation, we also include a scenario with an extremely high censoring level, therefore, $r = (0, 0.2, 0.4, 0.6, 0.8, 0.99)$. We ran 10 replicates for each scenario.

For the GTEx simulation, we used the default number of iterations provided by each

Genotype data	Sample size (n)	Number of variables (p)	prior variance (σ_0^2)
GTEx	574	1000	1
UK Biobank	50000	1000	0.1

Table 3.1: Summary of two data generation settings.

methods software. However, for the UKB simulation, since the sample size is much larger, running the default number of iterations for some methods was time-consuming. Therefore, we made these methods to run similar amount of time. For SuSiE.RSS, R2BGLiMS and BVSNLP, we still use the default number of iterations. For CoxPH-SuSiE and survival.svb, we ran at most 10 iterations and 100 iterations, respectively. For running CoxPH-SuSiE and SuSiE.RSS, we set $L = 5$.

Figure 3.3 and 3.4 visualize the distribution of posterior inclusion probabilities (PIPs) of different methods across two simulations. The PIPs of survival.svb show significant inconsistency with CoxPH-SuSiE PIPs. Survival.svb tends to have PIPs either close to 0 or 1, except when censor rate is 0.99, and it assigns a PIP value of 1 to many non-effect variables. This suggests survival.svb may have a high false positive rate. The PIPs of other methods (BVSNLP, R2BGLiMS and SuSiE.RSS) are slightly more consistent with CoxPH-SuSiE PIPs. When comparing CoxPH-SuSiE PIPs to SuSiE.RSS PIPs in GTEx simulation, we observe more effect variables in the bottom-right half of the plots, indicating that SuSiE.RSS may have lower power than CoxPH-SuSiE.

To assess the quality of the PIPs, we check the calibration of each method. Variables were first grouped into bins based on their PIPs. And then we plot the mean reported PIP (x-axis) against the empirical proportion of effect variables in that bin, see Figure 3.5 and 3.6. A well-calibrated method should have mean PIP agrees with observed frequency. In GTEx simulation, CoxPH-SuSiE and SuSiE.RSS are the best calibrated methods, while survival.svb is the worst. In UKB simulation, CoxPH-SuSiE again is the best one and survival.svb is the worst. BVSNLP, SuSiE.RSS and R2BGLiMS showed poorer calibration compared with in GTEx simulation.

In addition to directly comparing PIPs, we plotted power versus false discovery rate (FDR) at different censoring levels, as shown in Figure 3.7 and Figure 3.8. In both simulations, survival.svb exhibited a very high FDR, consistent with the PIP results. As the censoring level increased, the performance of all methods declined. In the GTEx simulation, CoxPH-SuSiE and BVSNLP showed similar strong performance, followed by SuSiE.RSS. In the UKB simulation, CoxPH-SuSiE was the best-performing method, followed by BVSNLP and SuSiE.RSS. These results suggest that in fine-mapping settings where variables are highly correlated with tiny effect sizes, CoxPH-SuSiE is the most effective method. Even in scenarios with larger effect sizes, such as in the GTEx simulation, CoxPH-SuSiE has the advantage over BVSNLP that it provides credible sets for uncertainty quantification, which BVSNLP does not offer.

Among all the methods, only CoxPH-SuSiE and SuSiE.RSS directly output credible sets (Definition 1). Figure 3.9 and Figure 3.10 compare the coverage, power and mean absolute correlation of the 95% CSs across different numbers of non-zero effects and different levels of censoring. Coverage is the proportion of CSs that contain an effect variable, and power is the overall proportion of CSs that contain an effect variable. Mean absolute correlation reflects the purity of the CS. From Figure 3.9 and Figure 3.10, we can see under both simulation settings, the CSs of CoxPH-SuSiE have high coverage, close to or higher than 90%. The CSs of CoxPH-SuSiE not only achieve higher coverage, they also have higher power than CSs of SuSiE.RSS in all cases. In GTEx simulation, the CSs of CoxPH-SuSiE are purer than those of SuSiE.RSS, while in UK Biobank simulation, CSs of SuSiE.RSS have higher purity more often.

3.8 Real data analysis

To demonstrate our method on real data, we analyzed self-reports of doctor-diagnosed asthma (data field 3786 and 22147) on UK Biobank samples. The UK Biobank is a very

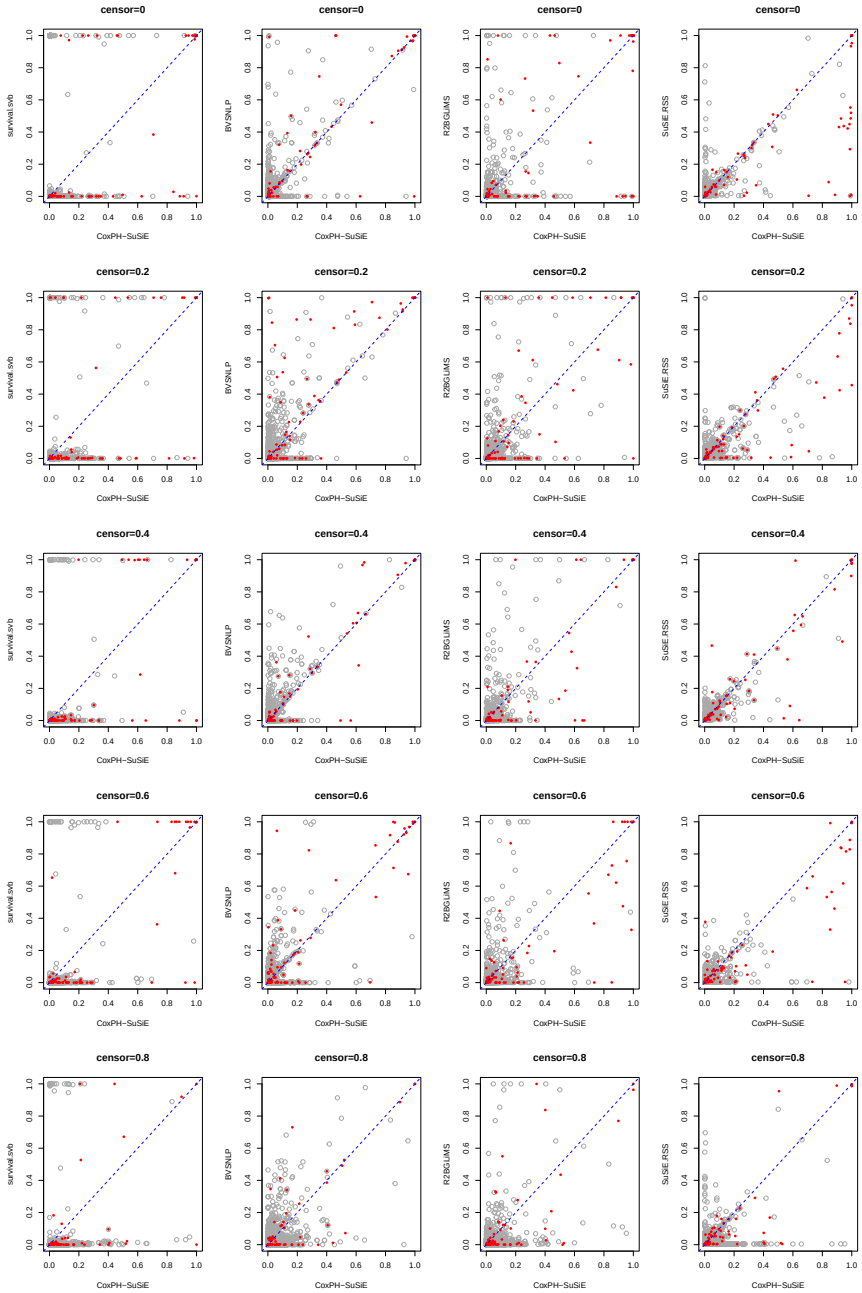

Figure 3.3: Comparison of posterior inclusion probabilities (PIPs) of different methods on GTEx genotype data. Grey circles represent zero effect variables and red dots represent non-zero effect variables. The blue dashed line represents $y = x$.

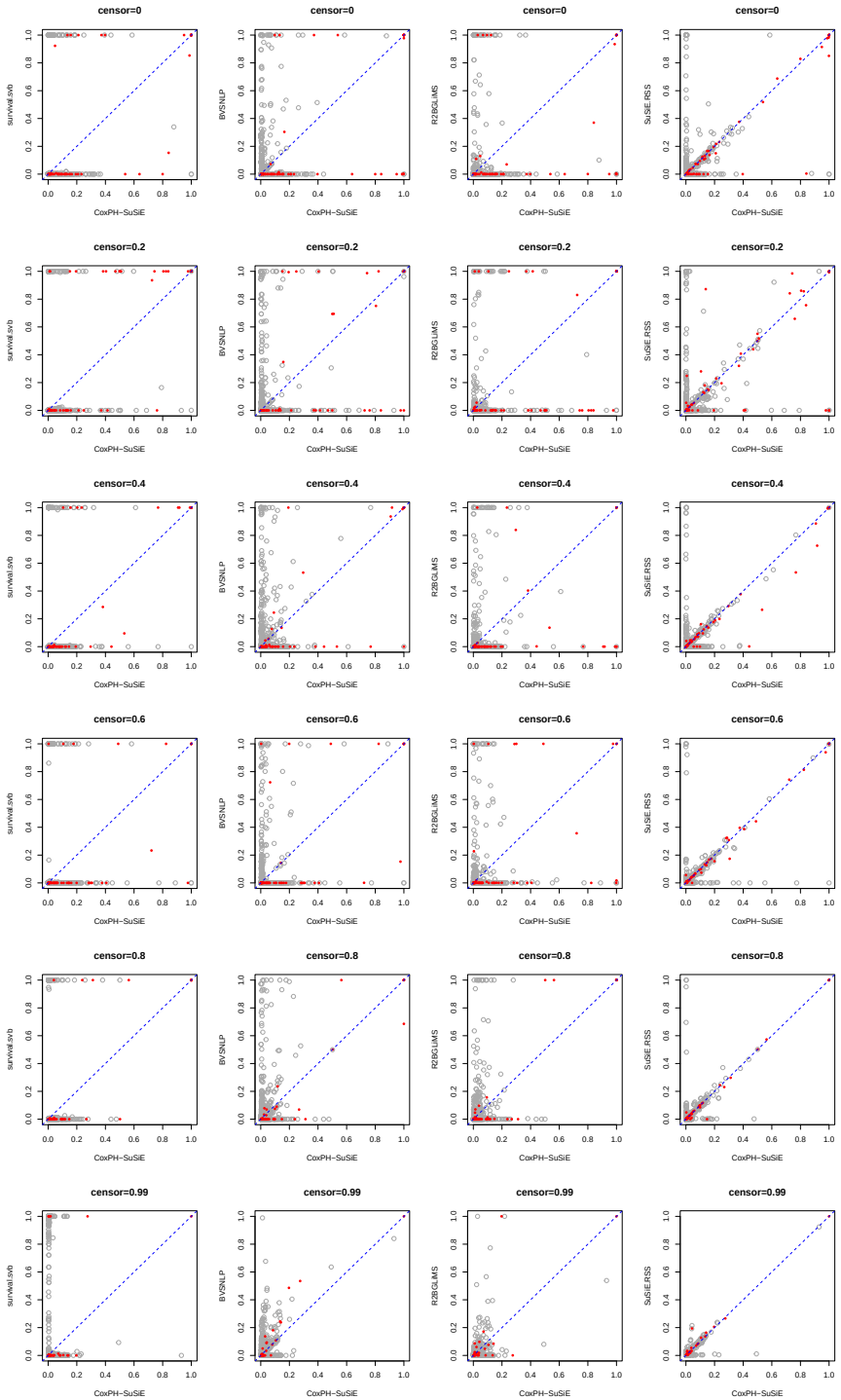

Figure 3.4: Comparison of posterior inclusion probabilities (PIPs) of different methods on UK biobank genotype data. Grey circles represent zero effect variables and red dots represent non-zero effect variables. The blue dashed line represents $y = x$.

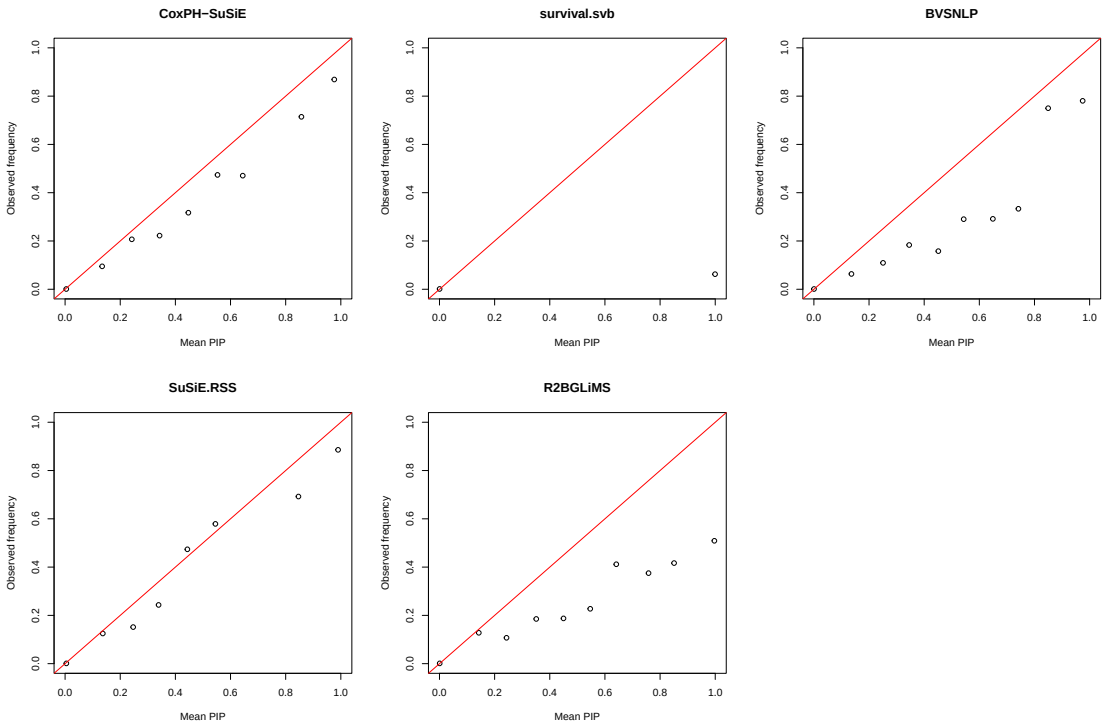

Figure 3.5: Assessment of PIP calibration in GTEx simulation. Variables across all simulations were grouped into 10 equal bins from 0 to 1 based on their PIP values. Bins with fewer than 10 observations are removed in plotting.

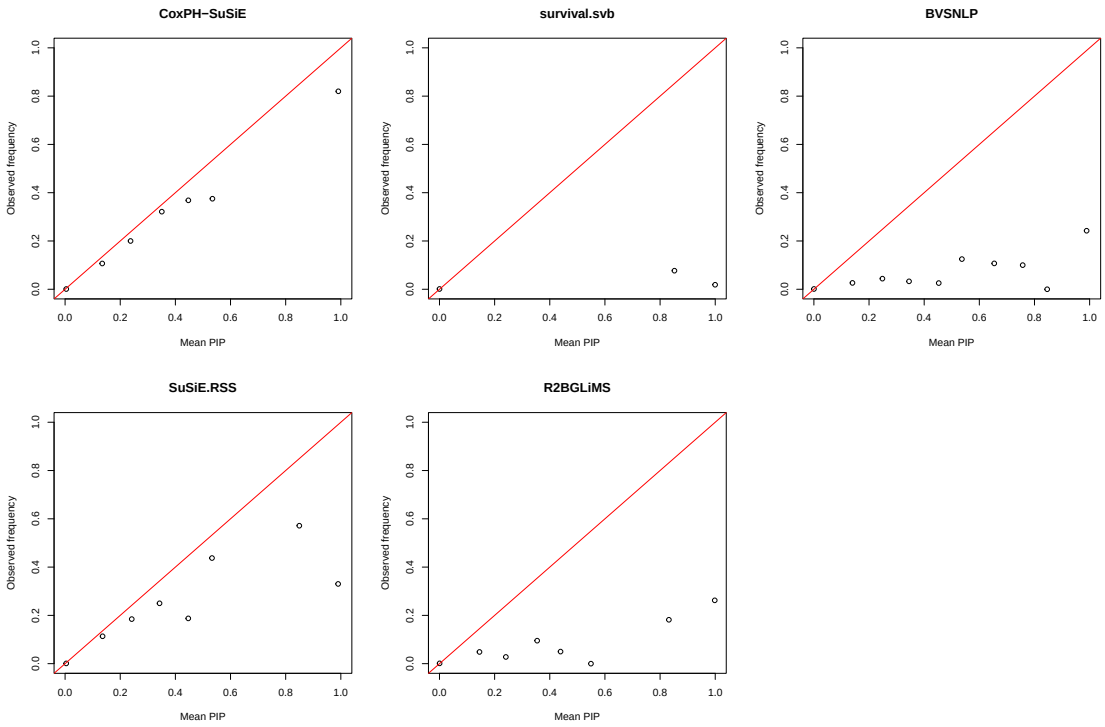

Figure 3.6: Assessment of PIP calibration in UKB simulation. Variables across all simulations were grouped into 10 equal bins from 0 to 1 based on their PIP values. Bins with fewer than 10 observations are removed in plotting.

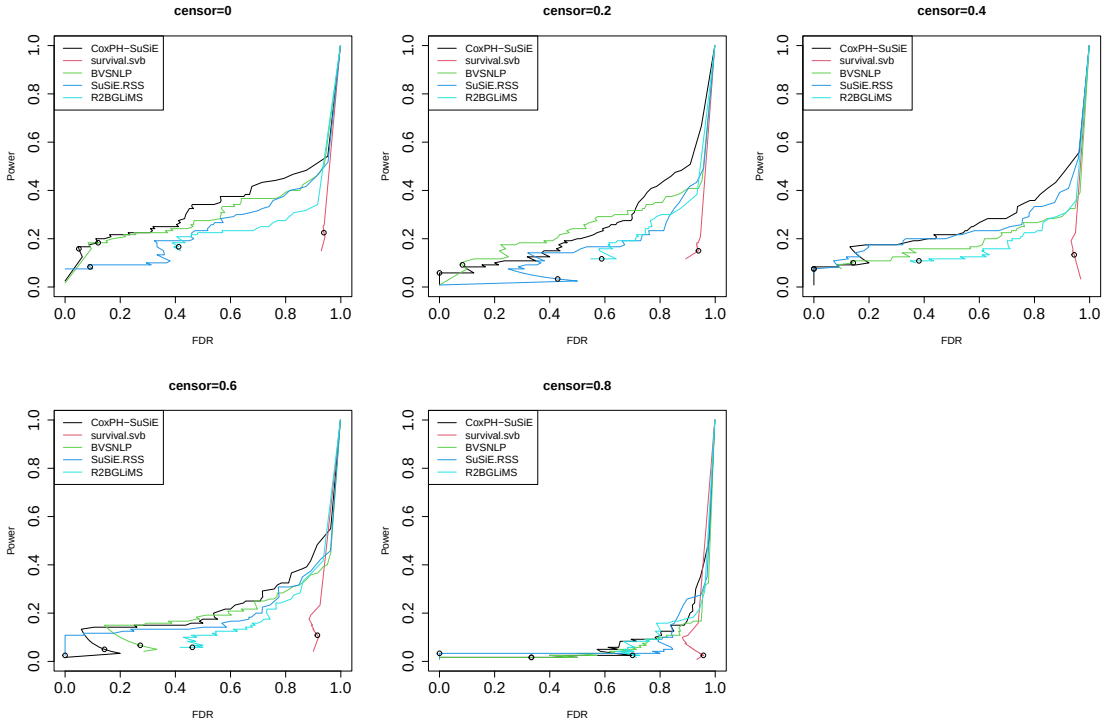

Figure 3.7: Power versus FDR at different censoring level in GTEx simulation. The open circles highlight power versus FDR at PIP threshold of 0.95. FDR := FP/(TP+FP) and power:=TP/(TP + FN) where FP, TP, FN and TN denote the number of False Positives, True Positives, False Negatives and True Negatives, respectively.

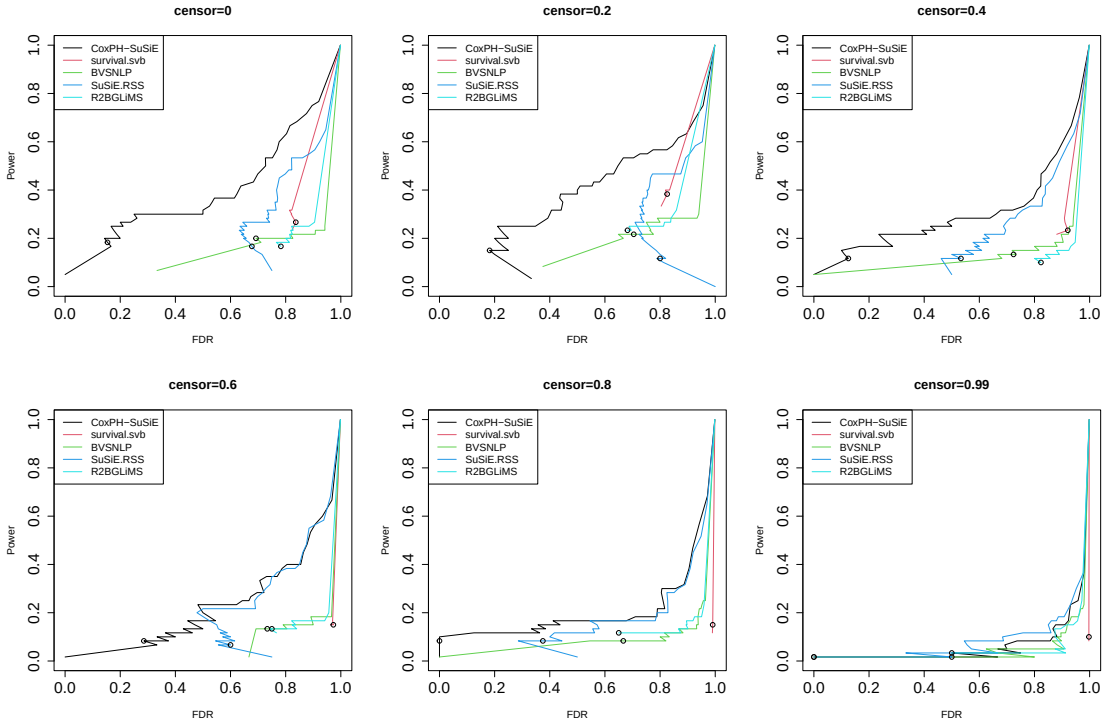

Figure 3.8: Power versus FDR at different censoring level in UKB simulation. The open circles highlight power versus FDR at PIP threshold of 0.95.

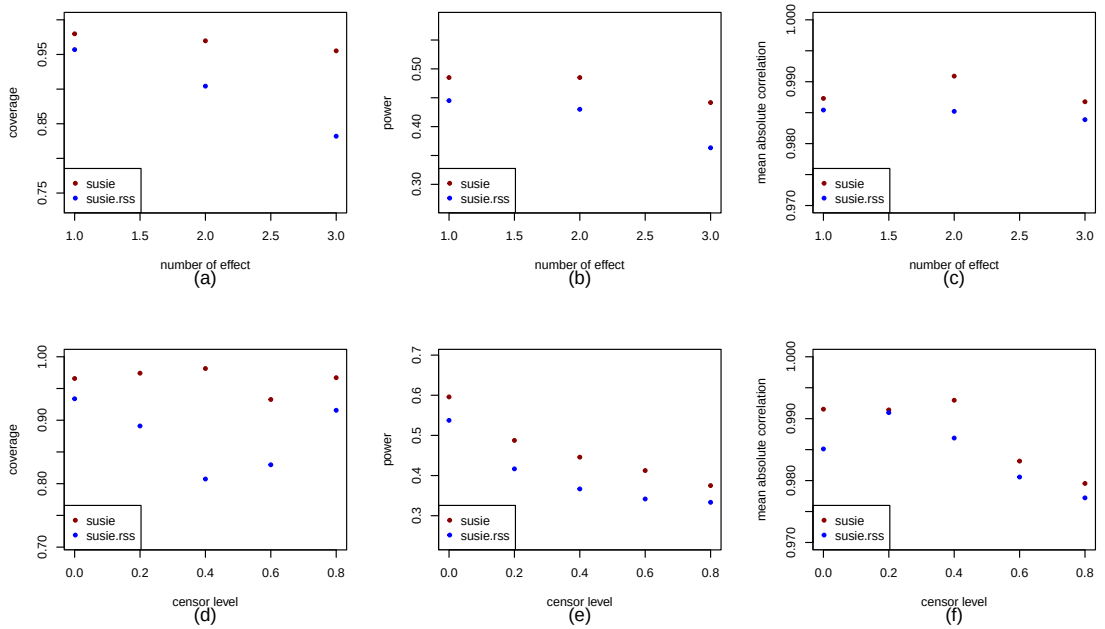

Figure 3.9: Credible sets assessment under GTEx simulation. Statistics (coverage, power and mean absolution correlation) are averaged across data replicates.

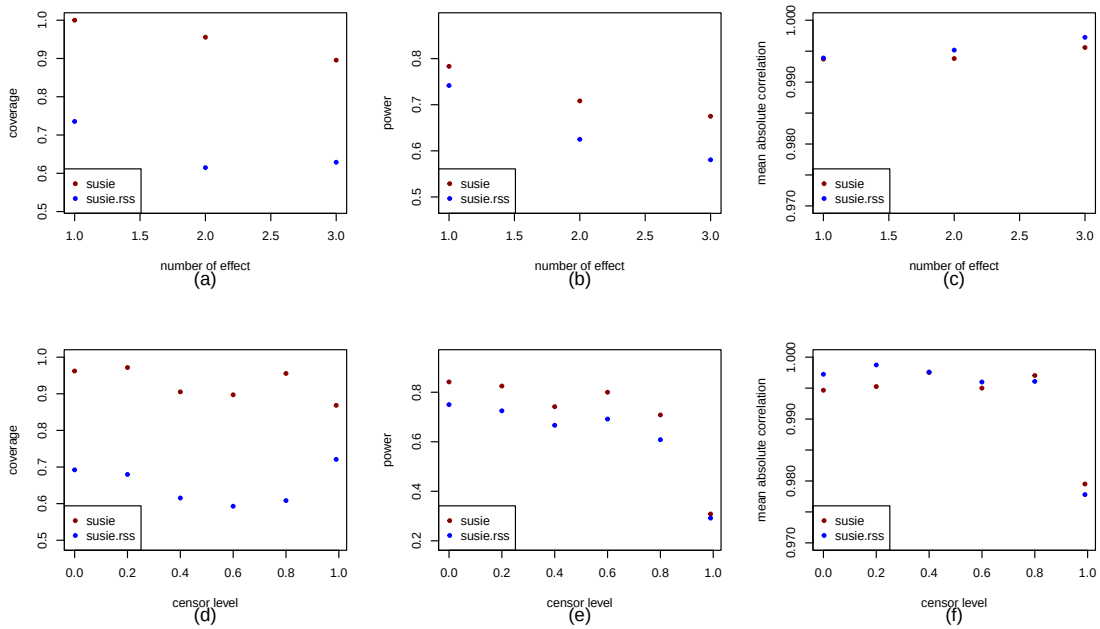

Figure 3.10: Credible sets assessment under UKB simulation. Statistics (coverage, power and mean absolution correlation) are averaged across data replicates.

large, population-based prospective study, with detailed phenotype and genotype data from over 500,000 participants in the United Kingdom, with ages between 40 and 69 at time of recruitment [Sudlow et al., 2015, Bycroft et al., 2018a]. Before conducting fine-mapping analysis, there are two questions we want to assess.

First, whether modeling time to disease onset using CoxPH regression is more powerful than modeling disease case-control status using logistic regression. Green and Symons [1983] described the theoretical relationship between CoxPH model and logistic model in prospective cohort studies. Under a constant baseline hazard, they show that when the follow-up period is short and the disease is relatively rare with not too great risk factors, the regression coefficients of logistic model approximate those of CoxPH model. In contrast, the approximation becomes poorer for more common diseases, longer follow-up time, and large effect risk factors. Staley et al. [2017] assessed the performance of the two models using simulated and real genetic data in cohort studies. They concluded that CoxPH model results

in a modest improvement in power to detect SNP–disease associations, and the improvement increased as the disease incidence increased. Additionally, logistic regression yielded inflated effect size estimates, especially for SNPs with larger effect.

Second, whether the proportional hazard (PH) assumption satisfies in asthma. Asthma is known to be a complex chronic respiratory disease, which may comprise many different conditions [Fuchs et al., 2017]. Growing evidence shows childhood onset and adulthood onset asthma have different sex ratios, triggers of symptoms, associated comorbidities, severity and potential genetic risk loci [Larsen, 2000, Pividori et al., 2019, Ferreira et al., 2019]. If the PH assumption holds, the hazard ratio among individuals with different genotype would be constant across lifetime in CoxPH model. However, there are risk loci that are seemingly specific to childhood onset or adulthood onset asthma [Pividori et al., 2019]. A locus that only affects childhood asthma risk won't have any effect in one's adulthood. Consequently, the hazard ratios across different genotype change over time, violating the PH assumption.

To evaluate the first question in asthma, we compared the log-likelihood ratios of top SNPs under CoxPH model and logistic model. For the second question, we perform three sets of association analysis: all asthma combined (AA), childhood onset asthma (COA) and adulthood onset asthma (AOA). Moreover, we plot Kaplan-Meier curves at SNPs with most significant p-values and large effect sizes.

3.8.1 Data preprocessing & association studies

To limit confounding due to population structure, we focus on "White British" (data field 22009) in UK biobank. We used genotype data of version 3, and removed samples based on the following exclusion criteria, similar to Zou et al. [2023]: (1) individuals who don't know or don't wish to answer their asthma diagnosis age; (2) individuals who withdraw from UK biobank; (3) mismatch between self-reported and genetic sex; (4) outlier genotype samples based on heterozygosity and/or rate of missing genotypes defined by UK Biobank data

field 22027; (5) individuals who have at least one relative in the cohort based on kinship calculations (samples with a value other than zero in data field 22021). After filtering genotype samples according to these criteria, 273543 samples remained.

We began by computing summary statistics of associations under logistic model and CoxPH model. We selected 9 genomic regions based on results in Pividori et al. [2019], which include loci shared between COA and AOA (2q12.1, 6p21.33, 10p14, 11q13.5, 12q13.11, 15q22.2), COA-specific loci (1q21.3, 17q12) and AOA-specific loci (2q22.3).

To prepare samples for AA logistic association analyses, we simply included all UK Biobank samples that met the filtering criteria. For COA and AOA logistic association analyses, we used the definitions outlined in Pividori et al. [2019]. Specifically, COA cases are individuals who developed asthma before age 12, while controls are those who either never developed asthma or developed it after age 26. For AOA, cases are those who developed asthma between ages 25 and 65, with controls being individuals who never developed asthma. A summary of samples used in the three sets of analysis is provided in Table 3.2.

For CoxPH association analysis, we used the same samples as logistic association analysis. For AA analysis, the time to event is the age of asthma onset for people who got asthma. For people who haven't got asthma, we use age at the most recent visit to assessment center as censoring time (data field 21003). For COA, the time to event is the age of asthma onset for COA cases, and is treated as censored at age 12 for controls. For AOA, the time to event is age of asthma onset for AOA cases, and is treated as censored at their current age (or age 65 if their age exceeds 65) for controls. To compute CoxPH associations, we used an R package "SPACox" [Bi et al., 2020], which employs saddlepoint approximation to calibrate the test statistics.

For both association analysis, we adjusted for 10 genetic principal components (PCs) stored in data field 22009 and the sex in data field 31.

	AA	COA	AOA
Case	People with asthma	Asthma onset ≤ 12	Asthma onset between 26-65
Control	People without asthma	People without asthma + Asthma onset ≥ 26	People without asthma

Table 3.2: A summary of samples used in AA/COA/AOA logistic analysis

3.8.2 Exploratory data analysis results

To evaluate whether the CoxPH model extracts more information compared to the logistic model, we compared the logarithm of the likelihood ratios (log(LR)) for SNPs under both models. The log(LR) is defined as the logarithm of the ratio of the likelihood of the single SNP model to that of the null model, which directly reflects the strength of evidence provided by the data under a particular model. We first identified the top SNPs from either the CoxPH or logistic associations within each genomic region and then calculated the log(LR) for each signal. As shown in Figure 3.11, a significant proportion of SNPs have a higher log(LR) under the CoxPH model, specifically 88.89%, 55.56%, and 61.11% in the AA, COA, and AOA analyses, respectively. This highlights the advantages of conducting time-to-event (TTE) analysis.

To illustrate the CoxPH association results in regions containing different types of signals, we plot 4 examples in Figure 3.12. Panels (b) and (d) highlight regions with shared signals between AOA and COA, while panel (a) shows COA-specific loci, and panel (c) shows AOA-specific loci. Additionally, we selected two highly significant SNPs from region 1q21.3 (COA-specific) and 12q13.11 (AOA-specific) and plotted Kaplan-Meier (KM) curves by genotype (see Figure 3.13). For the SNP rs12123821 in 1q21.3, the survival probabilities of different genotypes begin to diverge at a very young age, continuing until around age 20, after which the hazards do not appear to differ. For the SNP rs11168252 in 12q13.11, the survival probabilities of different genotypes remain similar before age 20, only beginning to diverge

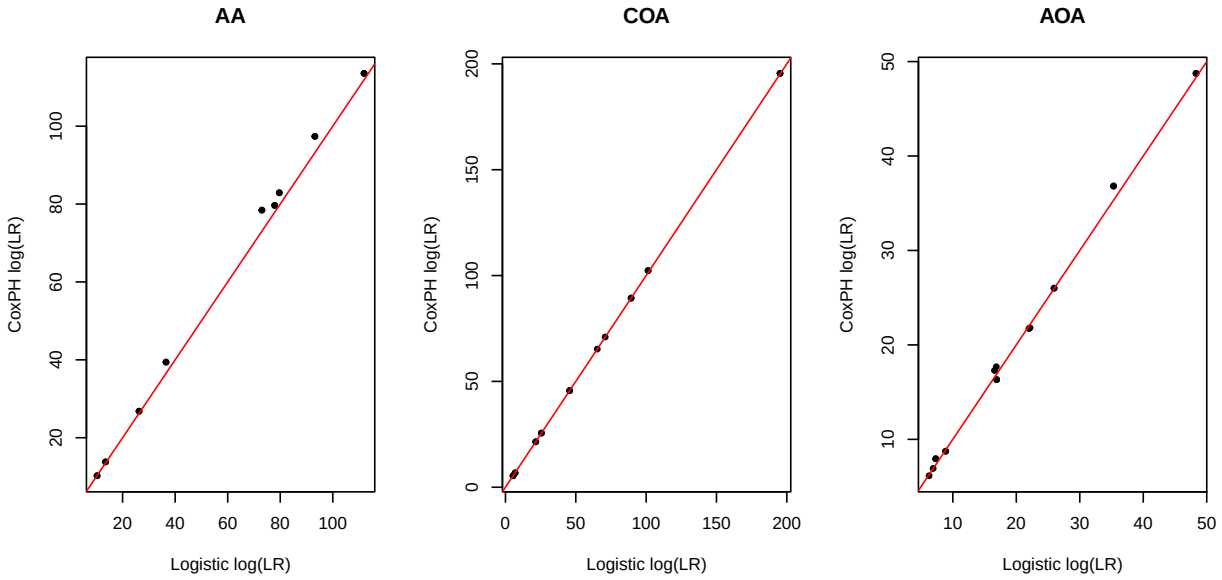

Figure 3.11: Log-likelihood ratios of top SNPs selected from AA, COA and AOA analysis under logistic regression model and CoxPH model. The red solid line indicates the line of $x = y$.

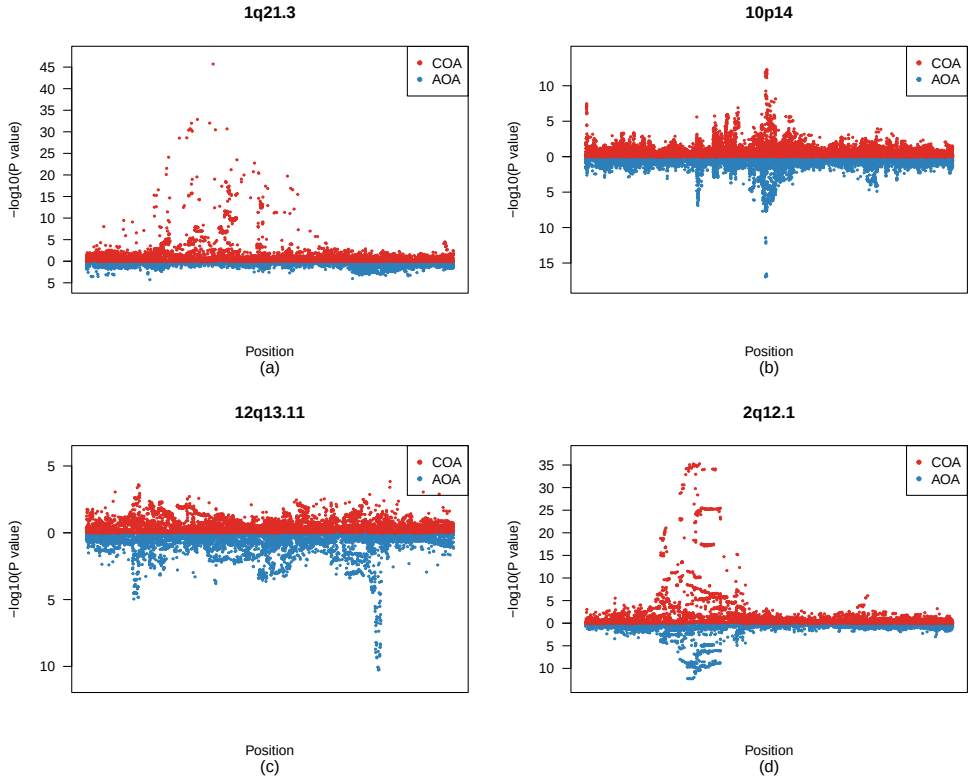

Figure 3.12: CoxPH AOA/COA GWAS results in 4 different regions.

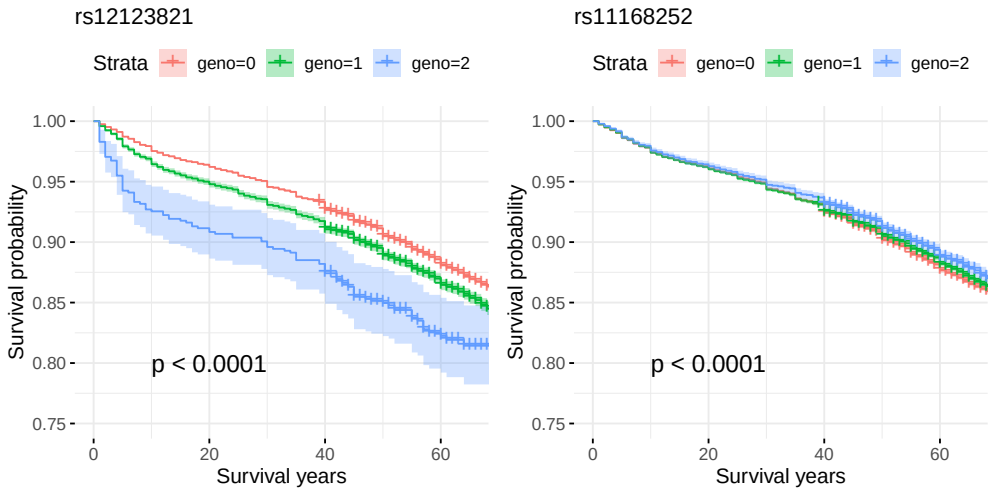

Figure 3.13: Kaplan-Meier plots at two SNPs by genotype. rs12123821 is in COA specific region 1q21.3 and rs11168252 is in AOA specific region 12q13.11.

afterward. These results suggest that the PH assumption doesn't hold for the two SNPs, and it is possible that the assumption may also fail for some other SNPs as well.

Given this consideration, we conduct three separate fine-mapping analyses: AA, COA, and AOA. The rationale is that if a region predominantly contains SNPs specific to COA, including AOA samples in the analysis could introduce noise rather than enhance power. Conversely, if a region mainly contains shared signals for both AOA and COA, analyzing all asthma samples together would be the most powerful approach.

3.8.3 Fine-mapping results

For the fine-mapping analysis, we used the same samples and methods for creating the time-to-event phenotype and adjusted for the same covariates (sex and 10 genetic PCs) as in the association analysis described in Section 3.8.1. Again, we focused our analysis on the 8 genomic regions: 1q21.3, 2q12.1, 2q22.3, 10p14, 11q13.5, 12q13.11, 15q22.2 and 17q12. To reduce computational burden, we only include SNPs within $\pm$ 250 kb of the top signal for each region. To run CoxPH-SuSiE, we use the Laplace Bayes Factor as described in Section 3.4.2 and set the number of single effect vectors to $L = 10$. The convergence criterion was

set to 0.001, with a maximum of 10 iterations. For the CoxPH-SuSiE credible sets (CSs), we reported CSs with a coverage of 0.95 and minimal absolute correlations within the CS greater than 0.5.

The fine-mapping results based on AA, COA and AOA samples are available in Figure 3.14 and 3.15. We can see the top signals are all included in the CSs if CoxPH-SuSiE output any. Additionally, the fine-mapping results across AA, COA and AOA samples are generally consistent with one another.

For each region, we then focus on the most powerful analysis, the one which generated the smallest p-value for association among the AA, COA and AOA samples. We call them the preferred analysis for each region, and we summarised the CoxPH-SuSiE credible sets of the preferred analysis in Table 3.3. From Figure 3.14 and 3.15, along with Table 3.3, we can see 4 regions contain more than one CS, suggesting the presence of multiple causal SNPs within these regions. The most notable region is 10p14, which has 4 CSs. Interestingly, no nearby protein-coding genes were identified around the top SNPs within the 10p14 CSs.

Additionally, we observed that in regions with multiple CSs, one CS typically harbors the strongest signals, while the other CSs include SNPs with less significant p-values, such as in region 11q13.5. Some of these SNPs do not even reach the GWAS significance threshold, as seen in region 2q12.1. This may occur when a region contains more than one causal SNP, say two, with one having a larger effect size and the other having a smaller effect. When analyzing SNPs individually, the one with the smaller effect may not achieve GWAS significance. However, when conditioning on the SNP with the larger effect (thus reducing noise), the weaker causal SNP can become more significant. To test this hypothesis, we performed conditional analysis in regions 11q13.5 and 2q12.1, where we conditioned on the top SNP when computing the association statistics of other SNPs in the region. The conditional p-values are summarized in Figure 3.16. After conditioning on the top SNP, the SNPs in the other CSs became the top signals in the region. These findings support our hypothesis and

further validate the CoxPH-SuSiE results.

We further examined the CoxPH-SuSiE CS summary in Table 3.3. Most of the top variants in the CSs are located near multiple protein-coding genes, except for those in region 10p14. Among these variants, rs61816761 is particularly interesting. This variant is within the coding region of the FLG gene, resulting in a premature stop codon, and is also located within 2KB upstream of the FLG-AS1 gene. ClinVar [Landrum et al., 2016] reports this variant as a pathogenic or likely pathogenic mutation associated with Dermatitis and Ichthyosis vulgaris [Churnosov et al., 2022, Sun et al., 2022, Smieszek et al., 2020, Smith et al., 2006]. Additionally, FLG is recognized as a susceptibility gene for asthma and related traits [Vercelli, 2008].

Other notable variants include rs72823641, rs11071559, rs4795399, and rs56389811, which are located within the intronic regions of the IL1RL1, RORA, GSDMB, and HDAC7 genes, respectively. For SNP rs11236797, Nasrallah et al. [2020] found that its polymorphisms were significantly associated with both basal and stimulation-driven GARP expression on $CD4^+, CD127^-, CD25^+$ regulatory T cells, and further demonstrated that this variant can drive GARP expression, identifying GARP as a potential target for immune-mediated disease therapy. The significance of the RORA SNP rs11071559 has also been replicated in multiple studies [Cai et al., 2018, Li et al., 2013, Hirota et al., 2011].

As for the nearby genes, HDAC7 is a histone deacetylase involved in transcriptional regulation. This gene plays a crucial role in the function of regulatory T cell [Axisa et al., 2022]. It has also been implicated in asthma and allergic diseases, potentially through epigenetic modifications [Morin et al., 2023]. All above biological evidence supports the validity of the CoxPH-SuSiE results. However, the role of the variants in other credible sets (CSs) with less significant p-values remains unclear.

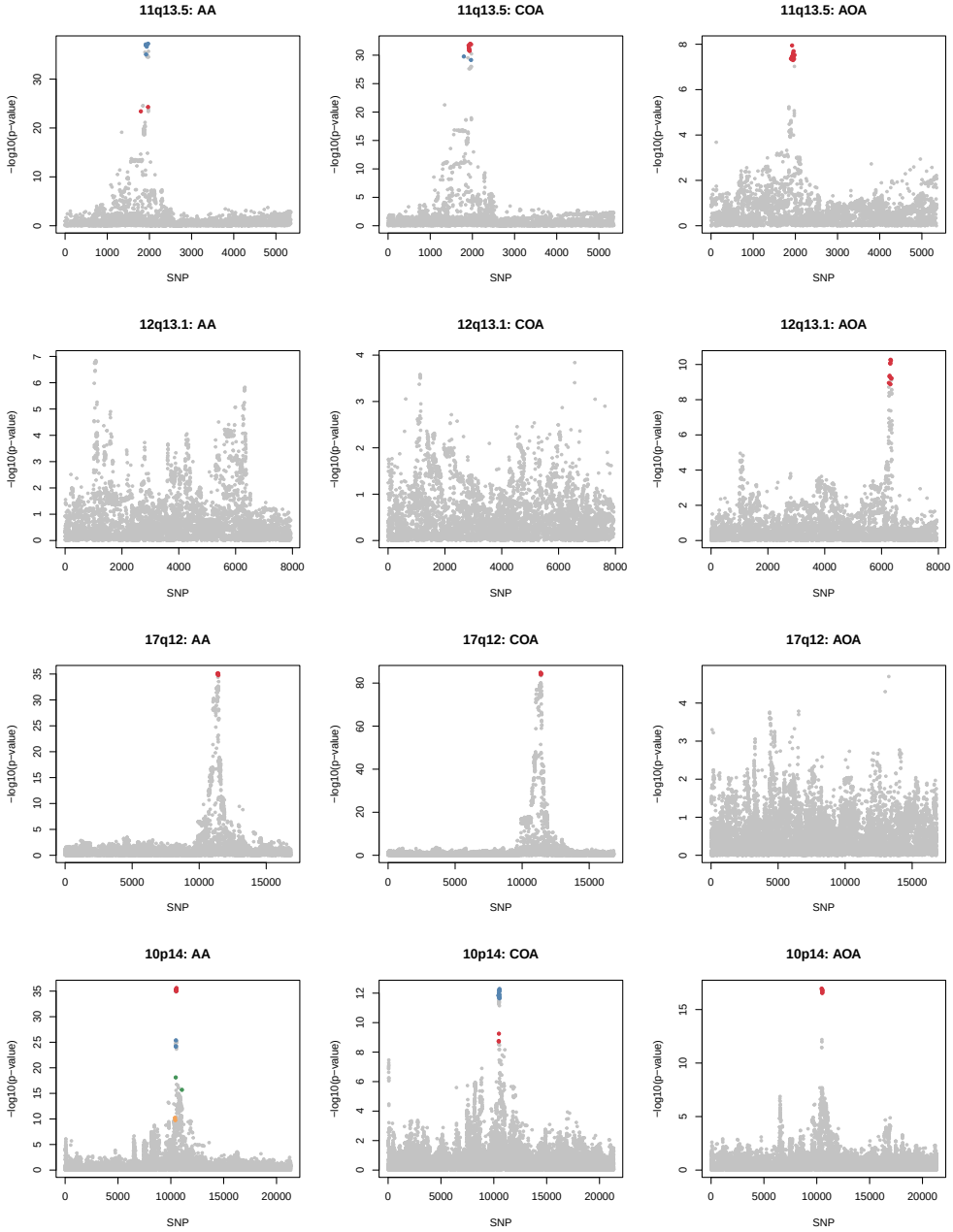

Figure 3.14: CoxPH-SuSiE results for region 11q13.5, 12q13.1, 17q12, 10p14. Each dot represents the p-value of single SNP association based on CoxPH regression. The colored dots represent SNPs in CoxPH-SuSiE CSs and different colors represent different CSs.

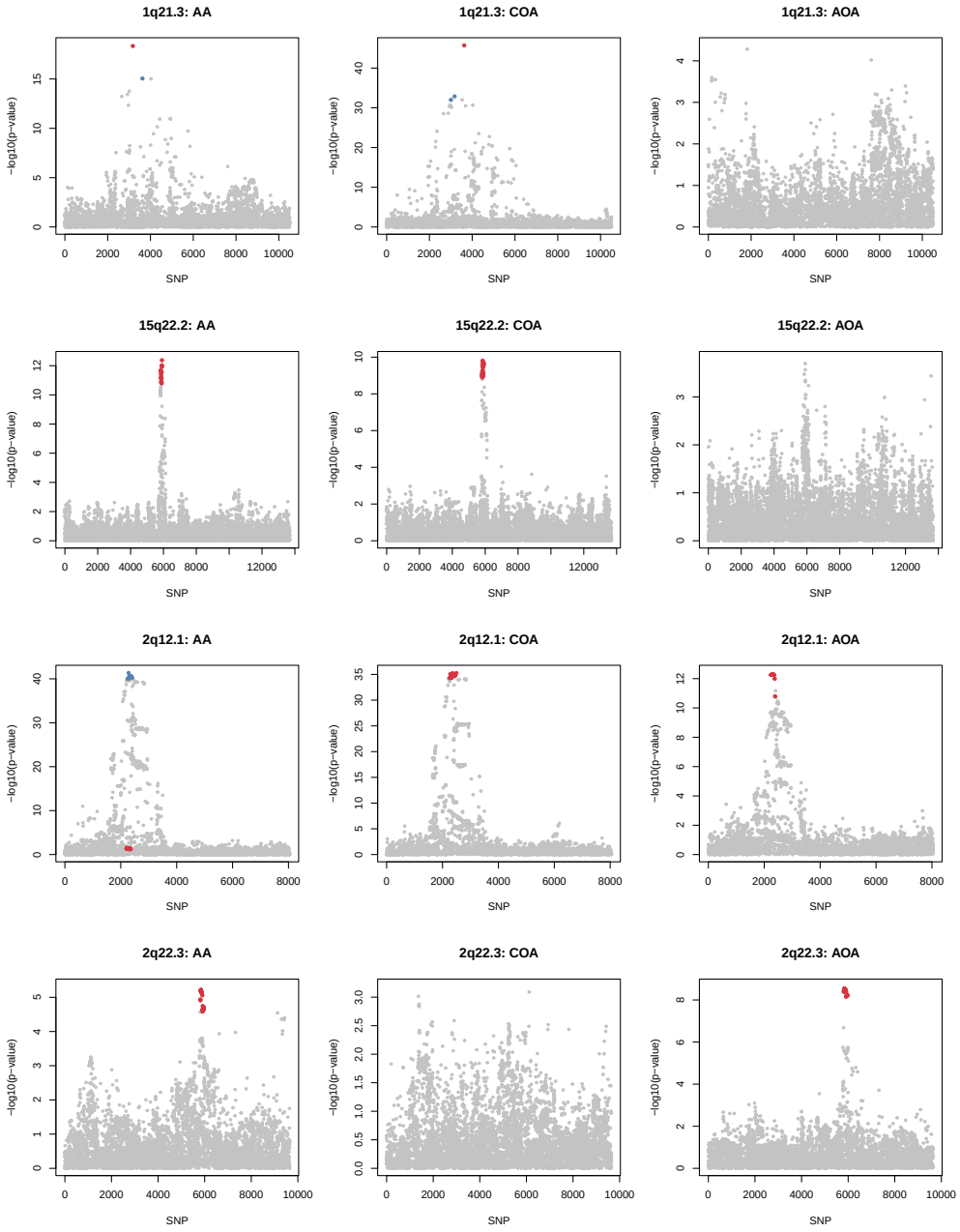

Figure 3.15: CoxPH-SuSiE results for region 1q21.3, 15q22.2, 2q12.1 and 2q22.3. Each dot represents the p-value of single SNP association based on CoxPH regression. The colored dots represent SNPs in CoxPH-SuSiE CSs and different colors represent different CSs.

Region	Analysis	Size	min.abs.corr	Variant	$-\log_{10}$(p.value)	Nearby genes
2q22.3	AOA	32	0.93	rs7571606	8.55	TEX41
2q12.1	AA	19	0.99	rs72823641	41.32	IL1R1, IL18R1, IL18RAP
		61	0.99	rs10179654	1.52	IL1R1, IL18R1, IL18RAP
15q22.2	AA	18	0.77	rs11071559	12.37	RORA, ICE2
1q21.3	COA	2	0.85	rs12123821	32.87	HRNR, RPTN, FLG
		1	1	rs61816761	45.71	FLG, FLG2, CRNN
10p14	AA	2	0.89	rs11256016	18.12	
		2	0.97	rs72782675	10.21	
		3	0.98	rs12413578	25.37	
		10	1	rs2197415	35.61	
17q12	COA	8	1	rs4795399	84.71	GSDMB, LRRC3C, ORMDL3
12q13.11	AOA	11	0.67	rs56389811	10.27	HDAC7, SLC48A1, VDR
11q13.5	AA	10	0.9	rs11236797	37.24	LRRC32, EMSY, TSKU
		2	0.94	rs55646091	24.29	LRRC32, EMSY, TSKU

Table 3.3: CoxPH-SuSiE credible sets summary based on the preferred analysis. Size: the size of the CS; min.abs.corr: minimum absolute correlation of the CS; Variant: the most significant SNP in the CS; $-\log_{10}$(p.value): $-\log_{10}$(p.value) of the most significant SNP. Nearby genes: protein-coding genes within $\pm$ 300 kb of the variant. If more than three genes fall within this window, we report only the three closest ones.

3.9 Discussion

As the analysis of time-to-event (TTE) phenotypes becomes more prevalent in genetics, there is a growing need for method development in downstream analyses, such as fine-mapping, which is critical for identifying causal variants. Fine-mapping is typically framed

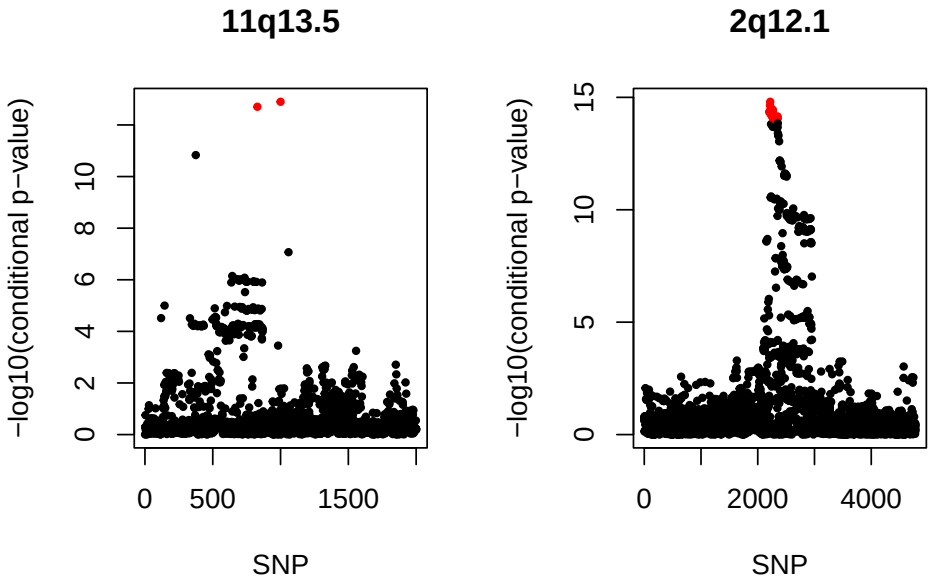

Figure 3.16: Conditional p-value plot in regions 11q13.5 and 2q12.1. All the SNP-trait association p-values are calculated by conditioning on the top SNP of the region. Red dots represent SNPs in the other CoxPH-SuSiE CS which doesn't include the top signal.

as a variable selection problem, and there are existing Bayesian variable selection methods for TTE phenotypes, including methods from Newcombe et al. [2017], Nikooienejad et al. [2020] and Komodromos et al. [2022]. However, these methods have not been applied to fine-mapping contexts where correlations among nearby variants can be extremely high, exceeding 0.99 or even equaling 1. To deal with this challenge, we build on the leading fine-mapping method, the "Sum of single effect" regression (SuSiE) developed by Wang et al. [2020], and extend it to the Cox proportional hazards (CoxPH) regression model. We call the new method for fine-mapping TTE phenotype CoxPH-SuSiE.

CoxPH-SuSiE incorporates the SuSiE parameterization for the effect vector and utilizes a similar modular fitting procedure, which we call the "Generalized Iterative Bayesian Stepwise Selection" (GIBSS). We also introduce a novel Laplace Bayes factor (BF), which outperforms the Wakefield BF in SuSiE for CoxPH model. Through two simulation scenarios—one with large variable effect sizes and a small sample size, and another with tiny variable effect sizes

but a large sample size—we demonstrate that CoxPH-SuSiE consistently delivers superior performance compared to other methods. When applied to fine-map self-reported asthma in the UK Biobank, several top SNPs within CoxPH-SuSiE credible sets (CSs) were corroborated by findings from other studies.

Despite its advantages, CoxPH-SuSiE has some limitations. First, in SuSiE, the fitting algorithm is shown to be a coordinate ascent algorithm for optimizing a variational approximation to the SuSiE posteriors. However, the theoretical understanding of the GIBSS procedure is currently insufficient, thus, the convergence cannot be guaranteed. Second, the current implementation of CoxPH-SuSiE has some speed limitation. It takes about [xx] time per iteration on a dataset with a sample size of $n = 50000$ and the number of variants $p = 1000$. The most time-consuming part is fitting the single variable CoxPH regression model pL times per iteration, where L is the pre-specified number of single effect vectors. Although CoxPH-SuSiE typically requires only a few iterations (fewer than 10) to achieve optimal performance among compared methods, this process can still be slow when applied to large datasets, such as the entire UK Biobank. Our current solution is to parallelize the computation of fitting the single variable CoxPH regression.

When applying CoxPH-SuSiE to real data, it is important to consider whether the model's proportional hazards (PH) assumption holds for the phenotype of interest. As observed in the UK Biobank's self-reported asthma data (Section 3.8), the PH assumption may fail for many SNPs. With this consideration, our approach is to conduct both pooled and separate fine-mapping analyses by dividing asthma cases into childhood-onset asthma (COA) and adult-onset asthma (AOA), using an age threshold based on Pividori et al. [2019]. However, this threshold is somewhat arbitrary, and determining a reasonable age threshold is itself could be an interesting problem in the study of asthma subtypes. For TTE phenotype fine-mapping where the PH assumption does not hold, a better approach might be to incorporate time-varying effects into the CoxPH model, such as in Ojavee et al. [2023]. This shows the

potential for further methodological advancements in survival fine-mapping analysis with time-varying effects.

CHAPTER 4

IMPROVING ESTIMATION EFFICIENCIES FOR FAMILY-BASED GWAS BY INTEGRATING LARGE EXTERNAL DATA

Abstract

Genome-wide association studies (GWASs) have identified numerous genetic variants linked to complex traits. However, the marginal associations not only reflect direct genetic effects but also contributions from nearby causal variants and the genotype of individuals' relatives. For instance, the genetic nurture effect [Kong et al., 2018], where parental genotypes influence offspring outcomes by shaping the environment provided to them. These indirect genetic effects complicate the interpretation of GWAS findings.

Family-based genetic studies offer a robust method to disentangle these effects, as the random segregation of parental genotypes and genetic differences between siblings simulate the conditions of a randomized controlled trial. However, the limited availability of family-based data results in less precise estimates due to smaller sample sizes. To address this, we introduce a calibration method that leverages external population-based data, such as GWAS summary statistics, to improve the precision of family-based estimates. By integrating biased but more precise estimates from large-scale population data, our method reduces the variance of family-based estimates. We validate this approach through theoretical analysis, simulations, and real data from the UK Biobank, demonstrating its utility in improving the accuracy of genetic effect estimates and the downstream analysis, such as Mendelian randomization.

4.1 Introduction

Genome-wide association studies (GWASs) have discovered thousands of genetic variants linked to complex human traits. Typically, a GWAS is conducted on nearly unrelated individuals to estimate the marginal association between each single-nucleotide polymorphism (SNP) and a particular phenotype. While the primary focus of GWAS is to identify the SNP's direct effect on the trait, the marginal association also reflects other factors beyond the direct effect of the variant [Veller and Coop, 2024, Davies et al., 2019].

Firstly, the effect size estimate of a SNP includes contributions from other causal SNPs in linkage disequilibrium (LD). Secondly, it captures effects of demography, such as assortative mating and residual population structure. More recently, researchers including Lee et al. 2018, Kong et al. 2018, Howe et al. 2022 have identified a non-negligible proportion of indirect genetic effects in GWAS effect size estimates, where the genotype of an individual's relatives can affect the individual's phenotype. An example of this is the genetic nurture effect, where parental genotype influences offspring outcomes by shaping the environment provided by the parents [Kong et al., 2018].

Understanding the sources of variant-trait associations is critical for several reasons. First, quantifying the direct contribution of genetic variants helps to identify and prioritize causal variants. Secondly, there is a growing interest in understanding how indirect genetic effects shape phenotypes [Balbona et al., 2021, Wu et al., 2021, Tubbs et al., 2020, Evans et al., 2019]. Thirdly, downstream analyses, such as Mendelian randomization (MR), can be biased without accounting for indirect genetic effects and other confoundings. The existence of indirect genetic effects and demographic confounders violate the independence assumption of MR [Brumpton et al., 2020, Davies et al., 2019].

Young et al. [2019] suggest that ideally, GWAS should be performed with parental or sibling genotypes as controls and using models that account for indirect genetic effects. The reasoning is that given the parental genotypes, an offspring's genotype results from the ran-

dom segregation of genetic material during meiosis, which is close to randomized controlled trial. This random segregation is uncorrelated with indirect genetic effects from relatives and other confounding effects. Similarly, genetic differences between siblings are also random and not confounded by indirect genetic effects from parents, population stratification, or assortative mating.

However, family-based genotype data is much less prevalent than population genotype data around the world. For example, the UK biobank contains approximately 500,000 individuals, whereas it only has 1,066 sets of trios (two parents and an offspring) and 22,666 sibling pairs [Bycroft et al., 2018b]. As a result, estimates derived from family data have much larger standard errors and are much less precise. In this paper, we focus on reducing the variance of estimates derived from family-based genotype data. We developed a calibration method in the regression context, where we leverage information from external large data, which may not contain family genotype information, such as population-based GWAS summary statistics. We use the biased estimates based on large population studies to improve the efficiency of family data based estimates.

This chapter is structured as follows. Section 4.2 presents the regression models to which our variance reduction method can be applied. Section 4.3 describes the approach to variance reduction in details and Section 4.4 gives theoretical results on variance reduction. Section 4.5 assesses the performance of our new method and its impact on Mendelian randomization using simulated data. Section 4.6 applies our method to analyze family data from UK Biobank. Section 4.7 discusses the promise (and limitations) of our methods.

4.2 Regression Models

Our approach is derived from single locus model, where the effects of SNPs are estimated one at a time. We assume a regression model has been built on a large external data to understand the association between an outcome of interest Y and genotype G. Typically,

we have access to the summary statistics of these associations. Meanwhile, we assume we have full access to a smaller GWAS data with family information, referred to as “internal" data. The internal data contains Y, G and family genotype F, such as the genotypes of an individual's parents or siblings. Therefore, we can build another regression model on internal data to adjust for family genotype information F. We refer such a model as the “full" model. We assume individuals in the internal data are i.i.d. observations, therefore, for an individual i at SNP locus j, the full model is expressed as follows:

$$\mu_{ij} := \mathrm{E}(Y_i | G_{ij}, F_{ij}) \tag{4.1}$$

$$g(\mu_{ij}) = \eta_{ij} = b_{0j} + b_{1j} G_{ij} + b_{2j} F_{ij} \tag{4.2}$$

where G_{ij} and F_{ij} represent the genotype of individual i and his/her family genotype at locus j. Here, Y_i denotes the phenotypic value of individual i and μ_{ij} is the conditional expectation of Y_i given (G_{ij}, F_{ij}). $\boldsymbol{b}_j := (b_{0j}, b_{1j}, b_{2j})$ represents the effect sizes, with b_{0j} being the intercept. The conditional expectation μ_{ij} is related to the linear predictor η_{ij} via a link function $g(\cdot)$. We use either identity link or logit link for $g(\cdot)$, which correspond to linear regression or logistic regression.

Depending on the specific family study design, the family genotype F_{ij} for individual i can take different information. We summarized common choices for F_{ij} in genetics literature, see Table 4.1. In a parent+offspring or mother+offspring design, an individual's genotype G_{ij} is often referred to as transmitted alleles T_{ij}, as these alleles are inherited from the parents. On the contrary, alleles that were not passed on are non-transmitted alleles NT_{ij}. Here, G_{ij}^m and G_{ij}^f denote the genotype of individual i's mother and father at locus j. In sibling design, G_{ij}^{sibs} denotes the genotype summation of all the siblings (K in total) within the family, including individual i. In scenario (1), the difference between the coefficients of G_{ij} and NT_{ij}, $b_{1j} - b_{2j}$, represents the direct genetic effect of G_{ij} on Y_i, which is also equivalent to b_{1j} in scenario (2). Scenario (3) is more flexible than (1) and (2), as it allows

for different family genotype effects from maternal side and paternal side. For (5) or (6) based on sibling data, b_{1j} can also be interpreted as the direct genetic effect. When using linear model on sibling data, though the latent effect model is more popular, we can always replace f_j by the measurable G_j^{sibs} to identify the same direct effect, see Supplementary C.1 for more discussion.

	Study design	F_{ij}	References
(1)	Parents + offspring	NT_{ij}	Kong et al. [2018]
(2)	Parents + offspring	$G_{ij}^{pa} = G_{ij}^{m} + G_{ij}^{f}$	Young et al. [2022]
(3)	Parents + offspring	(G_{ij}^{m}, G_{ij}^{f})	Brumpton et al. [2020], Young et al. [2022], Wu et al. [2021]
(4)	Mother + offspring	G_{ij}^{m}	Evans et al. [2019]
(5)	Siblings	Latent f_j shared by siblings	Brumpton et al. [2020]
(6)	Siblings	$G_j^{sibs} = \sum_{k=1}^{K} G_{kj}$	Howe et al. [2022]

Table 4.1: Common choices for family genotype information in genetic literature.

4.3 Method for Variance Reduction

In this section, we describe our new method for variance reduction in details. The central idea is to incorporate regression analysis results that use partial covariate information. Specifically, we utilize the information from a reduced model, which excludes the family genotype F in both the internal and external datasets. For an individual i at SNP j, the reduced model we consider is:

$$\tilde{\mu}_{ij} := \mathrm{E}(Y_i|G_{ij}) \tag{4.3}$$

$$g(\tilde{\mu}_{ij}) = \tilde{\eta}_{ij} = \alpha_{0j} + \alpha_{1j} G_{ij}, \tag{4.4}$$

where $\tilde{\mu}_{ij}$ is the conditional expectation of Y_i given G_{ij}. $\boldsymbol{\alpha}_j := (\alpha_{0j}, \alpha_{1j})$ contains the intercept and the effect size of G_{ij}. We specify the same link function $g(\cdot)$ for both the full model (4.2) and the reduced model. Again, the reduced model (4.4) has already been built on external data in most cases, and we have access to the summary statistics. Therefore, we only need to fit the full model and the reduced model on internal data.

For a SNP j, let τ_j represent the quantity of interest. For instance, the direct genetic effect $\tau_j = b_{1j} - b_{2j}$ in scenario (1) and $\tau_j = b_{1j}$ in other cases listed in Table (4.1). We can fit the full model (4.2) on internal data and obtain an estimate $\hat{\tau}_j$, which we refer to as the raw estimator. Our goal is to construct a new estimator for τ_j, which has smaller (asymptotic) variance than the raw estimator $\hat{\tau}_j$. Our approach for constructing the new estimator is closely aligned with the methods described in Chen and Chen [2000]. Our method makes the following assumption:

Assumption 1. *Both the families in the internal data and unrelated individuals in the external data are randomly sampled from the same population.*

Under this assumption, for a single SNP j, we consider a class of estimators of the following form:

$$\{\hat{\tau}_j - \lambda(\hat{\alpha}_{1j} - \hat{\alpha}'_{1j}), \lambda \in \mathbb{R}\}, \tag{4.5}$$

where $\hat{\alpha}_{1j}$ and $\hat{\alpha}'_{1j}$ denote the estimates of α_{1j} by fitting the reduced model on internal data and external data. We solve for an optimal λ^* such that the resulting estimator has the lowest asymptotic variance in this class. We call such an estimator the calibrated estimator for SNP j, denoted as $\tilde{\tau}_j$. The term $(\hat{\alpha}_{1j} - \hat{\alpha}'_{1j})$ helps to reduce the variance as it is correlated with $\hat{\tau}_j$. Furthermore, the calibrated estimator remains unbiased because $\hat{\alpha}_{1j} - \hat{\alpha}'_{1j}$ has mean 0.

To find the optimal λ^*, we follow the strategy in Chen and Chen [2000] by first finding the joint asymptotic distribution of $(\hat{\tau} - \tau^*, \hat{\alpha}_1 - \hat{\alpha}'_1)$. Unlike the models discussed in Chen and Chen [2000], our internal data can contain correlated individuals who are from the same family. When the number of families go to infinity, we can still show that the joint distribution of $(\hat{\tau} - \tau^*, \hat{\alpha}_1 - \hat{\alpha}'_1)$ is multivariate normal (we omit index j here for notation

convenience):

$$\sqrt{n}\begin{pmatrix}\hat{\tau}-\tau^* \\ \hat{\alpha}_1-\hat{\alpha}_1'\end{pmatrix} \to \mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{bmatrix}v_{11} & v_{12}\\ v_{21} & v_{22}\end{bmatrix}\right), \tag{4.6}$$

where τ^* denotes the true value of τ. The derivation is available in Supplementary C.2. $v_{11}, v_{12}, v_{21}, v_{22}$ constitute the variance-covariance matrix. They are all scalar values, which can be estimated from data, and we denote the estimates as $\hat{v}_{11}, \hat{v}_{12}, \hat{v}_{21}, \hat{v}_{22}$. The conditional asymptotic distribution of $\hat{\tau}-\tau^*|\hat{\alpha}_1-\hat{\alpha}_1'$ is also normal:

$$\sqrt{n}(\hat{\tau}-\tau^*)|\sqrt{n}(\hat{\alpha}_1-\hat{\alpha}_1') \sim \mathcal{N}(\sqrt{n}v_{12}v_{22}^{-1}(\hat{\alpha}_1-\hat{\alpha}_1), v_{11}-v_{12}v_{22}^{-1}v_{21}). \tag{4.7}$$

Equating $\sqrt{n}(\hat{\tau}-\hat{\tau}^*)$ to its estimated conditional mean, we obtain the calibrated estimator,

$$\hat{\tau}-\tau^* = \hat{v}_{12}\hat{v}_{22}^{-1}(\hat{\alpha}_1-\hat{\alpha}_1') \tag{4.8}$$

$$\tilde{\tau} = \hat{\tau}-\hat{v}_{12}\hat{v}_{22}^{-1}(\hat{\alpha}_1-\hat{\alpha}_1'). \tag{4.9}$$

Therefore, we set λ to $\hat{v}_{12}\hat{v}_{22}^{-1}$ where $\hat{v}_{12}$ and $\hat{v}_{22}$ are estimated from data, see Supplementary C.3 for more details. Chen and Chen [2000] claimed that the estimator found by this approach attains highest asymptotic efficiency within this class of estimators.

4.4 Theoretical Variance Reduction for the Calibrated Estimator

Denote the sample size of external data as N and internal data as n. To understand the amount of variance reduction, we derived theoretical results for a single SNP under linear regression models. Define the variance reduction comparing the variance of calibrated estimator and that of the raw estimator as $\text{VR} := 1-\text{var}(\tilde{\tau})/\text{var}(\hat{\tau})$, by plugging (4.6) and

(4.9), we have:

$$\mathrm{VR} = \frac{v_{12}^2 v_{22}^{-1}}{v_{11}}. \tag{4.10}$$

When there is no sample overlapping between the external and internal data, under parent+offspring design with scenario (1) in Table 4.1:

$$\mathrm{VR} \approx \frac{\sigma^2(1-r)}{2\left(\sigma^2 + b_2^2(1-r^2)\right)\left(1+\frac{n}{N}\right)} \tag{4.11}$$

where σ^2 is the variance of Y not explained by the single SNP. b_2 is the coefficient of F in the full model and r is the correlation between G and F. Supplementary C.4 shows the derivation in this case. For complex human traits, σ^2 is usually much larger than b_2. And under random mating, r is close to 0. When $n/N \to 0, \mathrm{VR} \to 50\%$. Therefore, the maximum variance reduction is 50% under parent+offspring design for a single SNP. This is essentially equivalent to double the sample size of internal data.

Under sibling design with scenario (5), if there are two siblings for each family,

$$\mathrm{VR} \approx \frac{(1-\pi)}{2}\frac{1-\rho}{1+\frac{n}{N}+\pi\rho}, \tag{4.12}$$

where $\pi := \mathrm{cor}(G, G^{sib})$ is the genetic correlation between two siblings at a loci and $\rho := \mathrm{cor}(Y, Y^{sib})$ is the phenotypic correlation between two siblings. Again, under random mating and random segregation, π is close to 0.5. If the phenotypic correlation is 0, the maximum variance reduction is 25%. If the phenotypic correlation is 0.5, the maximum variance reduction is 10%.

Note that to derive these theoretical results, we assume the genotype G and the noise terms (both in full and reduced models) are independent. Although this assumption may not satisfy in reality, these theoretical results are served as references for simulations and real data analysis.

4.5 Simulation

In this section, we conduct simulation to assess the performance of our calibrated estimator. We first assess the variance reduction for single SNPs. Then, we evaluate how variants with reduced variance might influence Mendelian randomization analysis.

4.5.1 Data Generation

To generate family genotype data, we apply the following data generation procedure. For i^{th} family,

1. Generate maternal and paternal genotype $(\boldsymbol{G}_i^f, \boldsymbol{G}_i^m)$ from Binomial distribution $\boldsymbol{G}_i^f, \boldsymbol{G}_i^m \sim \text{Bin}(2, \boldsymbol{f})$, where $\boldsymbol{f}$ is a vector of length p, containing the allele frequencies of p SNPs.

2. Generate parents phenotype using the following model:

$$\boldsymbol{b}^{pa} \sim N_p(\boldsymbol{0}, \sigma^2 \boldsymbol{I}) \tag{4.13}$$

$$Y_i^f = (\boldsymbol{G}_i^f)^T \boldsymbol{b}^{pa} + \epsilon_i^f, \ \epsilon_i^f \sim N(0, \sigma^2) \tag{4.14}$$

$$Y_i^m = (\boldsymbol{G}_i^m)^T \boldsymbol{b}^{pa} + \epsilon_i^m, \ \epsilon_i^m \sim N(0, \sigma^2) \tag{4.15}$$

   where $\boldsymbol{b}^{pa}$ denotes the causal effects of p SNPs in parents' generation and $(\epsilon_i^f, \epsilon_i^m)$ denote the noise in parental phenotypes. (Y_i^f, Y_i^m) represents some arbitrary parental phenotype which could affect offspring outcomes.

3. Generate children's genotype $\boldsymbol{G}_i$ based on Mendelian law of inheritance. That is, at each locus, we randomly select one allele from mother and one from father to create the child genotype.

4. Generate exposure trait and outcome trait in children's generation using the following

model:

$$\boldsymbol{b} \sim N_p(\mathbf{0}, \sigma^2 \boldsymbol{I}) \tag{4.16}$$

$$Y_{i1} = \boldsymbol{G}_i^T \boldsymbol{b} + r_1(Y_i^m + Y_i^f) + \epsilon_{i1} \ \epsilon_{i1} \sim N(0, \sigma^2) \tag{4.17}$$

$$Y_{i2} = \beta Y_{i1} + r_2(Y_i^m + Y_i^f) + \epsilon_{i2}, \ \epsilon_{i2} \sim N(0, \sigma^2) \tag{4.18}$$

where $\boldsymbol{b}$ denotes the casual effects of p SNPs in children generation, and Y_{i1}, Y_{i2} denote the exposure and outcome trait for the child in family i. ϵ_{i1} and ϵ_{i2} denote the noise term of exposure and outcome. β is the causal effect of exposure on outcome. $r_1, r_2 \in [0, 1)$ control the level of indirect genetic effects on exposure trait and outcome trait. To create binary exposure and outcome traits, we simply binarize continuous Y_{i1} and Y_{i2} at a threshold where the resulting cases and controls are balanced.

The above procedure generates trio data, mother-father-child for each family. To create sibling data, we simply repeat Step 3 to generate multiple siblings and then discard the parents for each family. In our simulation, we set the number of SNPs $p = 10$ and $\boldsymbol{f} = (0.61, 0.78, 0.15, 0.41, 0.37, 0.46, 0.51, 0.21, 0.64, 0.19)$. The sample size for external data is $N = 10000$ and internal data is $n = 1000$. For trio design, we run simulations for different values of r_1 and r_2, which control the contribution of indirect genetic effects in offspring phenotype. σ^2 is set to 1 all the time. For sibling design, we simulate two siblings for each family. Addition to r_1 and r_2, we also vary the size of σ^2 in offspring generation, which affects phenotypic correlation between siblings. For each simulation setting, we run 500 replicates. Across the replicates, $\boldsymbol{b}^{pa}$ and $\boldsymbol{b}$ are generated once and kept the same.

4.5.2 Simulation results

First, we assess variance reduction empirically for both trio data and sibling data across 500 simulation replicates. For each SNP, we compute the empirical variance of the raw

estimator and the calibrated estimator for its association with the exposure trait Y_1 using linear regression. Then, we compute the empirical variance reduction, see Figure 4.1. We can see from Figure 4.1, the empirical variance reduction results are consistent with the theoretical derivations for both trio and sibling data. For trio data, when the sample size ratio between internal data and external data is 0.1, the amount of theoretical variance reduction is around 0.45. In Figure 4.1 (a), the empirical variance reduction is also around 0.45 for all SNPs and is similar across different levels of indirect genetic effect. For sibling data, as phenotypic correlation increases, the empirical variance reduction decreases. The mean of empirical variance reduction is around 20% when the phenotypic correlation is near zero, and about 10% when the phenotypic correlation is near 0.4. The theoretical values are consistent with the empirical values, after accounting for the fact the actual sample size ratio between internal data and external data is 0.1 in simulation.

We then conducted Mendelian randomization (MR) on the simulated data using two MR methods: the Inverse-Variance Weighted Method (IVW) from Lawlor et al. [2008] and MR.RAPS by Zhao et al. [2020], which provides robust inference for MR with many weak instruments. We compare the MR results based on three different sets of summary statistics: (1) calibrated estimators with corresponding standard errors, (2) raw estimators with corresponding standard errors, and (3) external GWAS summary statistics. The first two sets of summary statistics adjust for family genotype information using internal data, while the external GWAS summary statistics do not. As a result, the MR estimates based on the first two sets of summary statistics are unbiased, whereas those based on the external GWAS summary statistics are biased when the indirect genetic effect on the outcome trait is non-zero. For each single SNP, the calibrated estimator has a smaller variance compared to the raw estimator, therefore, MR estimates based on the calibrated estimators should have smaller variance than those based on the raw estimators. The MR results for both trio data and sibling data are available in Figure 4.2.

From plot (a)-(d) in Figure 4.2, we can see when $r_2 = 0$, meaning there is no indirect genetic effect on outcome trait, the MR estimates based on external GWAS summary statistics is unbiased. As r_2 increases, the bias in MR becomes more obvious. In trio data, the reduction in the variance of MR estimates is around 40%-50% for different combinations of (r_1, r_2). In sibling data, the reduction in the variance of MR estimates depends on the phenotypic correlations of exposure and outcome traits. The maximum variance reduction is achieved when phenotypic correlations of exposure and outcome traits are both close to 0, which is nearly 25%, see plot (f). When phenotypic correlations are very high, the variance reduction in MR estimates can be less than 5%. In sibling MR plot of IVW, plot (c) of Figure 4.2, we can see the mean of MR estimates computed from calibrated estimators and raw estimators are biased towards 0 when $(r_1, r_2, \sigma^2) = (0, 0, 6)$. This is because the noise level σ^2 is too high, making the instruments too weak, and IVW will shrink the point estimates towards 0. We don't see this behavior in MR.RAPS, as it is a robust method for weak instruments.

4.6 Data Analysis

We applied our new method to UK biobank family genotype data. The UK Biobank is a very large, population-based prospective study, collected detailed phenotype and genotype data from over 500,000 participants in the United Kingdom, with ages between 40 and 69 at time of recruitment Sudlow et al. [2015], Bycroft et al. [2018b]. We applied our variance calibration approach to five different phenotypes: body mass index (BMI, data field 21001), Diastolic blood pressure (DBP, data field 4079), Systolic blood pressure (SBP, data field 4080), diabetes (data field 2443) and education years (EduYrs, converted from data field 6138 following Howe et al. [2023]). We also performed Mendelian Randomization for BMI on diabetes, BMI on EduYrs and height on EduYrs.

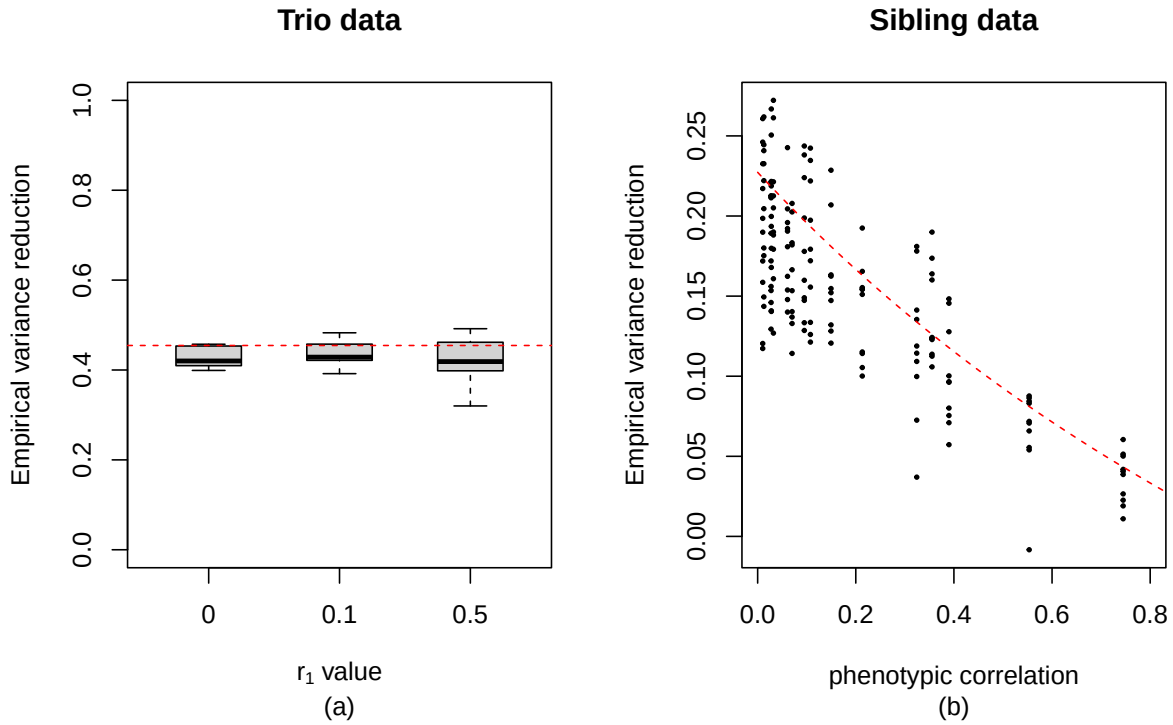

Figure 4.1: Simulation results for empirical variance reduction of 10 SNPs based on linear regression model. (a): empirical variance reduction across different r_1 values on trio data. (b): empirical variance reduction across different pheotypic correlation based on sibling data. Each black dot represents empirical variance reduction of a single SNP. The red dashed lines on both panels represent the value of theoretical variance reduction.

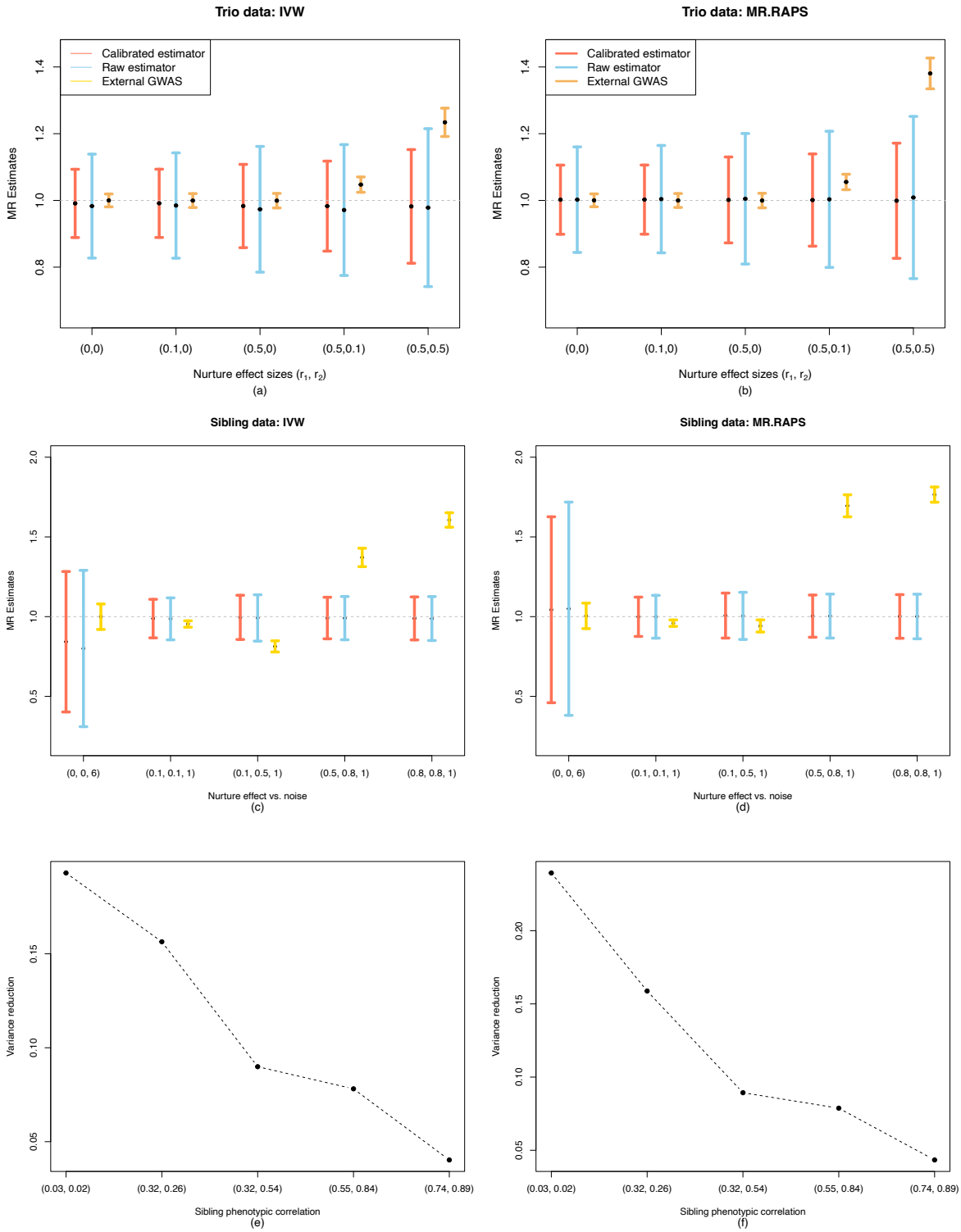

Figure 4.2: Mendelian Randomization results on simulated data. The true causal effect in simulation is $\beta = 1$. Plot (a),(b) contain results on trio data and plot (c)-(f) are results on sibling data. In plot (a)-(d), the black dots are the mean of causal effect estimates across 500 replicates and the error bars are 1.96 times the standard deviation of the causal estimates. The x-axes of plot (a) and (b) are different combinations of (r_1, r_2) values. The x-axes of plot (c), (d) are different combinations of (r_1, r_2, σ^2) values and the x-axes of (e), (f) are the resulting average phenotypic correlations between sibling pairs for the exposure and the outcome trait from corresponding (r_1, r_2, σ^2).

4.6.1 Data preprocessing

To get family data from UK biobank, we first filter individuals based on kinship value and identical-by-descent (IBD) value, suggested by Manichaikul et al. [2010], Bycroft et al. [2018b]. Parent-offspring pairs and full sibling pairs both have kinship value between $\frac{1}{2^{5/2}}$ and $\frac{1}{2^{3/2}}$. Parent-offspring pairs have IBD0 value ≤ 0.0012 and full sibling pairs have IBD0 value in the range of $(0.0012, 0.365)$. These criteria result in 6263 parent-offspring pairs and 22646 sibling pairs. To divide these relative pairs into families, we implemented a graph search algorithm, breath-first search (BFS) to find connected components in the data. In total, there are 4958 families in parent-offspring data and 20128 families in sibling data. For trio data, we use families with three individuals, father-mother-offspring and there are 1172 such families. To create external data, we remove individuals who have at least one relative in the cohort based on kinship (samples with a value other than zero in data field 22021).

4.6.2 Result

For assessing variance reduction on the real data, we focused on independent SNPs associated with BMI. We used 1000 genomes [Consortium et al., 2015] as the LD reference panel and selected SNPs with p-value smaller than 0.01 based on summary statistics from GIANT consortium [Locke et al., 2015, Wood et al., 2014] (a study of more than 250,000 European descents for anthropometric traits). This results in 783 nearly independent SNPs. Then we apply our variance calibration approach to these SNPs based on linear regression and Figure 4.3 summarizes the results. From (a) and (c), we can see the difference between calibrated estimator and raw estimator are centered around zero. This is expected as our calibrated estimator is unbiased. Both trio and sibling real data variance reduction are consistent with our simulation results and theoretical results, see Figure 4.3 (b)(d). For trio data, the average estimated variance reduction across all traits are around 50%. For sibling data, the average estimated variance reduction varies in the range of 15% and 25%, depending on the

phenotypic correlations.

Next, we performed Mendelian randomization analysis. We began with LD clumping, where we used 1000 genomes as the LD reference panel. We selected SNPs based on summary statistics of GIANT consortium using different p-value thresholds, that is, p-value smaller than $1e^{-2}, 1e^{-5}$ and $1e^{-8}$. Although SNPs with p-values larger than $1e^{-5}$ are considered relatively weak instruments, there are MR methods that can handle weak instrumental variables. They allow us to select more SNPs to perform MR, increasing the robustness of the results potentially. For BMI as the exposure trait, the different selection thresholds result in 783, 180 and 57 SNPs. For height as the exposure trait, the selection results in 1150, 564 and 346 SNPs, see the x-axes of Figure 4.4. To obtain the summary statistics on selected SNPs, we use the linear regression model for all the traits (BMI, height, EduYrs and diabetes).

We experimented with three MR methods, Inverse-Variance Weighted (IVW), MR.RAPS (mle option) and MR.RAPS (shrinkage option). The results are available in Figure 4.4. For trio data results of BMI on diabetes, we can see obvious variance reduction (27.25% - 85.93%) comparing the standard error resulted from calibrated estimators and from raw estimators, plot (b)-(c) of Figure 4.4. However, no variance reduction was observed in the IVW result, Figure 4.4 (a). Besides, there are two additional concerns with these results. First, there are noticeable differences in the point estimates derived from calibrated estimators compared to those from raw estimators, plot (a) - (c) of Figure 4.4. Second, the causal effect estimate of BMI on diabetes by MR.RAPS (mle) and MR.RAPS (shrinkage) is larger than those reported in other MR studies [Corbin et al., 2016, Wang et al., 2021]. These discrepancies are not fully understood yet and may be due to the small sample size of the trio data, resulting in very noisy estimates.

For sibling MR results, Figure 4.4 (d) - (i), the variance reduction comparing the variance of the causal effect estimate derived from calibrated estimators and raw estimators (estimated

by IVW) ranges from 5.98% to 17.90% for BMI on EduYrs and 5.83% - 8.11% for height on EduYrs. The other two MR.RAPS methods estimated the variance reduction to be 21.14% - 30.19% for BMI on EduYrs and 12.72% - 21.26% for height on EduYrs. The MR point estimates based on calibrated estimators and those based on raw estimators are not significantly different from zero. These results make sense intuitively as changing one's BMI or height are not likely to cause a change in one's education years. On the contrary, the MR point estimates based on external GWAS summary statistics all significantly differ from zero. We believe this shows the bias in MR rather than the truth since EduYrs is a phenotype with a stronger genetic nurture effect. When performing MR without adjusting for indirect genetic effect on the outcome trait, this would introduce bias to the MR results.

4.7 Discussion

Growing evidence suggests that a non-negligible proportion of indirect genetic effects contribute to complex human traits, such as educational attainment and cognition [Kong et al., 2018, Warrington et al., 2018, Selzam et al., 2019, Howe et al., 2022]. These indirect genetic effects are of particular interest for estimating parental influences and understanding the broader genetic architecture of complex traits. Furthermore, Mendelian randomization analyses may be biased if indirect genetic effects in outcome traits are not accounted for. Conducting GWAS using family genotype data is a promising approach to address these issues. However, this approach usually lacks of statistical power because large samples of genotyped trios or siblings are rare. To address this limitation, we have developed a new method that improves the estimation efficiency of family-based GWAS by integrating information from standard GWAS summary statistics.

Our approach to variance reduction is very general and does not depend on specific models. It only requires to fit a full and a reduced model on internal data, in addition to the summary statistics from external data. This flexibility allows our method to be applied

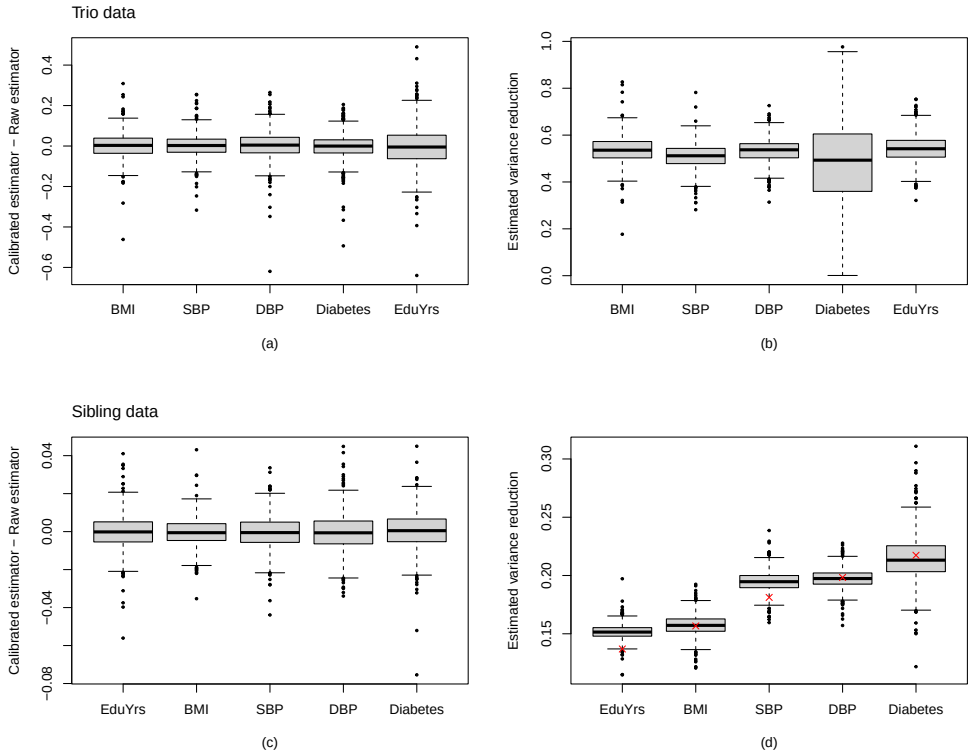

Figure 4.3: Variance reduction of five UK Biobank traits on 783 SNPs using linear regression. All phenotypes were standardized and we adjusted for 10 genetic PCs, sex, age and age-squared. Plot (a), (b) show results on trio data and (c), (d) show results on sibling data. (a) and (c) show the difference between calibrated estimator and raw estimator. (b) and (d) are boxplots of estimated variance reduction for all SNPs. Red crosses on plot (d) are theoretical values where estimated phenotypic correlations were plugged in.

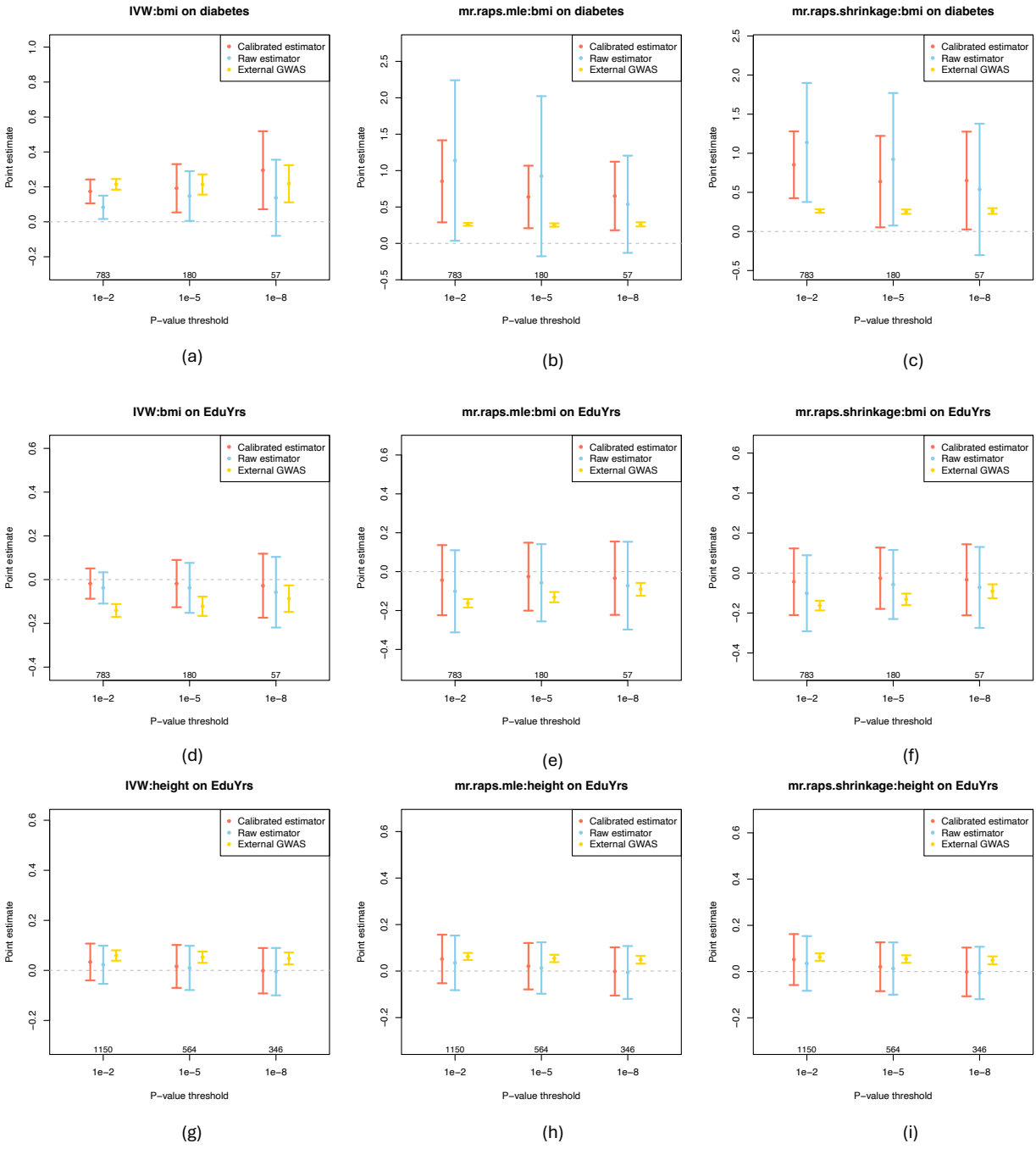

Figure 4.4: Mendelian randomization results on UK biobank real data. Plot (a)-(c) use UKB trio data as the internal data and (d)-(i) use sibling data as the internal data. The x-axes indicate the p-value thresholds and the corresponding number of instrumental variables. The error bar indicates 1.96 times the standard error output by the corresponding MR method. IVW: Inverse-Variance Weighted, mr.raps.mle: MR.RAPS with the mle option and mr.raps.shrinkage: MR.RAPS with the shrinkage option.

across various family data designs and different covariate choices. Additionally, since our method is built on single SNP models, similar to standard GWAS, it can be effectively used for family-based GWAS. The method provides effect size estimates from a family genotype-adjusted model of the user's choice, along with calibrated variance for each SNP. Based on our theoretical results and simulation results, the maximum variance reduction for a single SNP is 50% for trio data, and 25% for sibling data. This is equivalent to increase the sample size of internal data by a factor of 2 and 1.33. Using SNPs with reduced variance also lead to variance reduction in Mendelian randomization results, which could potentially alter the conclusions drawn about the causal relationship between traits.

However, assessing the variance reduction in MR is challenging. MR results can be complicated by issues such as weak instruments, uncorrelated pleiotropy and correlated pleiotropy [Morrison et al., 2020, Richmond and Davey Smith]. Furthermore, each MR method has its own advantages and limitations, making it difficult to directly compare results from external GWAS, family-based GWAS, and those using calibrated versus uncalibrated estimators.

Although considerable amount of variance reduction can be achieved for individual SNPs, it is still constrained by the size of the available family data. For example, the UK Biobank includes only about 1,000 trios, and even with variance reduction doubling the effective sample size to around 2,000, it is still not comparable with siling data sample size, which is approximately ten times larger. As our MR results shown, causal estimates based on trio data tend to be quite noisy, making it difficult to draw reliable conclusions. Therefore, despite the smaller variance reduction, sibling data may still be preferable in practice due to its much larger sample size.

REFERENCES

Abdel Abdellaoui, Loic Yengo, Karin JH Verweij, and Peter M Visscher. 15 years of gwas discovery: realizing the promise. *The American Journal of Human Genetics*, 110(2):179–194, 2023.

François Aguet, Kaur Alasoo, Yang I Li, Alexis Battle, Hae Kyung Im, Stephen B Montgomery, and Tuuli Lappalainen. Molecular quantitative trait loci. *Nature Reviews Methods Primers*, 3(1):4, 2023.

Daniel S Araujo, Chris Nguyen, Xiaowei Hu, Anna V Mikhaylova, Chris Gignoux, Kristin Ardlie, et al. Multivariate adaptive shrinkage improves cross-population transcriptome prediction and association studies in underrepresented populations. *Human Genetics and Genomics Advances*, 4(4):100216, 2023.

Hugues Aschard, Bjarni J Vilhjálmsson, Nicolas Greliche, Pierre-Emmanuel Morange, David-Alexandre Trégouët, and Peter Kraft. Maximizing the power of principal-component analysis of correlated phenotypes in genome-wide association studies. *The American Journal of Human Genetics*, 94(5):662–676, 2014.

Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.

Pierre-Paul Axisa, Tomomi M Yoshida, Liliana E Lucca, Herbert G Kasler, Matthew R Lincoln, Giang H Pham, Dante Del Priore, Jean-Marie Carpier, Carrie L Lucas, Eric Verdin, et al. A multiple sclerosis–protective coding variant reveals an essential role for hdac7 in regulatory t cells. *Science Translational Medicine*, 14(675):eabl3651, 2022.

Jared V Balbona, Yongkang Kim, and Matthew C Keller. Estimation of parental effects using polygenic scores. *Behavior genetics*, 51(3):264–278, 2021.

Jenna Lee Ballard and Luke Jen O'Connor. Shared components of heritability across genetically correlated traits. *American Journal of Human Genetics*, 109(6):989–1006, 2022.

Alvaro N Barbeira, Owen J Melia, Yanyu Liang, Rodrigo Bonazzola, Gao Wang, Heather E Wheeler, François Aguet, Kristin G Ardlie, Xiaoquan Wen, and Hae K Im. Fine-mapping and qtl tissue-sharing information improves the reliability of causal gene identification. *Genetic Epidemiology*, 44(8):854–867, 2020.

Mathew J Barber, Lara M Mangravite, Craig L Hyde, Daniel I Chasman, Joshua D Smith, Catherine A McCarty, Xiaohui Li, Russell A Wilke, Mark J Rieder, Paul T Williams, et al. Genome-wide association of lipid-lowering response to statins in combined study populations. *PloS one*, 5(3):e9763, 2010.

Stephen Bates, Matteo Sesia, Chiara Sabatti, and Emmanuel Candès. Causal inference in genetic trio studies. *Proceedings of the National Academy of Sciences*, 117(39):24117–24126, 2020.

Anindya Bhadra, Jyotishka Datta, Nicholas G. Polson, and Brandon Willard. Lasso meets horseshoe: a survey. *Statistical Science*, 34(3):405–427, 2019.

Wenjian Bi, Lars G Fritsche, Bhramar Mukherjee, Sehee Kim, and Seunggeun Lee. A fast and accurate method for genome-wide time-to-event data analysis and its application to uk biobank. *The American Journal of Human Genetics*, 107(2):222–233, 2020.

John D Blischak, Peter Carbonetto, and Matthew Stephens. Creating and sharing reproducible research code the workflowr way [version 1; peer review: 3 approved]. *F1000Research*, 8(1749), 2019.

Marc Jan Bonder, Craig Smail, Michael J Gloudemans, Laure Frésard, David Jakubosky, Matteo D'Antonio, et al. Identification of rare and common regulatory variants in pluripotent cells using population-scale transcriptomics. *Nature Genetics*, 53(3):313–321, 2021.

Jo Bovy, David W. Hogg, and Sam T. Roweis. Extreme Deconvolution: Inferring complete distribution functions from noisy, heterogeneous and incomplete observations. *Annals of Applied Statistics*, 5(2B):1657–1677, 2011.

Ben Brumpton, Eleanor Sanderson, Karl Heilbron, Fernando Pires Hartwig, Sean Harrison, Gunnhild Åberge Vie, Yoonsu Cho, Laura D Howe, Amanda Hughes, Dorret I Boomsma, et al. Avoiding dynastic, assortative mating, and population stratification biases in mendelian randomization through within-family analyses. *Nature communications*, 11(1):1–13, 2020.

Brendan Bulik-Sullivan, Hilary K Finucane, Verneri Anttila, Alexander Gusev, Felix R Day, Po-Ru Loh, Laramie Duncan, John RB Perry, Nick Patterson, Elise B Robinson, et al. An atlas of genetic correlations across human diseases and traits. *Nature genetics*, 47(11): 1236, 2015.

Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, et al. The uk biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209, 2018a.

Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, et al. The uk biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209, 2018b.

Richard H Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208, 1995.

Tony Cai and Weidong Liu. Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Association*, 106(494):672–684, 2011.

Xulong Cai, Mali Lin, Shan Cao, Yunguang Liu, and Na Lin. The association of rar-related orphan receptor a (rora) gene polymorphisms with the risk of asthma. *Annals of human genetics*, 82(3):158–164, 2018.

Peter W Callas, Harris Pastides, and David W Hosmer. Empirical comparisons of proportional hazards, poisson, and logistic regression modeling of occupational cohort data. *American journal of industrial medicine*, 33(1):33–47, 1998.

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

Nilanjan Chatterjee, Yi-Hau Chen, Paige Maas, and Raymond J Carroll. Constrained maximum likelihood estimation for model calibration using summary-level information from external big data sources. *Journal of the American Statistical Association*, 111(513):107–117, 2016.

Yi-Hau Chen and Hung Chen. A unified approach to regression analysis under double-sampling designs. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(3):449–460, 2000.

Eric C Chi and Kenneth Lange. Stable estimation of a covariance matrix guided by nuclear norm penalties. *Computational statistics and Data Analysis*, 80:117–128, 2014.

Mikhail Churnosov, Tatyana Belyaeva, Evgeny Reshetnikov, Volodymyr Dvornyk, and Irina Ponomarenko. Polymorphisms of the filaggrin gene are associated with atopic dermatitis in the caucasian population of central russia. *Gene*, 818:146219, 2022.

Taane G Clark, Michael J Bradburn, Sharon B Love, and Douglas G Altman. Survival analysis part i: basic concepts and first analyses. *British journal of cancer*, 89(2):232–238, 2003.

Selene M Clay, Nathan Schoettler, Andrew M Goldstein, Peter Carbonetto, Matthew Dapas, Matthew C Altman, Mario G Rosasco, James E Gern, Daniel J Jackson, Hae Kyung Im, et al. Fine-mapping studies distinguish genetic risks for childhood-and adult-onset asthma in the hla region. *Genome Medicine*, 14(1):55, 2022.

Genomes Project Consortium, A Auton, LD Brooks, RM Durbin, EP Garrison, and HM Kang. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 2015.

Laura J Corbin, Rebecca C Richmond, Kaitlin H Wade, Stephen Burgess, Jack Bowden, George Davey Smith, and Nicholas J Timpson. Bmi as a modifiable risk factor for type 2 diabetes: refining and understanding causal estimates using mendelian randomization. *Diabetes*, 65(10):3002–3007, 2016.

David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

Neil M Davies, Laurence J Howe, Ben Brumpton, Alexandra Havdahl, David M Evans, and George Davey Smith. Within family mendelian randomization studies. *Human Molecular Genetics*, 28(R2):R170–R179, 2019.

Philip J Davis and Philip Rabinowitz. *Methods of numerical integration*. Courier Corporation, 2007.

Christiaan A de Leeuw, Joris M Mooij, Tom Heskes, and Danielle Posthuma. Magma: generalized gene-set analysis of gwas data. *PLoS Comput Biol*, 11(4):e1004219, 2015.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1): 1–22, 1977.

Rounak Dey, Wei Zhou, Tuomo Kiiskinen, Aki Havulinna, Amanda Elliott, Juha Karjalainen, Mitja Kurki, Ashley Qin, FinnGen, Seunggeun Lee, et al. Efficient and accurate frailty model approach for genome-wide survival association analysis in large-scale biobanks. *Nature communications*, 13(1):5437, 2022.

Thomas A DiPrete, Casper AP Burik, and Philipp D Koellinger. Genetic instrumental variable regression: Explaining socioeconomic and health outcomes in nonexperimental data. *Proceedings of the National Academy of Sciences*, 115(22):E4970–E4979, 2018.

Bradley Efron and Carl Morris. Limiting the risk of Bayes and empirical Bayes estimators—Part II: the empirical Bayes case. *Journal of the American Statistical Association*, 67(337): 130–139, 1972.

Jason Ernst and Manolis Kellis. Chromhmm: automating chromatin-state discovery and characterization. *Nature methods*, 9(3):215–216, 2012.

David M Evans, Gunn-Helen Moen, Liang-Dar Hwang, Debbie A Lawlor, and Nicole M Warrington. Elucidating the role of maternal environmental exposures on offspring health and disease using two-sample mendelian randomization. *International journal of epidemiology*, 48(3):861–875, 2019.

Jianqing Fan, Yuan Liao, and Han Liu. An overview of the estimation of large covariance and precision matrices. *Econometrics Journal*, 19(1):C1–C32, 2016. URL `https://academic.oup.com/ectj/article/19/1/C1/5056252`.

Zhanying Feng, Xianwen Ren, Zhana Duren, and Yong Wang. Human genetic variants associated with covid-19 severity are enriched in immune and epithelium regulatory networks. *Phenomics*, 2(6):389–403, 2022.

Manuel AR Ferreira and Shaun M Purcell. A multivariate test of association. *Bioinformatics*, 25(1):132–133, 2009.

Manuel AR Ferreira, Riddhima Mathur, Judith M Vonk, Agnieszka Szwajda, Ben Brumpton, Raquel Granell, Bronwyn K Brew, Vilhelmina Ullemar, Yi Lu, Yunxuan Jiang, et al. Genetic architectures of childhood-and adult-onset asthma are partly distinct. *The American Journal of Human Genetics*, 104(4):665–684, 2019.

Jason Fletcher, Yuchang Wu, Tianchang Li, and Qiongshi Lu. Interpreting polygenic score effects in sibling analysis. *bioRxiv*, 2021.

Timothée Flutre, Xiaoquan Wen, Jonathan Pritchard, and Matthew Stephens. A statistical framework for joint eqtl analysis in multiple tissues. *PLoS Genetics*, 9(5):e1003486, 2013.

Centers for Disease Control, Prevention, et al. National diabetes statistics report, 2020. *Atlanta, GA: Centers for Disease Control and Prevention, US Department of Health and Human Services*, pages 12–15, 2020.

Chris Fraley and Adrian E Raftery. Bayesian regularization for normal mixture estimation and model-based clustering. *Journal of Classification*, 24:155–181, 2007.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

Oliver Fuchs, Thomas Bahmer, Klaus F Rabe, and Erika von Mutius. Asthma transition from childhood into adulthood. *The Lancet Respiratory Medicine*, 5(3):224–234, 2017.

Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian data analysis*. CRC Press, Boca Raton, FL, third edition, 2013.

Zoubin Ghahramani and Geoffrey E Hinton. The EM algorithm for mixtures of factor analyzers. Technical report, University of Toronto, 1996.

Greg Gibson, Joseph E Powell, and Urko M Marigorta. Expression quantitative trait locus analysis for translational medicine. *Genome medicine*, 7:1–14, 2015.

Manfred S Green and Michael J Symons. A comparison of the logistic risk function and the proportional hazards model in prospective epidemiologic studies. *Journal of chronic diseases*, 36(10):715–723, 1983.

Andrew D Grotzinger, Mijke Rhemtulla, Ronald de Vlaming, Stuart J Ritchie, Travis T Mallard, W David Hill, Hill F Ip, Riccardo E Marioni, Andrew M McIntosh, Ian J Deary, et al. Genomic structural equation modelling provides insights into the multivariate genetic architecture of complex traits. *Nature human behaviour*, 3(5):513–525, 2019.

GTEx Consortium. The GTEx consortium atlas of genetic regulatory effects across human tissues. *Science*, 369(6509):1318–1330, 2020.

GTEx Consortium, Kristin G Ardlie, David S Deluca, Ayellet V Segrè, Timothy J Sullivan, Taylor R Young, Ellen T Gelfand, Casandra A Trowbridge, Julian B Maller, Taru Tukiainen, et al. The genotype-tissue expression (gtex) pilot analysis: multitissue gene regulation in humans. *Science*, 348(6235):648–660, 2015.

Yongtao Guan and Matthew Stephens. Bayesian variable selection regression for genome-wide association studies and other large-scale problems. 2011.

Paul Gustafson, Peter Carbonetto, Natalie Thompson, and Nando de Freitas. Bayesian feature weighting for unsupervised learning, with application to object recognition. In Christopher M. Bishop and Brendan J. Frey, editors, *Proceedings of the 9th International Workshop on Artificial Intelligence and Statistics*, pages 124–131, 2003.

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

Sara A Hart, Callie Little, and Elsje van Bergen. Nurture might be nature: Cautionary tales and proposed solutions. *NPJ science of learning*, 6(1):2, 2021.

Fernando Pires Hartwig, Neil Martin Davies, and George Davey Smith. Bias in mendelian randomization due to assortative mating. *Genetic epidemiology*, 42(7):608–620, 2018.

Tomomitsu Hirota, Atsushi Takahashi, Michiaki Kubo, Tatsuhiko Tsunoda, Kaori Tomita, Satoru Doi, Kimie Fujita, Akihiko Miyatake, Tadao Enomoto, Takehiko Miyagawa, et al. Genome-wide association study identifies three new susceptibility loci for adult asthma in the japanese population. *Nature genetics*, 43(9):893–896, 2011.

T. Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd Annual International ACM SIGIR Conference*, pages 50–57, 1999.

Thomas Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42(1):177–196, 2001.

Farhad Hormozdiari, Emrah Kostem, Eun Yong Kang, Bogdan Pasaniuc, and Eleazar Eskin. Identifying causal variants at loci with multiple signals of association. In *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 610–611, 2014.

Laurence J Howe, Thomas Battram, Tim T Morris, Fernando P Hartwig, Gibran Hemani, Neil M Davies, and George Davey Smith. Assortative mating and within-spouse pair comparisons. *PLoS genetics*, 17(11):e1009883, 2021.

Laurence J Howe, Michel G Nivard, Tim T Morris, Ailin F Hansen, Humaira Rasheed, Yoonsu Cho, Geetha Chittoor, Rafael Ahlskog, Penelope A Lind, Teemu Palviainen, et al. Within-sibship genome-wide association analyses decrease bias in estimates of direct genetic effects. *Nature genetics*, 54(5):581–592, 2022.

Laurence J Howe, Humaira Rasheed, Paul R Jones, Dorret I Boomsma, David M Evans, Alexandros Giannelis, Caroline Hayward, John L Hopper, Amanda Hughes, Hannu Lahtinen, et al. Educational attainment, health outcomes and mortality: a within-sibship mendelian randomization study. *International Journal of Epidemiology*, 52(5):1579–1591, 2023.

Yiming Hu, Qiongshi Lu, Wei Liu, Yuhua Zhang, Mo Li, and Hongyu Zhao. Joint modeling of genetically correlated diseases and functional annotations increases accuracy of polygenic risk prediction. *PLoS Genetics*, 13(6):e1006836, 2017.

Joseph G Ibrahim, Ming-Hui Chen, and Steven N MacEachern. Bayesian variable selection for proportional hazards models. *Canadian Journal of Statistics*, 27(4):701–717, 1999.

Joseph G Ibrahim, Ming-Hui Chen, Debajyoti Sinha, JG Ibrahim, and MH Chen. *Bayesian survival analysis*, volume 2. Springer, 2001.

Nallathambi Jeyabalan and James P Clement. Syngap1: mind the gap. *Frontiers in Cellular Neuroscience*, 10:32, 2016.

Iain Johnstone. Gaussian estimation: sequence and wavelet models, 2019. Available at `https://imjohnstone.su.domains/GE_08_09_17.pdf`.

Iain M. Johnstone and Debashis Paul. PCA in high dimensions: an orientation. *Proceedings of the IEEE*, 106(8):1277–1292, 2018.

John D Kalbfleisch. Non-parametric bayesian analysis of survival time data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 40(2):214–221, 1978.

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

Philipp D Koellinger and Ronald De Vlaming. Mendelian randomization: the challenge of unobserved environmental confounds, 2019.

Michael Komodromos, Eric O Aboagye, Marina Evangelou, Sarah Filippi, and Kolyan Ray. Variational bayes for high-dimensional proportional hazards models with applications within gene expression. *Bioinformatics*, 38(16):3918–3926, 2022.

Augustine Kong, Gudmar Thorleifsson, Michael L Frigge, Bjarni J Vilhjalmsson, Alexander I Young, Thorgeir E Thorgeirsson, Stefania Benonisdottir, Asmundur Oddsson, Bjarni V Halldorsson, Gisli Masson, et al. The nature of nurture: Effects of parental genotypes. *Science*, 359(6374):424–428, 2018.

Anshul Kundaje, Wouter Meuleman, Jason Ernst, Misha Bilenky, Angela Yen, Alireza Heravi-Moussavi, Pouya Kheradpour, Zhizhuo Zhang, Jianrong Wang, Michael J Ziller, et al. Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–330, 2015.

Mitja I Kurki, Juha Karjalainen, Priit Palta, Timo P Sipilä, Kati Kristiansson, Kati M Donner, Mary P Reeve, Hannele Laivuori, Mervi Aavikko, Mari A Kaunisto, et al. Finngen provides genetic insights from a well-phenotyped isolated population. *Nature*, 613(7944): 508–518, 2023.

Melissa J Landrum, Jennifer M Lee, Mark Benson, Garth Brown, Chen Chao, Shanmuga Chitipiralla, Baoshan Gu, Jennifer Hart, Douglas Hoffman, Jeffrey Hoover, et al. Clinvar: public archive of interpretations of clinically relevant variants. *Nucleic acids research*, 44 (D1):D862–D868, 2016.

Gary L Larsen. Differences between adult and childhood asthma. *Journal of allergy and clinical immunology*, 106(3):S153–S157, 2000.

Debbie A Lawlor, Roger M Harbord, Jonathan AC Sterne, Nic Timpson, and George Davey Smith. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Statistics in medicine*, 27(8):1133–1163, 2008.

Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004.

Olivier Ledoit and Michael Wolf. The power of (non-)linear shrinking: a review and guide to covariance matrix estimation. *Journal of Financial Econometrics*, 20(1):187–218, 2022.

James J Lee, Robbee Wedow, Aysu Okbay, Edward Kong, Omeed Maghzian, Meghan Zacher, Tuan Anh Nguyen-Viet, Peter Bowers, Julia Sidorenko, Richard Karlsson Linnér, et al. Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nature genetics*, 50(8):1112–1121, 2018.

Phil H Lee, Verneri Anttila, Hyejung Won, Yen-Chen A Feng, Jacob Rosenthal, Zhaozhong Zhu, Elliot M Tucker-Drob, Michel G Nivard, Andrew D Grotzinger, Danielle Posthuma, et al. Genomic relationships, novel loci, and pleiotropic mechanisms across eight psychiatric disorders. *Cell*, 179(7):1469–1482, 2019.

Lihua Lei and William Fithian. Adapt: an interactive procedure for multiple testing with side information. *Journal of the Royal Statistical Society, Series B*, 80(4):649–679, 2018.

Lin Li, Michael Kabesch, Emmanuelle Bouzigon, Florence Demenais, Martin Farrall, Miriam F Moffatt, Xihong Lin, and Liming Liang. Using eqtl weights to improve power for genome-wide association studies: a genetic study of childhood asthma. *Frontiers in genetics*, 4:103, 2013.

Qin Li, Michael J Gloudemans, Jonathan M Geisinger, Boming Fan, François Aguet, Tao Sun, Gokul Ramaswami, Yang I Li, Jin-Biao Ma, Jonathan K Pritchard, et al. Rna editing underlies genetic risk of common inflammatory diseases. *Nature*, 608(7923):569–577, 2022.

Wenhe Lin, Jeffrey D. Wall, Ge Li, Deborah Newman, Yunqi Yang, Mark Abney, John L. VandeBerg, Michael Olivier, Yoav Gilad, and Laura A. Cox. Genetic regulatory effects in response to a high-cholesterol, high-fat diet in baboons. *Cell Genomics*, 4(3):100509, 2024.

Chuanhai Liu and Donald B Rubin. Maximum likelihood estimation of factor analysis using the ECME algorithm with complete and incomplete data. *Statistica Sinica*, 8(3):729–747, 1998.

Qing Liu and Donald A Pierce. A note on gauss—hermite quadrature. *Biometrika*, 81(3): 624–629, 1994.

Yusha Liu, Peter Carbonetto, Michihiro Takahama, Adam Gruenbaum, Dongyue Xie, Nicolas Chevrier, and Matthew Stephens. A flexible model for correlated count data, with application to analysis of gene expression differences in multi-condition experiments. *arXiv*, 2210.00697, 2023.

Nerea Llamosas, Vineet Arora, Ridhima Vij, Murat Kilinc, Lukasz Bijoch, Camilo Rojas, et al. Syngap1 controls the maturation of dendrites, synaptic function, and network activity in developing human neurons. *Journal of Neuroscience*, 40(41):7980–7994, 2020.

Luke R Lloyd-Jones, Jian Zeng, Julia Sidorenko, Loïc Yengo, Gerhard Moser, Kathryn E Kemper, Huanwei Wang, Zhili Zheng, Reedik Magi, Tonu Esko, et al. Improved polygenic prediction by bayesian multiple regression on summary statistics. *Nature communications*, 10(1):1–11, 2019.

Adam E Locke, Bratati Kahali, Sonja I Berndt, Anne E Justice, Tune H Pers, Felix R Day, Corey Powell, Sailaja Vedantam, Martin L Buchkovich, Jian Yang, et al. Genetic studies of body mass index yield new insights for obesity biology. *Nature*, 518(7538):197–206, 2015.

Po-Ru Loh, George Tucker, Brendan K Bulik-Sullivan, Bjarni J Vilhjalmsson, Hilary K Finucane, Rany M Salem, Daniel I Chasman, Paul M Ridker, Benjamin M Neale, Bonnie Berger, et al. Efficient bayesian mixed-model analysis increases association power in large cohorts. *Nature genetics*, 47(3):284, 2015.

Lan Luo, Judong Shen, Hong Zhang, Aparna Chhibber, Devan V Mehrotra, and Zheng-Zheng Tang. Multi-trait analysis of rare-variant association summary statistics using mtar. *Nature Communications*, 11(1):2850, 2020.

Robert Maier, Gerhard Moser, Guo-Bo Chen, Stephan Ripke, Devin Absher, Ingrid Agartz, Huda Akil, Farooq Amin, Ole A Andreassen, Adebayo Anjorin, et al. Joint analysis of psychiatric disorders increases accuracy of risk prediction for schizophrenia, bipolar disorder, and major depressive disorder. *The American Journal of Human Genetics*, 96 (2):283–294, 2015.

Timothy Shin Heng Mak, Robert Milan Porsch, Shing Wan Choi, Xueya Zhou, and Pak Chung Sham. Polygenic scores via penalized regression on summary statistics. *Genetic epidemiology*, 41(6):469–480, 2017.

Ani Manichaikul, Josyf C Mychaleckyj, Stephen S Rich, Kathy Daly, Michèle Sale, and Wei-Min Chen. Robust relationship inference in genome-wide association studies. *Bioinformatics*, 26(22):2867–2873, 2010.

Geoffrey McLachlan and David Peel. *Finite mixture models*. John Wiley & Sons, Inc., New York, NY, 2000.

Andrew McMartin and Dalton Conley. Commentary: Mendelian randomization and education–challenges remain. *International Journal of Epidemiology*, 49(4):1193–1206, 2020.

Leon Mirsky. A trace inequality of John von Neumann. *Monatshefte für mathematik*, 79(4): 303–306, 1975.

Andréanne Morin, Emma E Thompson, Britney A Helling, Lyndsey E Shorey-Kendrick, Pieter Faber, Tebeb Gebretsadik, Leonard B Bacharier, Meyer Kattan, George T O'Connor, Katherine Rivera-Spoljaric, et al. A functional genomics pipeline to identify high-value asthma and allergy cpgs in the human methylome. *Journal of Allergy and Clinical Immunology*, 151(6):1609–1621, 2023.

Carl N. Morris. Parametric empirical Bayes inference: theory and applications. *Journal of the American Statistical Association*, 78(381):47–55, 1983.

Jean Morrison, Nicholas Knoblauch, Joseph H Marcus, Matthew Stephens, and Xin He. Mendelian randomization accounting for correlated and uncorrelated pleiotropic effects using genome-wide summary statistics. *Nature genetics*, 52(7):740–747, 2020.

Gerhard Moser, Sang Hong Lee, Ben J Hayes, Michael E Goddard, Naomi R Wray, and Peter M Visscher. Simultaneous discovery, estimation and prediction analysis of complex traits using a bayesian mixture model. *PLoS Genet*, 11(4):e1004969, 2015.

Rabab Nasrallah, Charlotte J Imianowski, Lara Bossini-Castillo, Francis M Grant, Mikail Dogan, Lindsey Placek, Lina Kozhaya, Paula Kuo, Firas Sadiyah, Sarah K Whiteside, et al. A distal enhancer at risk locus 11q13. 5 promotes suppression of colitis by treg cells. *Nature*, 583(7816):447–452, 2020.

Heini M Natri, Christina B Del Azodi, Lance Peter, Chase J Taylor, Sagrika Chugh, Robert Kendle, et al. Cell-type-specific and disease-associated expression quantitative trait loci in the human lung. *Nature Genetics*, 56:595–604, 2024.

John C Naylor and Adrian FM Smith. Applications of a method for the efficient computation of posterior distributions. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 31(3):214–225, 1982.

R Neal and G Hinton. A view of the EM algorithm that justifies incremental, sparse, and other variants. In M. I. Jordan, editor, *Learning in Graphical Models*, pages 355–368. Springer, New York, 1998.

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

Paul J Newcombe, H Raza Ali, Fiona M Blows, Elena Provenzano, Paul D Pharoah, Carlos Caldas, and Sylvia Richardson. Weibull regression with bayesian variable selection to identify prognostic tumour markers of breast cancer survival. *Statistical methods in medical research*, 26(1):414–436, 2017.

Amir Nikooienejad, Wenyi Wang, and Valen E Johnson. Bayesian variable selection for survival data using inverse moment priors. *The annals of applied statistics*, 14(2):809, 2020.

Michel Guillaume Nivard, Daniel Belsky, Kathryn Paige Harden, Tina Baier, Ole A Andreassen, Eivind Ystrom, Elsje van Bergen, and Torkild Hovde Lyngstad. Neither nature nor nurture: Using extended pedigree data to understand indirect genetic effects on offspring educational outcomes. 2022.

Sven E Ojavee, Liza Darrous, Marion Patxot, Kristi Läll, Krista Fischer, Reedik Mägi, Zoltan Kutalik, and Matthew R Robinson. Genetic insights into the age-specific biological mechanisms governing human ovarian aging. *The American Journal of Human Genetics*, 110(9):1549–1563, 2023.

Paul F O'Reilly, Clive J Hoggart, Yotsawat Pomyen, Federico CF Calboli, Paul Elliott, Marjo-Riitta Jarvelin, and Lachlan JM Coin. Multiphen: joint model of multiple phenotypes can increase discovery in gwas. *PloS one*, 7(5):e34861, 2012.

Joseph K Pickrell, Tomaz Berisa, Jimmy Z Liu, Laure Ségurel, Joyce Y Tung, and David A Hinds. Detection and interpretation of shared genetic influences on 42 human traits. *Nature Genetics*, 48(7):709–717, 2016.

Milton Pividori, Nathan Schoettler, Dan L Nicolae, Carole Ober, and Hae Kyung Im. Shared and distinct genetic risk factors for childhood-onset and adult-onset asthma: genome-wide and transcriptome-wide studies. *The Lancet Respiratory Medicine*, 7(6):509–522, 2019.

Heather F Porter and Paul F O'Reilly. Multivariate simulation framework reveals performance of multi-trait gwas methods. *Scientific reports*, 7(1):1–12, 2017.

RC Richmond and G Davey Smith. Mendelian randomization: Concepts and scope. cold spring harb perspect med. 2022; 12 (1).

Herbert Robbins. Asymptotically subminimax solutions of compound statistical decision problems. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, 1951, vol. II*, pages 131–149. University of California Press, Berkeley and Los Angeles, CA, 1951.

Donald B Rubin and Dorothy T Thayer. EM algorithms for ML factor analysis. *Psychometrika*, 47:69–76, 1982.

Douglas M Ruderfer, Stephan Ripke, Andrew McQuillin, James Boocock, Eli A Stahl, Jennifer M Whitehead Pavlides, Niamh Mullins, Alexander W Charney, Anil PS Ori, Loes M Olde Loohuis, et al. Genomic dissection of bipolar disorder and schizophrenia, including 28 subphenotypes. *Cell*, 173(7):1705–1715, 2018.

USING ROBUST ADJUSTED PROFILE SCORE. Statistical inference in two-sample summary-data mendelian randomization using robust adjusted profile score. *arXiv preprint arXiv:1801.09652*, 2018.

Saskia Selzam, Stuart J Ritchie, Jean-Baptiste Pingault, Chandra A Reynolds, Paul F O'Reilly, and Robert Plomin. Comparing within-and between-family polygenic score prediction. *The American Journal of Human Genetics*, 105(2):351–363, 2019.

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

Andrey A. Shabalin. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics*, 28(10):1353–1358, 2012.

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

Debajyoti Sinha, Joseph G Ibrahim, and Ming-Hui Chen. A bayesian justification of cox's partial likelihood. *Biometrika*, 90(3):629–641, 2003.

SP Smieszek, Sarah Welsh, C Xiao, J Wang, Christos Polymeropoulos, Gunther Birznieks, and MH Polymeropoulos. Correlation of age-of-onset of atopic dermatitis with filaggrin loss-of-function variant status. *Scientific Reports*, 10(1):2721, 2020.

Frances JD Smith, Alan D Irvine, Ana Terron-Kwiatkowski, Aileen Sandilands, Linda E Campbell, Yiwei Zhao, Haihui Liao, Alan T Evans, David R Goudie, Sue Lewis-Jones, et al. Loss-of-function mutations in the gene encoding filaggrin cause ichthyosis vulgaris. *Nature genetics*, 38(3):337–342, 2006.

Gordon Smyth, Yifang Hu, Peter Dunn, Belinda Phipson, Yunshun Chen, and Maintainer Gordon Smyth. Package 'statmod'. *R Documentation. Package for R programming*, 2017.

Marcus M Soliai, Atsushi Kato, Britney A Helling, Catherine T Stanhope, James E Norton, Katherine A Naughton, et al. Multi-omics colocalization with genome-wide association studies reveals a context-specific genetic mechanism at a childhood onset asthma risk locus. *Genome Medicine*, 13:157, 2021.

Elliot Sollis, Abayomi Mosaku, Ala Abid, Annalisa Buniello, Maria Cerezo, Laurent Gil, et al. The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Research*, 51(D1):D977–D985, 2022.

James R Staley, Edmund Jones, Stephen Kaptoge, Adam S Butterworth, Michael J Sweeting, Angela M Wood, and Joanna MM Howson. A comparison of cox and logistic regression for use in genome-wide association studies of cohort and case-cohort design. *European Journal of Human Genetics*, 25(7):854–862, 2017.

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

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

Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, et al. Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine*, 12(3):e1001779, 2015.

Benjamin B Sun, Mitja I Kurki, Christopher N Foley, Asma Mechakra, Chia-Yen Chen, Eric Marshall, Jemma B Wilk, Biogen Biobank Team Sun Benjamin B. 1 2 Ghen Chia-Yen 1 Marshall Eric 1 Wilk Jemma B. 1 Runz Heiko 1, Mohamed Chahine, Philippe Chevalier, et al. Genetic associations of protein-coding variants in human disease. *Nature*, 603(7899): 95–102, 2022.

Lei Sun. *Topics on Empirical Bayes Normal Means*. PhD thesis, University of Chicago, Chicago, IL, 2020.

Michael E Tipping and Christopher M Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society, Series B*, 61(3):611–622, 1999.

Justin D Tubbs, Robert M Porsch, Stacey S Cherny, and Pak C Sham. The genes we inherit and those we don't: maternal genetic nurture and child bmi trajectories. *Behavior genetics*, 50(5):310–319, 2020.

Michael C Turchin and Matthew Stephens. Bayesian multivariate reanalysis of large genetic studies identifies many new associations. *PLoS Genetics*, 15(10):e1008431, 2019.

Patrick Turley, Raymond K Walters, Omeed Maghzian, Aysu Okbay, James J Lee, Mark Alan Fontana, Tuan Anh Nguyen-Viet, Robbee Wedow, Meghan Zacher, Nicholas A Furlotte, et al. Multi-trait analysis of genome-wide association summary statistics using mtag. *Nature Genetics*, 50(2):229–237, 2018.

Miriam S Udler, Jaegil Kim, Marcin von Grotthuss, Sílvia Bonàs-Guarch, Joanne B Cole, Joshua Chiou, Christopher D. Anderson on behalf of METASTROKE and the ISGC, Michael Boehnke, et al. Type 2 diabetes genetic loci informed by multi-trait associations point to disease mechanisms and subtypes: a soft clustering analysis. *PLoS Medicine*, 15 (9):e1002654, 2018.

Emil Uffelmann, Qin Qin Huang, Nchangwi Syntia Munung, Jantina De Vries, Yukinori Okada, Alicia R Martin, Hilary C Martin, Tuuli Lappalainen, and Danielle Posthuma. Genome-wide association studies. *Nature Reviews Methods Primers*, 1(1):59, 2021.

Sarah M Urbut, Gao Wang, Peter Carbonetto, and Matthew Stephens. Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. *Nature Genetics*, 51(1):187–195, 2019.

Sarah M Urbut, Margaret Sunitha Selvaraj, Akhil Pampana, Lillian Dattilo, Benjamin Neale, Christopher J ODonnell, Gina Peloso, and Pradeep Natarajan. Bayesian multivariate genetic analysis of lipids improves translational insight. *Circulation*, 144(Suppl_1):A9855–A9855, 2021.

Carl Veller and Graham M Coop. Interpreting population-and family-based genome-wide association studies in the presence of confounding. *Plos Biology*, 22(4):e3002511, 2024.

Donata Vercelli. Discovering susceptibility genes for asthma and allergy. *Nature reviews immunology*, 8(3):169–182, 2008.

Peter M Visscher, Naomi R Wray, Qian Zhang, Pamela Sklar, Mark I McCarthy, Matthew A Brown, and Jian Yang. 10 years of gwas discovery: biology, function, and translation. *The American Journal of Human Genetics*, 101(1):5–22, 2017.

Theo Vos, Amanuel Alemu Abajobir, Kalkidan Hassen Abate, Cristiana Abbafati, Kaja M Abbas, Foad Abd-Allah, RS Abdulkader, AM Abdulle, TA Abebo, SF Abera, et al. Gbd 2016 disease and injury incidence and prevalence collaborators. global, regional, and national incidence, prevalence, and years lived with disability for 328 diseases and injuries for 195 countries, 1990-2016: a systematic analysis for the global burden of disease study 2016. *Lancet*, 390(10100):1211–59, 2017.

Johannes Waage, Marie Standl, John A Curtin, Leon E Jessen, Jonathan Thorsen, Chao Tian, Nathan Schoettler, 23andMe Research Team, AAGC collaborators, Carlos Flores, et al. Genome-wide association and hla fine-mapping studies identify risk loci and genetic pathways underlying allergic rhinitis. *Nature genetics*, 50(8):1072–1080, 2018.

Jon Wakefield. 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, 2009.

Chris Wallace, Antony J Cutler, Nikolas Pontikos, Marcin L Pekalski, Oliver S Burren, Jason D Cooper, Arcadio Rubio Garcia, Ricardo C Ferreira, Hui Guo, Neil M Walker, et al. Dissection of a complex disease susceptibility region using a bayesian stochastic search approach to fine mapping. *PLoS genetics*, 11(6):e1005272, 2015.

Bo Wang and D. M. Titterington. Inadequacy of interval estimates corresponding to variational bayesian approximations. In *Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics*, pages 373–380, 2005.

Gao Wang, Abhishek Sarkar, Peter Carbonetto, and Matthew Stephens. 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, 2020.

Jingshu Wang, Qingyuan Zhao, Jack Bowden, Gibran Hemani, George Davey Smith, Dylan S Small, and Nancy R Zhang. Causal inference for heritable phenotypic risk factors using heterogeneous genetic instruments. *PLoS genetics*, 17(6):e1009575, 2021.

Kai Wang, Mingyao Li, and Hakon Hakonarson. Annovar: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research*, 38(16):e164–e164, 2010.

Wei Wang and Matthew Stephens. Empirical bayes matrix factorization. *arXiv preprint arXiv:1802.06931*, 2018.

Wei Wang and Matthew Stephens. Empirical bayes matrix factorization. *Journal of Machine Learning Research*, 22(120):1–40, 2021.

Nicole M Warrington, Rachel M Freathy, Michael C Neale, and David M Evans. Using structural equation modelling to jointly estimate maternal and fetal effects on birthweight in the uk biobank. *International journal of epidemiology*, 47(4):1229–1241, 2018.

Kyoko Watanabe, Erdogan Taskesen, Arjen Van Bochoven, and Danielle Posthuma. Functional mapping and annotation of genetic associations with fuma. *Nature communications*, 8(1):1–11, 2017.

Kyoko Watanabe, Sven Stringer, Oleksandr Frei, Maša Umićević Mirkov, Christiaan de Leeuw, Tinca JC Polderman, Sophie van der Sluis, Ole A Andreassen, Benjamin M Neale, and Danielle Posthuma. A global overview of pleiotropy and genetic architecture in complex traits. *Nature genetics*, 51(9):1339–1348, 2019.

Xiaoquan Wen and Matthew Stephens. Bayesian methods for genetic association analysis with heterogeneous subgroups: from meta-analyses to gene-environment interactions. *The Annals of Applied Statistics*, 8(1):176–203, 2014.

Xiaoquan Wen, Yeji Lee, Francesca Luca, and Roger Pique-Regi. Efficient integrative multi-snp association analysis via deterministic approximation of posteriors. *The American Journal of Human Genetics*, 98(6):1114–1129, 2016.

Halbert White. Maximum likelihood estimation of misspecified models. *Econometrica: Journal of the econometric society*, pages 1–25, 1982.

Cristen J Willer, Yun Li, and Gonçalo R Abecasis. Metal: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics*, 26(17):2190–2191, 2010.

Jason Willwerscheid and Matthew Stephens. ebnm: An r package for solving the empirical bayes normal means problem using a variety of prior families. *arXiv preprint arXiv:2110.00152*, 2021.

Joong-Ho Won, Johan Lim, Seung-Jean Kim, and Bala Rajaratnam. Condition-number-regularized covariance estimation. *Journal of the Royal Statistical Society, Series B*, 75 (3):427–450, 2013.

Andrew R Wood, Tonu Esko, Jian Yang, Sailaja Vedantam, Tune H Pers, Stefan Gustafsson, Audrey Y Chu, Karol Estrada, Jian'an Luan, Zoltán Kutalik, et al. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nature genetics*, 46(11):1173–1186, 2014.

Yuchang Wu, Xiaoyuan Zhong, Yunong Lin, Zijie Zhao, Jiawen Chen, Boyan Zheng, James J Li, Jason M Fletcher, and Qiongshi Lu. Estimating genetic nurture with summary statistics of multigenerational genome-wide association studies. *Proceedings of the National Academy of Sciences*, 118(25), 2021.

Yulu Wu, Hongbao Cao, Ancha Baranova, Hailiang Huang, Sheng Li, Lei Cai, et al. Multi-trait analysis for genome-wide association study of five psychiatric disorders. *Translational Psychiatry*, 10:209, 2020.

Dongyue Xie and Matthew Stephens. Discussion of "confidence intervals for nonparametric empirical bayes analysis". *Journal of the American Statistical Association*, 117(539):1186–1191, 2022.

Jian Yang, S Hong Lee, Michael E Goddard, and Peter M Visscher. Gcta: a tool for genome-wide complex trait analysis. *The American Journal of Human Genetics*, 88(1): 76–82, 2011.

Shu Yang and Peng Ding. Combining multiple observational data sources to estimate causal effects. *Journal of the American Statistical Association*, 2019.

Yunqi Yang, Peter Carbonetto, and Matthew Stephens. Supplementary materials for "improved methods for empirical bayes multivariate multiple testing and effect size estimation", 2024a.

Yunqi Yang, Peter Carbonetto, and Matthew Stephens. R package and source code reproducing the analyses for "improved methods for empirical bayes multivariate multiple testing and effect size estimation", 2024b.

Alexander I Young, Stefania Benonisdottir, Molly Przeworski, and Augustine Kong. Deconstructing the sources of genotype-phenotype associations in humans. *Science*, 365(6460): 1396–1400, 2019.

Alexander I Young, Seyed Moeen Nehzati, Stefania Benonisdottir, Aysu Okbay, Hariharan Jayashankar, Chanwook Lee, David Cesarini, Daniel J Benjamin, Patrick Turley, and Augustine Kong. Mendelian imputation of parental genotypes improves estimates of direct genetic effects. *Nature genetics*, 54(6):897–905, 2022.

Haoyu Zhang, Jianan Zhan, Jin Jin, Jingning Zhang, Wenxuan Lu, Ruzhang Zhao, et al. A new method for multiancestry polygenic prediction improves performance across diverse populations. *Nature Genetics*, 55(10):1757–1768, 2023.

J-H Zhao, LH Philip, and Qibao Jiang. ML estimation for factor analysis: EM or non-EM? *Statistics and Computing*, 18(2):109–123, 2008.

Qingyuan Zhao, Jingshu Wang, Gibran Hemani, Jack Bowden, and Dylan S. Small. Statistical inference in two-sample summary-data Mendelian randomization using robust adjusted profile score. *The Annals of Statistics*, 48(3):1742 – 1769, 2020. doi:10.1214/19-AOS1866. URL `https://doi.org/10.1214/19-AOS1866`.

Xiang Zhou and Matthew Stephens. Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nature Methods*, 11(4):407–409, 2014.

Zhaozhong Zhu, Xi Zhu, Cong-Lin Liu, Huwenbo Shi, Sipeng Shen, Yunqi Yang, Kohei Hasegawa, Carlos A Camargo, and Liming Liang. Shared genetics of asthma and mental health disorders: a large-scale genome-wide cross-trait analysis. *European Respiratory Journal*, 54(6), 2019.

Zhaozhong Zhu, Kohei Hasegawa, Carlos A Camargo Jr, and Liming Liang. Investigating asthma heterogeneity through shared and distinct genetics: insights from genome-wide cross-trait analysis. *Journal of Allergy and Clinical Immunology*, 147(3):796–807, 2021.

Yuxin Zou, Peter Carbonetto, Gao Wang, and Matthew Stephens. Fine-mapping from summary data with the "sum of single effects" model. *PLoS genetics*, 18(7):e1010299, 2022.

Yuxin Zou, Peter Carbonetto, Dongyue Xie, Gao Wang, and Matthew Stephens. Fast and flexible joint fine-mapping of multiple traits via the sum of single effects model. *bioRxiv*, 2023.

Yuxin Zou, Peter Carbonetto, Dongyue Xie, Gao Wang, and Matthew Stephens. Fast and flexible joint fine-mapping of multiple traits via the Sum of Single Effects model. *bioRxiv*, page doi:10.1101/2023.04.14.536893, 2024.

CHAPTER 5

APPENDICES

A Derivations, proofs, and additional definitions for Chapter 2

A.1 Posterior distribution for unknown means

Urbut et al. [2019] gives the posterior distributions only for the case when the prior covariances $\boldsymbol{U}_k$ are invertible, so here we give the slightly more general expressions that allow for one or more of the $\boldsymbol{U}_k$ to be singular.

For $\boldsymbol{x} \mid \boldsymbol{\theta} \sim N_R(\boldsymbol{\theta}, \boldsymbol{V})$ and $\boldsymbol{\theta} \sim N_R(\boldsymbol{0}, \boldsymbol{U})$, we have

$$\boldsymbol{\theta} \mid \boldsymbol{x} \sim N_R(\boldsymbol{\mu}^*, \boldsymbol{U}^*) \tag{1}$$

in which

$$\boldsymbol{U}^* := \boldsymbol{U}^*(\boldsymbol{U}, \boldsymbol{V}) = \boldsymbol{U}(\boldsymbol{I}_R + \boldsymbol{V}^{-1}\boldsymbol{U})^{-1} \tag{2}$$

$$\boldsymbol{\mu}^* := \boldsymbol{\mu}^*(\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{x}) = \boldsymbol{U}^*\boldsymbol{V}^{-1}\boldsymbol{x}. \tag{3}$$

For the mixture prior, the posterior distribution is

$$p(\boldsymbol{\theta}_j \mid \boldsymbol{x}_j, \boldsymbol{\pi}, \mathcal{U}) = \sum_{k=1}^{K} \pi_{jk}^* N_R(\boldsymbol{\theta}_j; \boldsymbol{\mu}_{jk}^*, \boldsymbol{U}_{jk}^*) \tag{4}$$

where $\boldsymbol{U}_{jk}^* := \boldsymbol{U}^*(\boldsymbol{U}_k, \boldsymbol{V}_j)$ and $\boldsymbol{\mu}_{jk}^* := \boldsymbol{\mu}^*(\boldsymbol{U}_k, \boldsymbol{V}_j, \boldsymbol{x}_j)$, and

$$\pi_{jk}^* := \frac{\pi_k N_R(\boldsymbol{x}_j; 0, \boldsymbol{U}_k + \boldsymbol{V}_j)}{\sum_{k'=1}^{K} \pi_{k'} N_R(\boldsymbol{x}_j; 0, \boldsymbol{U}_{k'} + \boldsymbol{V}_j)}. \tag{5}$$

These expressions are the same as the expressions in Urbut et al. [2019] when all the $\boldsymbol{U}_k$ are invertible.

A.2 EM for weighted log-likelihoods

In this section we derive an "EM-like" algorithm for maximizing a weighted log-likelihood of the form:

$$\begin{aligned}\phi(\Theta; \boldsymbol{w}) &:= \sum_{j=1}^{n} w_j \log p_j(\boldsymbol{x}_j \mid \Theta) \\ &= \sum_{j=1}^{n} w_j l_j(\Theta), \end{aligned} \tag{6}$$

where the w_j are known (fixed) weights, the $\boldsymbol{x}_j$ denote the independently observed data, and Θ denotes the unknowns to be estimated. Note that here we do not make specific modeling assumptions; in particular, the results are not specific to the EBMNM model.

We assume, as in usual applications of EM, that the likelihoods can be written using an "augmented data" form; that is, $p_j(\boldsymbol{x}_j \mid \Theta) = \int p_j(\boldsymbol{x}_j, \boldsymbol{z}_j \mid \Theta)\, d\boldsymbol{z}_j$.

The following proposition gives an "EM-like" update that is guaranteed to increase (or not decrease) the weighted log-likelihood, ϕ.

Proposition 1. Given the current value of Θ, denoted $\Theta^{(0)}$, define a new value, $\Theta^{(1)}$, by applying the following steps:

1. E-step: For each $j = 1, \ldots, n$, compute the conditional distribution of $\boldsymbol{z}_j$, $p_j(\boldsymbol{z}_j \mid \boldsymbol{x}_j, \Theta^{(0)})$.

2. M-step: Set $\Theta^{(1)} = \text{argmax}_\Theta \sum_{j=1}^{n} w_j \mathrm{E}_{p_j(\boldsymbol{z}_j \mid \boldsymbol{x}_j, \Theta^{(0)})}\big[\log p(\boldsymbol{x}_j, \boldsymbol{z}_j \mid \Theta)\big]$.

Then $\phi(\Theta^{(1)}; \boldsymbol{w}) \geq \phi(\Theta^{(0)}; \boldsymbol{w})$.

Note that when the weights w_j are all 1, these steps are the standard E-step and M-step in EM.

Proof. Following Neal and Hinton [1998], we define

$$F(q_1, \ldots, q_n; \Theta) := \sum_{j=1}^{n} w_j \big\{ \mathrm{E}_{q_j}[\log p_j(\boldsymbol{x}_j, \boldsymbol{z}_j \mid \Theta)] + H(q_j) \big\}, \tag{7}$$

where q_j is any distribution of $\boldsymbol{z}_j$ and $H(q_j) = -\mathrm{E}_{q_j}[\log q_j(\boldsymbol{z}_j)]$ is the entropy of distribution q_j. Using Lemma 1, 2 of Neal and Hinton [1998], we have

$$\hat{q}_j(\Theta) := \mathrm{argmax}_{q_j} F(q_1, \ldots, q_n; \Theta) = p_j(\boldsymbol{z}_j \mid \boldsymbol{x}_j, \Theta), \quad j = 1, \ldots, n. \tag{8}$$

$$F(\hat{q}(\Theta); \Theta) = \sum_{j=1}^{n} w_j l_j(\Theta) = \phi(\Theta; \boldsymbol{w}), \tag{9}$$

where we have introduced the notation $\hat{q}(\Theta)$ as shorthand for $\hat{q}_1(\Theta), \ldots, \hat{q}_n(\Theta)$.

The function being maximized in the M-step differs from $F(\hat{q}(\Theta^{(0)}); \Theta)$ only by terms that do not depend on Θ, so the M-step can be written as

$$\Theta^{(1)} = \mathrm{argmax}_{\Theta} F(\hat{q}(\Theta^{(0)}); \Theta). \tag{10}$$

Therefore,

$$\phi(\Theta^{(1)}; \boldsymbol{w}) = F(\hat{q}(\Theta^{(1)}), \Theta^{(1)}) \geq F(\hat{q}(\Theta^{(0)}), \Theta^{(1)}) \geq F(\hat{q}(\Theta^{(0)}), \Theta^{(0)}) = \phi(\Theta^{(0)}; \boldsymbol{w}),$$

where the equalities are due to (9); the first inequality is due to the optimality of $\hat{q}(\Theta^{(1)})$ from its definition, and the second inequality is due to the optimality of $\Theta^{(1)}$ in (10). This completes the proof.

□

A.3 Derivation of the EM algorithm for fitting the EBMNM model

Here we derive the general EM algorithm for EBMNM (Algorithm 1).

First, we introduce a latent variable $\boldsymbol{z}_j$ for each $j = 1, \ldots, n$. Each $\boldsymbol{z}_j$ is a binary vector of length K indicating the component k from which $\boldsymbol{x}_j$ arose. Following Neal and Hinton [1998], we introduce the function $F(q, \mathcal{U}, \boldsymbol{\pi}, \boldsymbol{s})$,

$$\begin{aligned} F(q, \mathcal{U}, \boldsymbol{\pi}, \boldsymbol{s}) &= E_q[\log p(\boldsymbol{x}, \boldsymbol{z}; \mathcal{U}, \boldsymbol{\pi}) - \textstyle\sum_{k=1}^{K} \rho(\boldsymbol{U}_k / s_k)] - E_q[\log q(\boldsymbol{z})] \\ &= \sum_{j=1}^{n} \sum_{k=1}^{K} E_q\big[\log(\pi_k) + \log N_R(\boldsymbol{x}_j; 0, \boldsymbol{U}_k + \boldsymbol{V}_j)\big] - \sum_{k=1}^{K} \rho(\boldsymbol{U}_k / s_k) - E_q[\log q(\boldsymbol{z})], \end{aligned} \tag{11}$$

where q is any distribution over $\boldsymbol{z}$. Note that the only difference between (11) and the function F in Neal and Hinton [1998] is a constant term with respect to q, the penalty $\sum_{k=1}^{K} \rho(\boldsymbol{U}_k / s_k)$. Therefore, we can use Lemma 1 and Lemma 2 in Neal and Hinton [1998], which shows that the log-likelihood is related to F by

$$F(\hat{q}, \mathcal{U}, \boldsymbol{\pi}, \boldsymbol{s}) = l(\mathcal{U}, \boldsymbol{\pi}, \boldsymbol{s}), \tag{12}$$

where $\hat{q} := \operatorname{argmax}_q F(q, \mathcal{U}, \boldsymbol{\pi}, \boldsymbol{s})$, and is the conditional distribution of $\boldsymbol{z}$,

$$\hat{q}(\boldsymbol{z}) := \operatorname{argmax}_q F(q, \mathcal{U}, \boldsymbol{\pi}) = p(\boldsymbol{z} \mid \boldsymbol{x}, \mathcal{U}, \boldsymbol{\pi}). \tag{13}$$

In our case, F is related to the penalized log-likelihood,

$$F(\hat{q}, \boldsymbol{s}, \mathcal{U}, \boldsymbol{\pi}) = l(\mathcal{U}, \boldsymbol{\pi}) - \sum_{k=1}^{K} \rho(\boldsymbol{U}_k / s_k). \tag{14}$$

Therefore, the maximum-likelihood estimates of $\boldsymbol{s}, \boldsymbol{\pi}$ and the penalized maximum-likelihood estimates of $\boldsymbol{U}$ can be obtained by maximizing F jointly over $q, \boldsymbol{s}, \boldsymbol{U}, \boldsymbol{\pi}$:

$$\begin{aligned} (\hat{\mathcal{U}}, \hat{\boldsymbol{\pi}}, \hat{\boldsymbol{s}}) :=\ & \operatorname{argmax}_{\boldsymbol{s}, \mathcal{U}, \boldsymbol{\pi}} l(\mathcal{U}, \boldsymbol{\pi}) - \textstyle\sum_{k=1}^{K} \rho(\boldsymbol{U}_k / s_k) \\ =\ & \operatorname{argmax}_{\boldsymbol{s}, \mathcal{U}, \boldsymbol{\pi}} \max_q F(q, \mathcal{U}, \boldsymbol{\pi}, \boldsymbol{s}). \end{aligned}$$

The standard EM algorithm can be thought of as iterating between optimizing over q (the E-step) and optimizing over $\mathcal{U}, \boldsymbol{\pi}, \boldsymbol{s}$ (the M-step). The E-step computes $\hat{q}$. The expected value of z_{jk} under $\hat{q}$ is

$$w_{jk} := E_{\hat{q}}[z_{jk}] = \frac{\pi_k N_R(\boldsymbol{x}_j; 0, \boldsymbol{U}_k + \boldsymbol{V}_j)}{\sum_{k'=1}^{K} \pi_{k'} N_R(\boldsymbol{x}_j; 0, \boldsymbol{U}_{k'} + \boldsymbol{V}_j)}. \tag{15}$$

The part of the M-Step optimizing F over $\boldsymbol{\pi}$ is straightforward and given by

$$\hat{\pi}_k = \frac{1}{n} \sum_{j=1}^{n} w_{jk}, \quad k = 1, \ldots, K. \tag{16}$$

Optimizing F over $\boldsymbol{U}_k$ and s_k is described below.

A.4 Data transformation for homoskedastic case of EBMNM

When $\boldsymbol{V}_j = \boldsymbol{V}$, we can simplify the fitting procedure of EBMNM model by performing a data transformation and then applying algorithms to the case of $\boldsymbol{V} = \boldsymbol{I}_R$ on transformed data. Here we describe this approach in more details.

Let $\boldsymbol{R}$ be any matrix such that $\boldsymbol{V} = \boldsymbol{R}\boldsymbol{R}^T$ (e.g., $\boldsymbol{R}$ could be the Cholesky decomposition of $\boldsymbol{V}$). Since $\boldsymbol{V}$ is invertible, $\boldsymbol{R}$ is also invertible, so consider the transformed data $\tilde{\boldsymbol{x}}_j := \boldsymbol{R}^{-1}\boldsymbol{x}_j$. The marginal model (2.3) for the transformed data becomes

$$p(\tilde{\boldsymbol{x}}_j \mid \boldsymbol{\pi}, \mathcal{U}, \boldsymbol{V}_j) = \sum_{k=1}^{K} \pi_k N_R(\tilde{\boldsymbol{x}}_j; \boldsymbol{0}, \tilde{\boldsymbol{U}}_k + \boldsymbol{I}_R), \tag{17}$$

where $\tilde{\boldsymbol{U}}_k := \boldsymbol{R}^{-1}\boldsymbol{U}_k\boldsymbol{R}^{-T}$. Thus, we can estimate $\boldsymbol{U}_k$ by first estimating $\tilde{\boldsymbol{U}}_k$—by fitting the EBMNM model to transformed data $\tilde{\boldsymbol{x}}_j$ with $\boldsymbol{V} = \boldsymbol{I}_R$, yielding estimates $\hat{\tilde{\boldsymbol{U}}}_k$ say—and then

reversing the transformation to obtain estimates $\hat{\boldsymbol{U}}_k$ for $\boldsymbol{U}_k$,

$$\hat{\boldsymbol{U}}_k = \boldsymbol{R}\hat{\tilde{\boldsymbol{U}}}_k\boldsymbol{R}^T, \quad k = 1, \ldots, K. \tag{18}$$

When taking this approach, any constraint on $\boldsymbol{U}_k$ ($\boldsymbol{U}_k \in P_R^{+,k}$) must be translated to an equivalent corresponding constraint on $\tilde{\boldsymbol{U}}_k$ ($\tilde{\boldsymbol{U}}_k \in \tilde{P}_R^{+,k}$ say) when fitting the EBMNM model. For example, a rank-1 constraint on $\boldsymbol{U}_k$ translates to a rank-1 constraint on $\tilde{\boldsymbol{U}}$.

Note that with this transformation, when a penalty function is included in the EBMNM problem, the penalty is imposed on $\tilde{\boldsymbol{U}}_k$ rather than on $\boldsymbol{U}_k$. Specifically, it can be shown that this transformation approach, instead of solving (2.9), solves

$$(\hat{\boldsymbol{\pi}}, \hat{\mathcal{U}}, \hat{\boldsymbol{s}}) := \underset{\boldsymbol{\pi}\in\boldsymbol{S}_K, \boldsymbol{U}_k\in P_R^{+,k}}{\operatorname{argmax}} l(\boldsymbol{\pi}, \mathcal{U}) - \min_{\boldsymbol{s}} \sum_{k=1}^{K} \rho\left(\boldsymbol{R}^{-1}\boldsymbol{U}_k\boldsymbol{R}^{-T}/s_k\right). \tag{19}$$

When no penalty is included, clearly this does not alter the problem; *i.e.,* (2.9) and (19) are equivalent. With either the IW or NN penalty included, (2.9) and (19) differ; whereas (2.9) encourages $\boldsymbol{U}_k/s_k$ to be close to $\boldsymbol{I}_R$, (19) encourages $\tilde{\boldsymbol{U}}_k/s_k$ to be close to $\boldsymbol{I}_R$, and in turn encourages $\boldsymbol{U}_k/s_k$ to be close to $\boldsymbol{V}$. Whether one or other is better is difficult to say in general, and may be context-dependent. Indeed, the transformation approach ensures a stronger version of the invariance property (2.7):

$$\hat{\theta}(\boldsymbol{A}\boldsymbol{x}, \boldsymbol{A}\boldsymbol{V}\boldsymbol{A}^T) = \boldsymbol{A}\hat{\theta}(\boldsymbol{x}, \boldsymbol{V}) \tag{20}$$

for any invertible matrix $\boldsymbol{A}$. Further (19) has the advantage that it can be solved by our TED approach when $\boldsymbol{V}_j = \boldsymbol{V} \neq \boldsymbol{I}_R$, whereas (2.9) cannot. We therefore take the transformation approach as the default approach in our software.

Note that, for the IW and NN penalties, the penalty term in (19) does not depend on

the exact choice of matrix $\boldsymbol{R}$ in $\boldsymbol{V} = \boldsymbol{R}\boldsymbol{R}^T$. For example,

$$\begin{aligned}\rho_\lambda^{\mathrm{IW}}(\boldsymbol{R}^{-1}\boldsymbol{U}\boldsymbol{R}^{-T}) &= \frac{\lambda}{2}\left[\log\det \boldsymbol{R}^{-1}\boldsymbol{U}\boldsymbol{R}^{-T} + \mathrm{tr}((\boldsymbol{R}^{-1}\boldsymbol{U}\boldsymbol{R}^{-T})^{-1})\right] \\ &= \frac{\lambda}{2}\left[\log\det \boldsymbol{U} + \log\det \boldsymbol{R}^{-1}\boldsymbol{R}^{-T} + \mathrm{tr}(\boldsymbol{R}^T\boldsymbol{U}^{-1}\boldsymbol{R})\right] \\ &= \frac{\lambda}{2}\left[\log\det \boldsymbol{U}\boldsymbol{V}^{-1} + \mathrm{tr}(\boldsymbol{U}^{-1}\boldsymbol{V})\right].\end{aligned}$$

A.5 Algorithms for a special case of the EBMNM model when $K = 1$

In the M-step of Algorithm 1, we maximize the part of F that depends on $\boldsymbol{U}_k$ for each k, which is (2.27):

$$\max_{\boldsymbol{U} \in P_R^{+,k},\, s>0} \phi(\boldsymbol{U}; \boldsymbol{w}_k) - \rho(\boldsymbol{U}/s), \tag{21}$$

where

$$\phi(\boldsymbol{U}; \boldsymbol{w}) := \sum_{j=1}^{n} w_j \log N_R(\boldsymbol{x}_j; \boldsymbol{0}, \boldsymbol{U} + \boldsymbol{V}_j).$$

We take an alternating optimization approach to solving (21) in which we alternate between maximizing over $\boldsymbol{U}$ with fixed s, and maximizing over s with fixed $\boldsymbol{U}$. We have three algorithms to maximize over $\boldsymbol{U}$ in (21), which are detail in the following subsections. The update for s is given in Section A.5.

Truncated eigenvalue decomposition

We derive TED algorithm in the case where $\boldsymbol{V}_j = \boldsymbol{I}_R$. When $\boldsymbol{V} \neq \boldsymbol{I}_R$, we can simplify to the special case of $\boldsymbol{V}_j = \boldsymbol{I}_R$ by performing a simple data transformation described in Section A.4. When $\boldsymbol{V}_j = \boldsymbol{I}_R$ for all j, ϕ in (2.26) simplifies. Specifically, dropping terms that do not depend on $\boldsymbol{U}$, we have

$$\phi(\boldsymbol{U}; \boldsymbol{w}) = -\frac{\bar{w}}{2}\big\{\log|\boldsymbol{U} + \boldsymbol{I}| + \mathrm{tr}[(\boldsymbol{U} + \boldsymbol{I})^{-1}\hat{\boldsymbol{S}}]\big\}, \tag{22}$$

where $\boldsymbol{I} = \boldsymbol{I}_R$, and $\hat{\boldsymbol{S}} := \sum_{j=1}^n \tilde{w}_j \boldsymbol{x}_j \boldsymbol{x}_j^T$ is the (weighted) sample covariance with renormalized weights $\tilde{w}_j := w_j/\bar{w}$, $\bar{w} := \sum_{j=1}^n w_j$. Differentiating (22) with respect to $\boldsymbol{U} + \boldsymbol{I}$ and setting the derivative to zero yields the solution $\boldsymbol{U} = \hat{\boldsymbol{S}} - \boldsymbol{I}$. However, the matrix $\hat{\boldsymbol{S}} - \boldsymbol{I}$ is not necessarily a covariance; that is, it may have one or more eigenvalues that are negative. Intuitively, one might propose to create a valid covariance matrix by setting the negative eigenvalues to zero. Indeed, this intuition is correct: setting the negative eigenvalues of $\hat{\boldsymbol{S}} - \boldsymbol{I}$ to zero can be justified on the grounds that it maximizes (22) subject to the constraint that $\boldsymbol{U} \in P_R^+$. This is stated more formally by the following result.

Result 1. Let $\phi(\boldsymbol{U}; \boldsymbol{w})$ be defined in (22), and let $\hat{\boldsymbol{S}} = \boldsymbol{L}\boldsymbol{D}\boldsymbol{L}^T$ be the eigenvalue decomposition of $\hat{\boldsymbol{S}}$, with $\boldsymbol{D} := \mathrm{diag}(d_1, \ldots, d_R)$. Then we have that

$$\mathrm{argmax}_{\boldsymbol{U} \in P^+} \phi(\boldsymbol{U}; \boldsymbol{w}) = \boldsymbol{L}(\boldsymbol{D} - \boldsymbol{I})_+ \boldsymbol{L}^T, \tag{23}$$

where $\boldsymbol{A}_+$ denotes the matrix constructed from $\boldsymbol{A}$ by setting any negative elements of $\boldsymbol{A}$ to zero.

Remark 1. *It is straightforward to generalize this result to include an additional constraint on the rank of* $\boldsymbol{U}$*. Specifically, optimizing over* $\boldsymbol{U}$ *subject to the constraint that its rank is less than* R' *(where* $R' \leq R$*) can be achieved by first setting all negative eigenvalues to zero, then, if needed, setting the smallest positive eigenvalues to zero until at least* $R - R'$ *eigenvalues are zero.*

With the addition of a penalty, $\rho_\lambda^{\mathrm{IW}}(\boldsymbol{U})$ or $\rho_\lambda^{\mathrm{NN}}(\boldsymbol{U})$, the subproblem (2.27) no longer has a closed-form solution. However, it can nonetheless be solved easily using numerical methods, as explained in the following proposition.

Proposition 2. Let $\rho(\boldsymbol{U})$ be a function of $R \times R$ matrix $\boldsymbol{U}$ that is separable in the eigenvalues of its argument; that is, $\rho(\boldsymbol{U}) = \sum_{r=1}^R \rho_r(e_r)$ for some $\rho_r(\cdot)$, in which e_r denotes the rth

eigenvalue of $\boldsymbol{U}$. (This separability property is satisfied by both the IW and NN penalties.) Then

$$\text{argmax}_{\boldsymbol{U} \in P^+} \phi(\boldsymbol{U}; \boldsymbol{w}) - \rho(\boldsymbol{U}/s) = \boldsymbol{L}\text{diag}(\hat{e}_1, \ldots, \hat{e}_R)\boldsymbol{L}^T, \tag{24}$$

where

$$\hat{e}_r := \underset{e_r \geq 0}{\text{argmax}} -\frac{\bar{w}}{2}\left\{ \log(e_r + 1) + \frac{d_r}{e_r + 1} \right\} - \rho_r(e_r/s). \tag{25}$$

Proof. The proof relies on the following result, which is a corollary to the Von Neumann–Fan trace inequality [Mirsky, 1975], and is also used by Chi and Lange [2014]:

Result 2. For Hermitian $n \times n$ positive semidefinite complex matrices $\boldsymbol{A}, \boldsymbol{B}$ where the eigenvalues are sorted in decreasing order, $a_1 \geq a_2 \geq \cdots \geq a_n$ and $b_1 \geq b_2 \geq \cdots \geq b_n$, respectively, we have

$$\sum_{i=1}^{n} a_i b_{n-i+1} \leq \text{tr}(\boldsymbol{AB}) \leq \sum_{i=1}^{n} a_i b_i, \tag{26}$$

with equality if and only if $\boldsymbol{A}$ and $\boldsymbol{B}$ share singular vectors.

We use Result 2 to prove the following lemma:

Lemma 1. Define $f(\boldsymbol{Q}, \boldsymbol{E}; \boldsymbol{w}) := \phi(\boldsymbol{QEQ}^T; \boldsymbol{w})$ where ϕ denotes the function defined in (22), and $\boldsymbol{QEQ}^T$ is the eigenvalue decomposition of $\boldsymbol{U}$, so $\boldsymbol{Q}$ is an orthonormal matrix, and $\boldsymbol{E}$ is a diagonal matrix with non-negative entries $e_1 \geq e_2 \geq \ldots e_R \geq 0$ (so $\boldsymbol{QEQ}^T \in P_R^+$). Let $\boldsymbol{LDL}^T$ be the eigenvalue decomposition of the matrix $\boldsymbol{S}$ appearing in ϕ. Then

$$\boldsymbol{L} = \underset{\boldsymbol{Q}}{\text{argmax}}\, f(\boldsymbol{Q}, \boldsymbol{E}; \boldsymbol{w}), \tag{27}$$

and

$$\max_{\boldsymbol{Q}} f(\boldsymbol{Q}, \boldsymbol{E}; \boldsymbol{w}) = -\frac{\bar{w}}{2} \sum_{r=1}^{R} \left[\log(e_r + 1) + \frac{d_r}{e_r + 1} \right]. \tag{28}$$

Proof. From the definition,

$$f(\boldsymbol{Q}, \boldsymbol{E}; \boldsymbol{w}) = -\frac{\bar{w}}{2}\big(\log|\boldsymbol{Q}\boldsymbol{E}\boldsymbol{Q}^T + \boldsymbol{I}_R| + \mathrm{tr}((\boldsymbol{Q}\boldsymbol{E}\boldsymbol{Q}^T + \boldsymbol{I}_R)^{-1}\boldsymbol{S})\big) \tag{29}$$

$$= -\frac{\bar{w}}{2}\big(\log|\boldsymbol{E} + \boldsymbol{I}_R| + \mathrm{tr}(\boldsymbol{Q}(\boldsymbol{E} + \boldsymbol{I}_R)^{-1}\boldsymbol{Q}^T\boldsymbol{S})\big) \tag{30}$$

so

$$\max_{\boldsymbol{Q}} f(\boldsymbol{Q}, \boldsymbol{E}; \boldsymbol{w}) = -\frac{\bar{w}}{2}\left[\sum_r \log(e_r + 1) + \min_{\boldsymbol{Q}} \mathrm{tr}(\boldsymbol{Q}(\boldsymbol{E} + \boldsymbol{I}_R)^{-1}\boldsymbol{Q}^T\boldsymbol{S})\right]. \tag{31}$$

From (26) (left inequality), we have:

$$\mathrm{tr}(\boldsymbol{Q}(\boldsymbol{E} + \boldsymbol{I}_R)^{-1}\boldsymbol{Q}^T\boldsymbol{S}) \geq \sum_{i=1}^{R} \frac{d_i}{e_i + 1}, \tag{32}$$

with equality if and only if $\boldsymbol{Q} = \boldsymbol{L}$, and the result follows. □

Proposition 2 then follows by parameterizing $\boldsymbol{U} = \boldsymbol{Q}\boldsymbol{E}\boldsymbol{Q}^T$ and optimizing over $\boldsymbol{U}$ by optimizing over $\boldsymbol{Q}, \boldsymbol{E}$. □

Remark 2. *For separable penalties, the high-dimensional optimization problem* (2.27) *reduces to solving several 1-d optimization problems of the form* (25)*. These 1-d optimization problems can be solved very efficiently using standard numerical algorithms. Result 1 follows as a simple corollary, by setting the penalty to zero and noting that the maximum of* (25) *is then* $\hat{e}_r = \max\{0, d_r - 1\}$*. Note that Tipping and Bishop [1999] proved a result similar to Result 1.*

Extreme Deconvolution

The ED algorithm for solving (2.27) is due to Bovy et al. [2011] and is indeed an EM algorithm for solving weighted log-likelihood based on the data augmentation representation

(2.16). In this case, the weighted "complete data" log-likelihood is

$$\phi^{\mathrm{ED}}(\boldsymbol{U}, \boldsymbol{\Theta}; \boldsymbol{w}) = \sum_{j=1}^{n} w_j \log p(\boldsymbol{x}_j, \boldsymbol{\theta}_j \mid \boldsymbol{U}, \boldsymbol{V}_j). \tag{33}$$

Following Proposition 1, the subproblem (2.27) can be solved by the following EM steps:

- E-step: compute the posterior mean and covariance of $\boldsymbol{\theta}_j$ given current estimate of $\boldsymbol{U}$:

$$\boldsymbol{b}_j = \boldsymbol{U}(\boldsymbol{U} + \boldsymbol{V}_j)^{-1}\boldsymbol{x}_j \tag{34}$$

$$\boldsymbol{B}_j = \boldsymbol{U} - \boldsymbol{U}(\boldsymbol{U} + \boldsymbol{V}_j)^{-1}\boldsymbol{U}. \tag{35}$$

- M-step:

$$\boldsymbol{U}^{\mathrm{new}} \leftarrow \mathrm{argmax}_{\boldsymbol{U}}\, E_{\boldsymbol{\Theta}|\boldsymbol{X}}\left[\phi^{\mathrm{ED}}(\boldsymbol{U}, \boldsymbol{\Theta}; \boldsymbol{w})\right] - \rho(\boldsymbol{U}/s; \lambda). \tag{36}$$

Without a penalty, the ED update (36) has the following closed-form solution,

$$\boldsymbol{U}^{\mathrm{new}} = \sum_{j=1}^{n} \tilde{w}_j(\boldsymbol{B}_j + \boldsymbol{b}_j\boldsymbol{b}_j^T), \tag{37}$$

where $\tilde{w}_j$ are the normalized weights, $\tilde{w}_j := w_j/\bar{w}$, $\bar{w} := \sum_{j=1}^{n} w_j$. With the IW penalty, the ED update also has a closed form, which is:

$$\boldsymbol{U}^{\mathrm{new}} = \frac{\sum_{j=1}^{n} w_j(\boldsymbol{B}_j + \boldsymbol{b}_j\boldsymbol{b}_j^T) + \lambda s \boldsymbol{I}_R}{\sum_{j=1}^{n} w_j + \lambda}. \tag{38}$$

This expression is derived below. Under the NN penalty, the ED updates are not closed form, so we have not implemented them.

For simplicity, we have presented ED as solving the subproblem (2.27), which would involve iterating the updates (37) until they have converged to a stationary point (within some specified convergence tolerance). Practically speaking, however, iterating the ED updates

typically suffer from "diminishing returns" in the sense that repeated updates make smaller and smaller improvements to the likelihood. Therefore, it is often more efficient to not try to solve the subproblem accurately, and perform only a few updates. In our implementation, we run ED for one iteration in each M-step of Algorithm 1. The same approach was adopted in Bovy et al. [2011].

Derivation for ED algorithm with IW penalty The M-step involves maximizing,

$$E\big[\phi^{\mathrm{ED}}(\boldsymbol{U},\boldsymbol{\Theta},\boldsymbol{w}\big] - \rho^{IW}(\boldsymbol{U}/s) = \sum_{j=1}^{n} w_j\big(-\frac{1}{2}\log|\boldsymbol{U}| - \frac{1}{2}\mathrm{tr}[(\boldsymbol{B}+\boldsymbol{b}_j\boldsymbol{b}_j^T)\boldsymbol{U}^{-1}]\big) \\ -\frac{\lambda}{2}[\log|\boldsymbol{U}| - R\log s + \mathrm{tr}(s\boldsymbol{U}^{-1})] + \text{constant}, \tag{39}$$

where $\boldsymbol{B}$ and $\boldsymbol{b}_j$ are defined in (2.19) and (34). Denote the part of (39) that depends on $\boldsymbol{U}$ as $f(\boldsymbol{U})$. We take the (matrix) derivative of $f(\boldsymbol{U})$ with respect to $\boldsymbol{U}^{-1}$ and find the $\boldsymbol{U}^{-1}$ that sets the derivative to 0.

$$\frac{f(\boldsymbol{U})}{\partial\boldsymbol{U}^{-1}} = \frac{\sum_j^n w_j + \lambda}{2}\boldsymbol{U} - \frac{1}{2}\big(\sum_{j=1}^{n} w_j(\boldsymbol{B}_j + \boldsymbol{b}_j\boldsymbol{b}_j^T) + s\lambda\boldsymbol{I}\big) = \boldsymbol{0}. \tag{40}$$

This gives the closed-form solution

$$\boldsymbol{U}^{\mathrm{new}} = \frac{\sum_{j=1}^n w_j(\boldsymbol{B}_j + \boldsymbol{b}_j\boldsymbol{b}_j^T) + \lambda s\boldsymbol{I}_R}{\sum_{j=1}^n w_j + \lambda}. \tag{41}$$

Subspace-preserving property of ED Although the ED update is seemingly very general, it has an important limitation: the ED updates (37) have the property that they are "subspace preserving". While this limitation is not a big issue for unconstrained matrices, it makes ED poorly suited for estimating low-rank matrices, and in particular matrices with the rank-1 constraint; the ED updates (without penalty) will leave $\boldsymbol{U}$ unchanged aside from a change in scale. As far as we are aware, we are the first to report this limitation of ED.

Fortunately, there is another iterative approach which is much better suited to updating covariances $\boldsymbol{U}$ with constraints on the rank of $\boldsymbol{U}$. This is described in the next section.

A feature of the ED algorithm is that, if $\boldsymbol{U}$ is initialized to a rank-1 matrix then the ED update does not change $\boldsymbol{U}$ (or only by a multiplicative constant). Thus the ED algorithm is not suited to estimating rank-1 matrices. More generally, the ED update does not change the column space of the matrix being updated.

This behavior can be seen directly from the form of the update (37), which can be written as $\boldsymbol{U}^{\text{new}} = \boldsymbol{U}\boldsymbol{A}$, where

$$\boldsymbol{A} := \sum_{j=1}^{n} \tilde{w}_j \big[(\boldsymbol{U} + \boldsymbol{V}_j)^{-1}\boldsymbol{x}_j\boldsymbol{x}_j^T(\boldsymbol{U} + \boldsymbol{V}_j)^{-1}\boldsymbol{U} + \boldsymbol{I} - (\boldsymbol{U} + \boldsymbol{V}_j)^{-1}\boldsymbol{U}\big]. \tag{42}$$

As a result, $\text{col}(\boldsymbol{U}^{\text{new}}) = \text{col}(\boldsymbol{U}\boldsymbol{A}) \subseteq \text{col}(\boldsymbol{U})$. The column space of $\boldsymbol{U}$ is defined as

$$\text{col}(\boldsymbol{U}) = \{\boldsymbol{y} \in \mathbb{R}^R : \boldsymbol{y} = \boldsymbol{U}\boldsymbol{x}, \boldsymbol{x} \in \mathbb{R}^R\}. \tag{43}$$

If $\boldsymbol{y}' \in \text{col}(\boldsymbol{U}\boldsymbol{A})$, we can find some $\boldsymbol{x}$ such that $\boldsymbol{y}' = \boldsymbol{U}\boldsymbol{A}\boldsymbol{x} = \boldsymbol{U}\boldsymbol{x}^*$, where $\boldsymbol{x}^* = \boldsymbol{A}\boldsymbol{x}$. Therefore $\text{col}(\boldsymbol{U}\boldsymbol{A}) \subseteq \text{col}(\boldsymbol{U})$.

Factor analysis

This last approach is motivated by our interest in fitting the EBMNM model with the restriction that some covariance matrices $\boldsymbol{U}_k$ are rank-1. This constraint was also used in Urbut et al. [2019], but they used a heuristic approach to estimate these rank-1 matrices.

To impose the rank-1 constraint, we reparameterize the covariance $\boldsymbol{U}$ as $\boldsymbol{U} = \boldsymbol{u}\boldsymbol{u}^T$, where $\boldsymbol{u} \in \mathbb{R}^R$. With this parameterization, the EBMNM subproblem becomes

$$\boldsymbol{x}_j \mid \boldsymbol{u}, \boldsymbol{V}_j \sim N_R(\boldsymbol{0}, \boldsymbol{u}\boldsymbol{u}^T + \boldsymbol{V}_j). \tag{44}$$

This model admits the augmented representation in (2.20). The weighted complete data log-likelihood in this case is:

$$\phi^{\mathrm{FA}}(\boldsymbol{u}, \boldsymbol{a}; \boldsymbol{w}) = \sum_{j=1}^{n} w_j \log p(\boldsymbol{x}_j, a_j \mid \boldsymbol{u}, \boldsymbol{V}_j). \tag{45}$$

Following Proposition 1, the subproblem (2.27) can be solved by iteratively maximizing an expected (weighted) complete data log-likelihood,

$$\boldsymbol{u}^{\mathrm{new}} = \mathrm{argmax}_{\boldsymbol{u}} E_{a|\boldsymbol{X}} \left[\phi^{\mathrm{FA}}(\boldsymbol{u}, a; \boldsymbol{w})\right], \tag{46}$$

in which the expectations are taken with respect to the posterior under model (2.20) at the current estimate of $\boldsymbol{u}$. (Recall, there is no penalty term because the rank-1 constraint is instead of a penalty.) The EM steps are the following:

1. E-step: Compute the posterior mean and variance of a_j, which are

$$\mu_j = \sigma_j^2 \boldsymbol{u}^T \boldsymbol{V}_j^{-1} \boldsymbol{x}_j \tag{47}$$

$$\sigma_j^2 = 1/(1 + \boldsymbol{u}^T \boldsymbol{V}_j^{-1} \boldsymbol{u}). \tag{48}$$

2. M-step: Maximize $\mathrm{E}[\phi^{\mathrm{FA}}(\boldsymbol{u}, \boldsymbol{a}; \boldsymbol{w})]$ with respect to $\boldsymbol{u}$, which has the closed-form solution

$$\boldsymbol{u}^{\mathrm{new}} = \left(\sum_{j=1}^{n} w_j (\mu_j^2 + \sigma_j^2) \boldsymbol{V}_j^{-1} \right)^{-1} \left(\sum_{j=1}^{n} w_j \mu_j \boldsymbol{V}_j^{-1} \boldsymbol{x}_j \right). \tag{49}$$

Updating the scaling parameter

In the M-step of Algorithm 1, we update s by maximizing the part that depends on s,

$$s^{\mathrm{new}} = \underset{s>0}{\mathrm{argmax}} -\rho(\boldsymbol{U}/s). \tag{50}$$

For both the IW and NN penalties, the updates have closed-form solutions. For the IW penalty, the update is

$$s^{\text{new}} = \frac{R}{\text{tr}(\boldsymbol{U}^{-1})}. \tag{51}$$

For the NN penalty, the update is

$$s^{\text{new}} = \sqrt{\frac{\text{tr}(\boldsymbol{U})}{\text{tr}(\boldsymbol{U}^{-1})}}. \tag{52}$$

Proof. For the IW penalty, based on (2.10), we have

$$\begin{aligned}\rho_\lambda^{\mathsf{IW}}(\boldsymbol{U}/s) &= \frac{\lambda}{2}\Big[\log|\boldsymbol{U}/s| + \text{tr}\big((\boldsymbol{U}/s)^{-1}\big)\Big] \\ &= \frac{\lambda}{2}\Big[\sum_{r=1}^{R}\log e_r - R\log s + s\text{tr}(\boldsymbol{U}^{-1})\Big].\end{aligned} \tag{53}$$

Taking the first derivative of $\rho_\lambda^{\mathsf{IW}}(\boldsymbol{U}/s)$ with respect to s and set it to zero, we obtain (51).

For the NN penalty, based on (2.12), we have

$$\begin{aligned}\rho_\lambda^{\mathsf{NN}}(\boldsymbol{U}/s) &= \frac{\lambda}{2}\big(0.5\|\boldsymbol{U}/s\|_* + 0.5\|(\boldsymbol{U}/s)^{-1}\|_*\big) \\ &= \frac{\lambda}{2}\big(\frac{0.5}{s}\text{tr}(\boldsymbol{U}) + 0.5s\text{tr}(\boldsymbol{U}^{-1})\big).\end{aligned} \tag{54}$$

Taking the first derivative of $\rho_\lambda^{\mathsf{NN}}(\boldsymbol{U}/s)$ with respect to s and setting it to zero, and requiring $s > 0$, we obtain (52). □

Updating $\boldsymbol{U}$ with a scaling constraint

For the scaling constraint, in which $\boldsymbol{U} = c\boldsymbol{U}_0$ such that $\boldsymbol{U}_0$ is specified and $c > 0$ is the scalar parameter to be estimated, it is straightforward to solve the subproblem (2.25) by standard numerical methods for 1-d optimization.

A.6 Computational complexity of different algorithms

	Unconstrained	Scaled	Rank-1
TED	$O(R^3 + nR^2)$	–	$O(R^3 + nR^2)$
FA	$O(R^3 + nR^2)$	$O(npR)$	$O(nR)$
ED	$O(R^3 + nR^2)$	–	–

Table A.1: Computational complexity for homoskedastic case (when $\boldsymbol{V}_j = \boldsymbol{V}$). Per-iteration computational complexity of different algorithms for solving the subproblem when n is much larger than R. p is the rank of the canonical covariance matrix, $p \leq R$.

	Unconstrained	Scaled	Rank-1
ED	$O(nR^3)$	–	–
FA	–	$O(npR^2)$	$O(R^3 + nR^2)$

Table A.2: Computational complexity for heteroskedastic case (when $\boldsymbol{V}_j$ varies). Per-iteration computational complexity of different algorithms for solving the subproblem when n is much larger than R, in the case where $\boldsymbol{V}_j = \boldsymbol{I}$ and $\boldsymbol{V}_j$ varies. p is the rank of the canonical covariance matrix, $p \leq R$. Note that TED algorithm doesn't work for the heteroskedastic case.

A.7 Proof that changing α is equivalent to changing λ with scale invariance in the nuclear norm penalty

Let's compute $\hat{s}$ using the original form of the NN penalty from Chi and Lange [2014]:

$$\begin{aligned}\hat{s} &= \operatorname*{argmin}_{s>0} \rho_\lambda^{\mathsf{NN}}(\boldsymbol{U}/s; \alpha) \\ &= \operatorname*{argmin}_{s>0} \frac{\lambda}{2}\big[\frac{\alpha}{s}\|\boldsymbol{U}\|_* + (1-\alpha)s\|\boldsymbol{U}^{-1}\|_*\big]. \end{aligned} \tag{55}$$

This results in

$$\hat{s} = \sqrt{\frac{\alpha \mathrm{tr}(\boldsymbol{U})}{(1-\alpha)\mathrm{tr}(\boldsymbol{U}^{-1})}}. \tag{56}$$

This computation is similar to 52. Plugging $\hat{s}$ into $\rho_\lambda^{\mathsf{NN}}(\boldsymbol{U}/s;\alpha)$, we can see the term that includes α can be absorbed into λ:

$$\rho_\lambda^{\mathsf{NN}}(\boldsymbol{U}/\hat{s};\alpha) = -\lambda\sqrt{(1-\alpha)\alpha}\sqrt{\mathrm{tr}(\boldsymbol{U})\mathrm{tr}(\boldsymbol{U}^{-1})}. \tag{57}$$

A.8 Power and FSR

Given an effect estimate $\hat{\theta}_{jr}$ and $lfsr_{jr}$ for each observation $j = 1,\ldots,n$ and dimension $r = 1,\ldots,R$, we define S as the set of significant effects at threshold $t \geq 0$, CS as the set of "correctly signed" results, T as the set of true nonzero effects, and N the set of true null (zero) effects:

$$\mathrm{S} = \{j, r : lfsr_{jr} \leq t\} \tag{58}$$

$$\mathrm{CS} = \{j, r : \hat{\theta}_{jr} \times \theta_{jr} > 0\} \tag{59}$$

$$\mathrm{N} = \{j, r : \theta_{jr} = 0\} \tag{60}$$

$$\mathrm{T} = \{j, r : \theta_{jr} \neq 0\}. \tag{61}$$

Then we define true positive rate (power) and false sign rate (FSR) at *lfsr* threshold t as

$$\mathrm{TPR} = \frac{|\mathrm{CS} \cap \mathrm{S}|}{|\mathrm{T}|} \tag{62}$$

$$\mathrm{FSR} = \frac{|\mathrm{S}| - |\mathrm{CS} \cap \mathrm{S}|}{|\mathrm{S}|}. \tag{63}$$

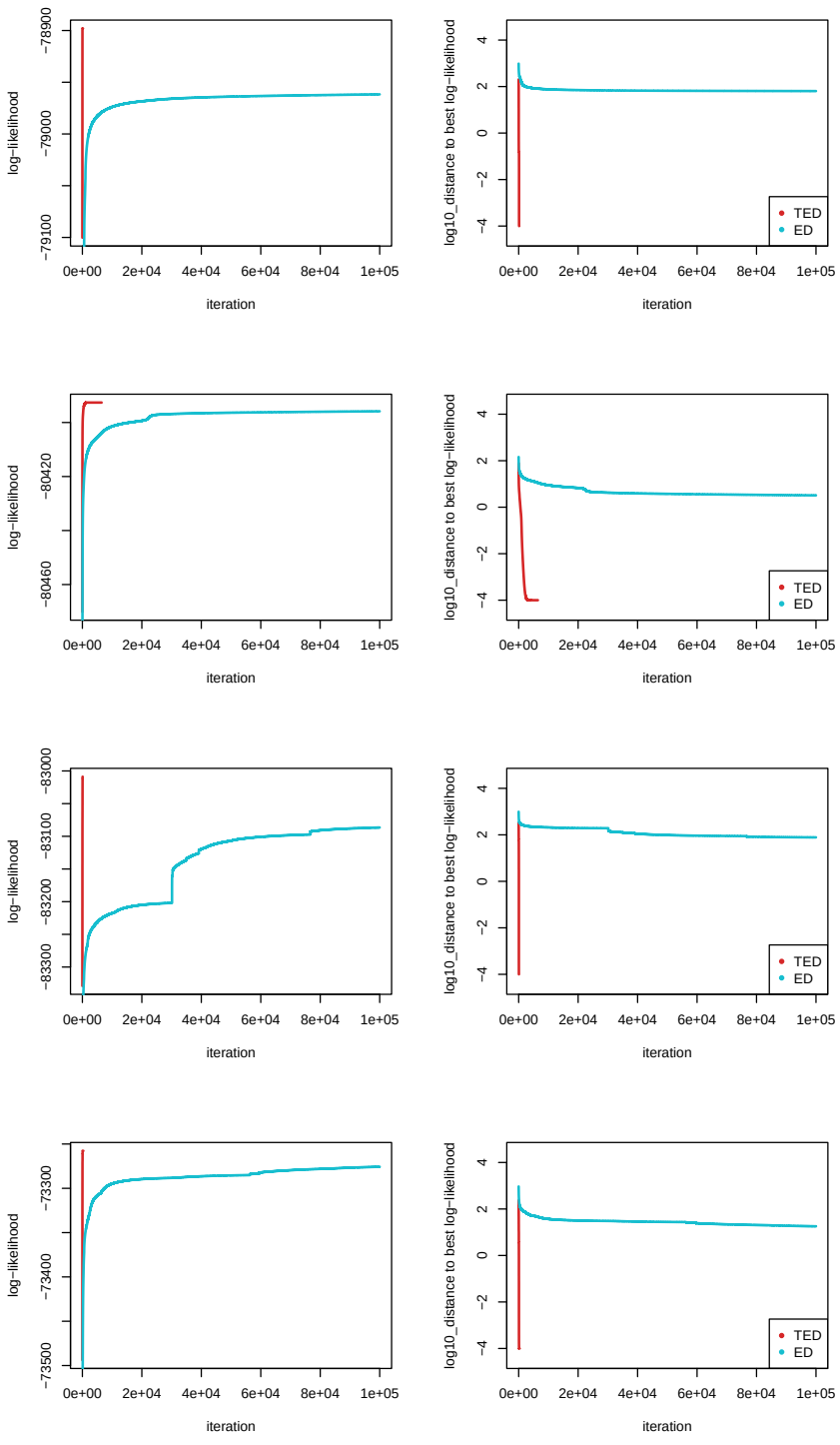

Figure A.1: Data examples for comparing the convergence between TED and ED. Each row represents one data example. We ran both algorithms for 100,000 iterations after running ED for 20 iterations initially.

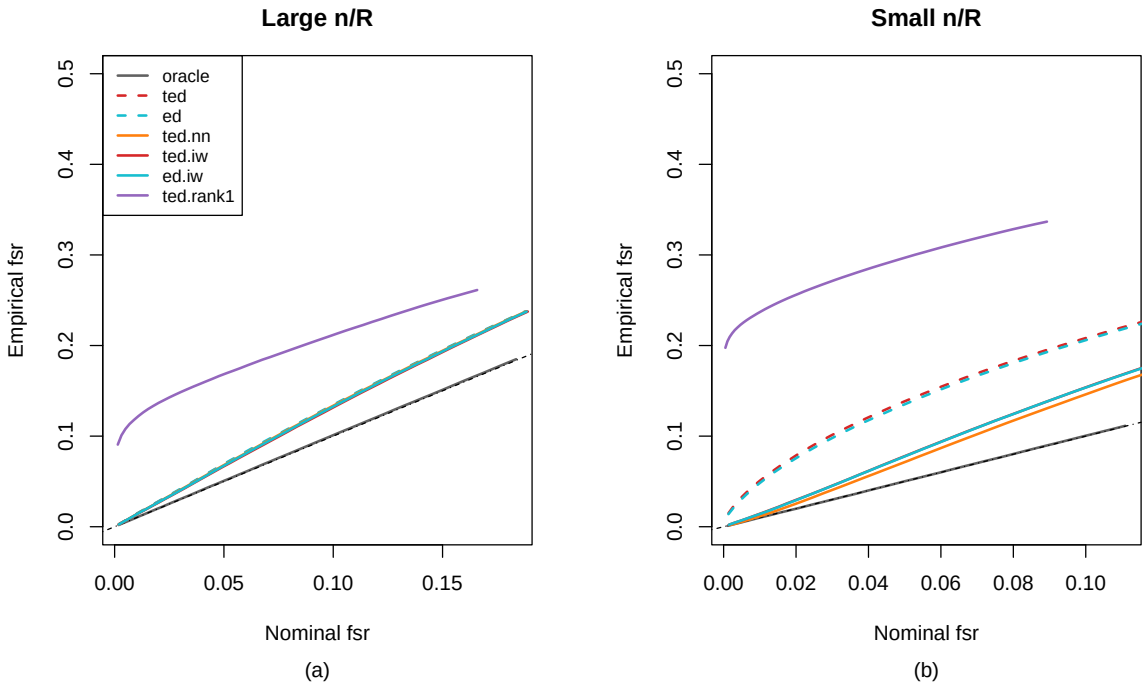

Figure A.2: Calibration of FSR for hybrid scenario. Other simulation parameters are as in Figure 2.4 in the main text. The dashed, black line represents the empirical FSR equals to nominal FSR.

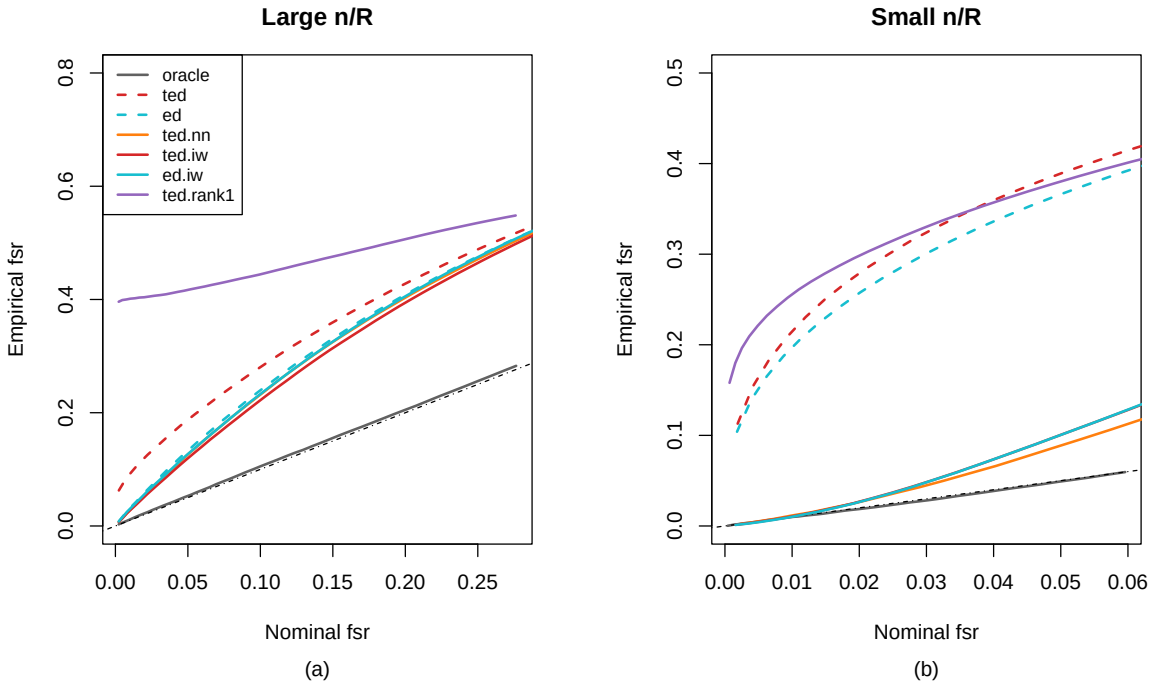

Figure A.3: Calibration of FSR where true covariances are all rank-1 matrices. Other simulation parameters are the same as Figure 2.5 in the main text. The dashed, black line represents empirical FSR equals the nominal FSR.

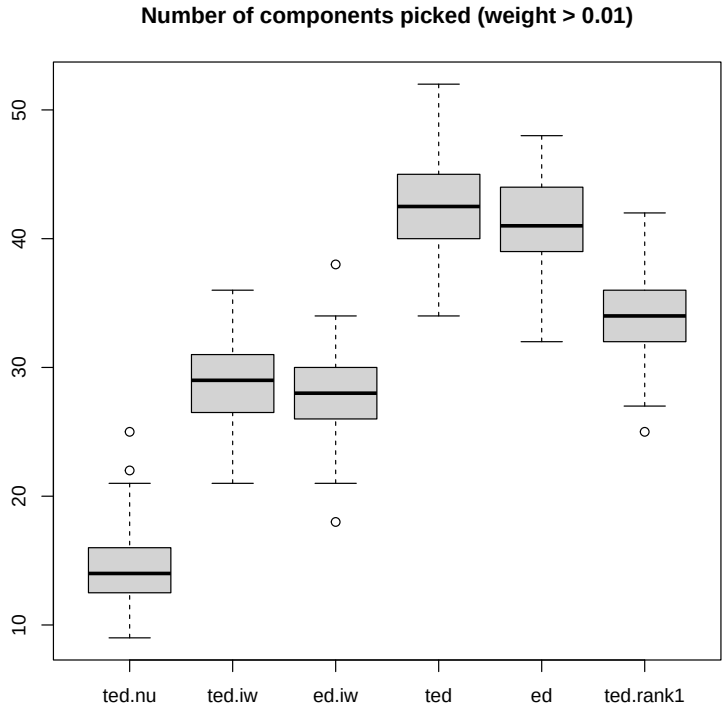

Figure A.4: The number of "important" components, defined as the components k with mixture weight $\pi_k > 0.01$. (The true K was 10.) Boxplots are based on 100 data replicates. For each simulated data set, $n = 1000$ and $R = 50$.

B CoxPH-SuSiE

B.1 Approximate Bayes factor calculation

To compute BF^{L}, we need to compute the integral in the numerator $I = \int \hat{L}_p(b)p(b)db$.

$$I = \frac{\exp\{l_p(\hat{b})\}}{\sqrt{2\pi}\sigma_0} \int \exp\{-\frac{(b-\hat{b})^2}{2s^2} - \frac{b^2}{2\sigma_0^2}\}db \tag{64}$$

$$= \frac{L_p(\hat{b})}{\sqrt{2\pi}\sigma_0} \exp\{-\hat{b}^2/2s^2\} \int \exp\{-(\frac{1}{2s^2} - \frac{1}{2\sigma_0^2})b^2 + \frac{\hat{b}b}{s^2}\}db. \tag{65}$$

Recognize the integrand is a Gaussian kernel with variance $\sigma_1^2 = \frac{1}{1/s^2+1/\sigma_0^2}$. Then,

$$I = L_p(\hat{b}) \exp\{-\frac{\hat{b}^2}{2s^2}\} \frac{\sigma_1}{\sigma_0} \exp\{\frac{\hat{b}^2\sigma_1^2}{2s^4}\} \tag{66}$$

$$= \sqrt{\frac{s^2}{\sigma_0^2+s^2}} \exp\big\{\frac{z^2}{2}\frac{\sigma_0^2}{\sigma_0^2+s^2}\big\} \exp\{-\hat{b}^2/2s^2\} L_p(\hat{b}). \tag{67}$$

B.2 An EM algorithm to estimate prior variance

With the data augmentation, the complete data likelihood is:

$$L(\sigma_0^2; \boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\gamma}, b) = p(\boldsymbol{y}, \boldsymbol{\gamma}, b|\boldsymbol{X}, \sigma_0^2) = p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\gamma}, b)p(b)p(\boldsymbol{\gamma}) \tag{68}$$

$$= \prod_{j=1}^{p} \big[p(\boldsymbol{y}|\boldsymbol{x}_j, b)p(\boldsymbol{\gamma}_j = 1)\big]^{I(\gamma_j=1)} p(b; \sigma_0^2). \tag{69}$$

The complete log-likelihood is:

$$l(\sigma_0^2) = \sum_{j=1}^{p} I(\boldsymbol{\gamma}_j = 1)\Big[\log p(\boldsymbol{y}|\boldsymbol{x}_j, b) + \log \pi_j\Big] + \log p(b; \sigma_0^2) \tag{70}$$

$$= \sum_{j=1}^{p} I(\boldsymbol{\gamma}_j = 1)\Big[\log p(\boldsymbol{y}|\boldsymbol{x}_j, b) + \log \pi_j\Big] - \frac{1}{2}\log\sigma_0^2 - \frac{b^2}{2\sigma_0^2} + \textit{constant}. \tag{71}$$

The **E-step** takes the expectation of $l(\sigma_0^2)$ w.r.t. the posterior of $\boldsymbol{\gamma}$ and the approximate posterior of $b|\boldsymbol{\gamma}$,

$$\mathrm{E}(l(\sigma_0^2)) = \sum_{j=1}^{p} \alpha_j \Big[\mathrm{E}_{b|\boldsymbol{\gamma}_j=1}\big(\log p(\boldsymbol{y}|\boldsymbol{x}_j, b)\big) + \log \pi_j\Big] - \frac{1}{2}\log\sigma_0^2 - \frac{\sum_{j=1}^{p} \alpha_j(\mu_{1j}^2 + \sigma_{1j}^2)}{2\sigma_0^2} + \textit{constant}. \tag{72}$$

The **M-step** maximizes $\mathrm{E}(l(\sigma_0^2))$ w.r.t. σ_0^2:

$$\frac{\partial \mathrm{E}(l(\sigma_0^2))}{\partial \sigma_0^2} = -\frac{1}{2\sigma_0^2} + \frac{\sum_{j=1}^{p} \alpha_j(\mu_{1j}^2 + \sigma_{1j}^2)}{2(\sigma_0^2)^2} \to 0 \tag{73}$$

$$\sigma_0^2 \leftarrow \sum_{j=1}^{p} \alpha_j(\mu_{1j}^2 + \sigma_{1j}^2). \tag{74}$$

B.3 Notes on simulating survival data

For individual i who has:

$$T_i \sim \exp(\lambda_i^s) \tag{75}$$

$$C_i \sim \exp(\lambda^c), \tag{76}$$

the probability $p(T_i > C_i)$ is $\lambda^c/(\lambda^c + \lambda_i^s)$.

Proof. For an arbitrary individual,

$$\begin{aligned}
p(T > C) &= \int_0^\infty \int_0^{t_s} \lambda^s \exp\{-\lambda^s t_s\} \lambda^c \exp\{-\lambda^c t_c\} dt_c dt_s && (77)\\
&= \int_0^\infty \lambda^s \exp\{-\lambda^s t_s\} \big(1 - \exp\{-\lambda^c t_s\}\big) dt_s && (78)\\
&= \int_0^\infty \lambda^s \exp\{-\lambda^s t_s\} dt_s - \int_0^\infty \lambda^s \exp\{-\lambda^s t_s - \lambda^c t_s\} dt_s && (79)\\
&= 1 - \frac{\lambda^s}{\lambda^s + \lambda^c} \int_0^\infty (\lambda^s + \lambda^c) \exp\{-(\lambda^s + \lambda^c) t_s\} dt_s && (80)\\
&= \frac{\lambda^c}{\lambda^s + \lambda^c}. && (81)
\end{aligned}$$

□

To generate λ^c for a corresponding censor level r, we want:

$$r = \frac{1}{n}\mathrm{E}(\sum_i^n \mathbb{1}_{C_i < T_i}) = \frac{1}{n}\sum_{i=1}^{n} P(C_i < T_i) = \frac{1}{n}\sum_{i=1}^{n} \frac{\lambda^c}{\lambda_i^s + \lambda^c}. \tag{82}$$

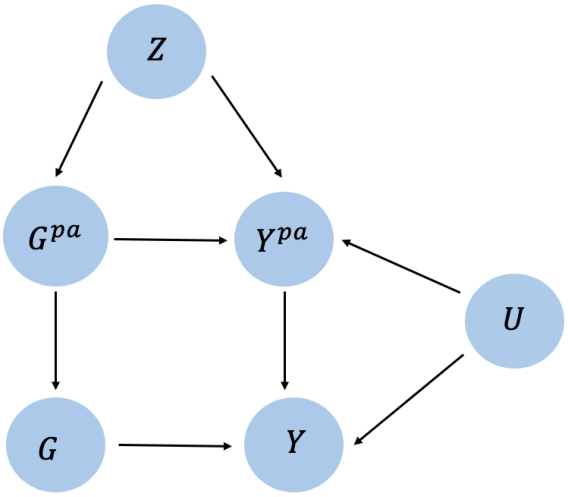

Figure C.5: Directed Acyclic Graph (DAG) of genetic nurture path. G is the whole genotype vector of an individual under consideration and Y is his/her phenotype of interest. G^{pa} is the genotype of this individual's parents and Z are the ancestry information. Y^{pa} are parental phenotypes that can be influential to this individual's phenotype Y. U is a set of unobserved non-heritable confounders.

We need to find a λ^c that solves (82), which is a little complicated. Instead, we use (3.74) for generating λ^c. Although it's not equivalent to what we really want, it simplifies the simulation procedure greatly.

C Derivations and proofs for Chapter 4

C.1 Discussion on models for sibling data

In this section, we prove the correctness of model (6) for sibling data in Table 4.1. We start from the Directed Acyclic Graph (DAG) of genetic nurture path (see Figure C.5) and the corresponding structural equation. We will also show the equivalence between using a latent familial effect f_j and using the measurable G_j^{sibs}.

The structural equation of the DAG in Figure C.5 is:

$$Y = f(G, Y^{pa}, Z, U, \epsilon), \tag{83}$$

where ϵ represents the error term. For sibling data, let's consider one family. Let $Y^{(i)}$ denote the measure of the phenotype Y on individual i; $G_j^{(i)}$ be the genotype j of individual i in the family, $\epsilon_{kj}^{(i)}$ be the residual term of genotype j for the individual i when we project onto k individuals.

Lemma 2. *Let* $\boldsymbol{G} := (G_j^{(1)}, \cdots, G_j^{(k)})$ *be the vector of children's SNP* j *in a family. Assume*

1. *The DAG in Figure C.5 is correct.*
2. *Conditioning on* G^{pa}, $G_j^{(i)}$ *are iid across* i.

Then

1. $(f(Z, U, \epsilon_{i_0}, Y^{pa}, G_j^{(i_0)}), G^{pa}, G_j^{(i)})$ *has the same distribution,* $\forall i \neq i_0$.
2. *Define*

$$\boldsymbol{\gamma} = \underset{\boldsymbol{\gamma} \in \mathbb{R}^k}{\operatorname{argmin}} \, Var\left(f(Z, U, \epsilon_{i_0}, Y^{pa}, G_j^{(i_0)}) - \sum_{i=1}^{k} \gamma_i G_j^{(i)} \right),$$

then for all $i \neq i_0$, *we have* $\gamma_i = \gamma$ *for some constant* $\gamma \in \mathbb{R}$

Using the result from Lemma 2, if we project Y onto k siblings in the family, we obtain model (6) in Table 4.1:

$$\begin{aligned}
Y^{(1)} &= \tilde{\gamma}_{1j} G_j^{(1)} + \tilde{\gamma}_{2j} \sum_{i=2}^{k} G_j^{(i)} + \epsilon_{kj}^{(1)} = (\tilde{\gamma}_{1j} - \tilde{\gamma}_{2j}) G_j^{(1)} + \tilde{\gamma}_{2j} \sum_{i=1}^{k} G_j^{(i)} + \epsilon_{kj}^{(1)} \\
Y^{(2)} &= \tilde{\gamma}_{1j} G_j^{(2)} + \tilde{\gamma}_{2j} \sum_{i \neq 2} G_j^{(i)} + \epsilon_{kj}^{(2)} = (\tilde{\gamma}_{1j} - \tilde{\gamma}_{2j}) G_j^{(2)} + \tilde{\gamma}_{2j} \sum_{i=1}^{k} G_j^{(i)} + \epsilon_{kj}^{(2)} \\
&\vdots
\end{aligned} \tag{84}$$

Then we can define $f_j = \tilde{\gamma}_{2j} \sum_{i=1}^{k} G_j^{(i)}$. Therefore, we obtain model (5) in Table 4.1 where f_j usually represents some arbitrary familial effect:

$$Y^{(i)} = (\tilde{\gamma}_{1j} - \tilde{\gamma}_{2j}) G_j^{(i)} + f_j + \epsilon_{kj}^{(i)}, \quad i = 1, 2, \ldots, k. \tag{85}$$

C.2 The joint asymptotic distribution of $(\hat{\tau} - \tau^*, \hat{\alpha}_1 - \hat{\alpha}_1')$

Let $\hat{\boldsymbol{b}}_j$ solves the estimating equations of the full model (4.2) on internal data,

$$\sum_{i=1}^{n} \boldsymbol{s}_{ij}(\boldsymbol{b}_j) := \sum_{i=1}^{n} (y_i - \mu_{ij}) \begin{pmatrix} 1 \\ G_{ij} \\ F_{ij} \end{pmatrix} = \sum_{i=1}^{n} (y_i - g^{-1}(\eta_{ij})) \begin{pmatrix} 1 \\ G_{ij} \\ F_{ij} \end{pmatrix} = \mathbf{0}. \tag{86}$$

Let $\hat{\boldsymbol{\alpha}}_j$ and $\hat{\boldsymbol{\alpha}}_j'$ solve the estimating equations of the reduced model (4.4) on internal and external data,

$$\sum_{i} \tilde{\boldsymbol{s}}_{ij}(\boldsymbol{\alpha}_j) := \sum_{i} (y_i - \tilde{\mu}_{ij}) \begin{pmatrix} 1 \\ G_{ij} \end{pmatrix} = \mathbf{0}. \tag{87}$$

Following Chen and Chen [2000], we present the following lemma which gives the asympototic joint distribution of $(\hat{\boldsymbol{b}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\alpha}}')$ (index j omitted here). The result is derived directly using Taylor expansion and the central limit theorem.

Lemma 3. *Let* $\boldsymbol{b}^*$ *be the true value of* $\boldsymbol{b}$*,* $\boldsymbol{\alpha}^*$ *be the value minimize the Kullback Leibler divergence between model* (4.2) *and model* (4.4)*.* $(\hat{\boldsymbol{b}}, \hat{\boldsymbol{\alpha}}', \hat{\boldsymbol{\alpha}})$ *is consistent for* $(\boldsymbol{b}^*, \boldsymbol{\alpha}^*, \boldsymbol{\alpha}^*)$ *and the joint distribution of* $(\hat{\boldsymbol{b}}, \hat{\boldsymbol{\alpha}}', \hat{\boldsymbol{\alpha}})$ *is asymptotically normal with mean* $\mathbf{0}$ *and variance given by the follows:*

$$\sqrt{n} \begin{pmatrix} \hat{\boldsymbol{b}} - \boldsymbol{b}^* \\ \hat{\boldsymbol{\alpha}}' - \boldsymbol{\alpha}^* \\ \hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^* \end{pmatrix} \to \mathcal{N} \left(\mathbf{0}, \begin{pmatrix} \boldsymbol{D}_1^{-1}\boldsymbol{C}_{11}\boldsymbol{D}_1^{-1} & \boldsymbol{D}_1^{-1}\boldsymbol{C}_{12}\boldsymbol{D}_2^{-1} & \rho\boldsymbol{D}_1^{-1}\boldsymbol{C}_{12}\boldsymbol{D}_2^{-1} \\ \boldsymbol{D}_2^{-1}\boldsymbol{C}_{21}\boldsymbol{D}_1^{-1} & \boldsymbol{D}_2^{-1}\boldsymbol{C}_{22}\boldsymbol{D}_2^{-1} & \rho\boldsymbol{D}_2^{-1}\boldsymbol{C}_{22}\boldsymbol{D}_2^{-1} \\ \rho\boldsymbol{D}_2^{-1}\boldsymbol{C}_{21}\boldsymbol{D}_1^{-1} & \rho\boldsymbol{D}_2^{-1}\boldsymbol{C}_{22}\boldsymbol{D}_2^{-1} & \frac{n}{N}\boldsymbol{D}_2^{-1}\boldsymbol{C}_{22}\boldsymbol{D}_2^{-1} \end{pmatrix} \right) \equiv \mathcal{N}(\mathbf{0}, \boldsymbol{V}), \tag{88}$$

where

$$\boldsymbol{D}_1 \equiv E\left(\frac{\partial \boldsymbol{s}(\boldsymbol{b}^*)}{\partial \boldsymbol{b}}\right), \tag{89}$$

$$\boldsymbol{D}_2 \equiv E\left(\frac{\partial \tilde{\boldsymbol{s}}(\boldsymbol{\alpha}^*)}{\partial \boldsymbol{\alpha}}\right), \tag{90}$$

$$\boldsymbol{C} \equiv E\Big[\left\{\boldsymbol{s}(\boldsymbol{b}^*)', \tilde{\boldsymbol{s}}(\boldsymbol{\alpha}^*)'\right\}'\left\{\boldsymbol{s}(\boldsymbol{b}^*)', \tilde{\boldsymbol{s}}(\boldsymbol{\alpha}^*)'\right\}\Big]. \tag{91}$$

And $\boldsymbol{C}_{11}, \boldsymbol{C}_{12}$ *and* $\boldsymbol{C}_{22}$ *denote the submatrices of* $\boldsymbol{C}$*, where*

$$\boldsymbol{C}_{11} \equiv E\left\{\mathbf{s}\left(\boldsymbol{b}^*\right)\mathbf{s}'\left(\boldsymbol{b}^*\right)\right\}$$

$$\boldsymbol{C}_{12} \equiv E\left\{\mathbf{s}\left(\boldsymbol{b}^*\right)\tilde{\mathbf{s}}'\left(\boldsymbol{\alpha}^*\right)\right\}$$

$$\boldsymbol{C}_{22} \equiv E\left\{\tilde{\mathbf{s}}\left(\boldsymbol{\alpha}^*\right)\tilde{\mathbf{s}}'\left(\boldsymbol{\alpha}^*\right)\right\}.$$

n and N denote the sample size of internal and external data. ρ denotes the sample overlapping ratio between internal and external data.

To find the asymptotic distribution of $(\hat{\tau} - \tau^*, \hat{\alpha}_1 - \hat{\alpha}_1')$, we multiply a suitable matrix $\boldsymbol{A}$ to $(\hat{\boldsymbol{b}} - \boldsymbol{b}^*, \hat{\boldsymbol{\alpha}}' - \boldsymbol{\alpha}^*, \hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*)$. Therefore,

$$\sqrt{n}\begin{pmatrix} \hat{\tau} - \tau^* \\ \hat{\alpha}_1' - \hat{\alpha}_1 \end{pmatrix} \to \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \boldsymbol{A}\boldsymbol{V}\boldsymbol{A}^T\right) \equiv \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} v_{11} & v_{12} \\ v_{21} & v_{22} \end{bmatrix}\right). \tag{92}$$

C.3 Covariance estimation in the asymptotic normal distribution

The estimated $\hat{v}_{11}, \hat{v}_{12}$ and $\hat{v}_{22}$ are obtained by estimating $\boldsymbol{D}_1, \boldsymbol{D}_2, \boldsymbol{C}_{11}, \boldsymbol{C}_{12}, \boldsymbol{C}_{22}$ using sample quantities, which are as follows when observations are i.i.d. in the internal data:

$$\hat{\boldsymbol{D}}_1 = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \boldsymbol{s}_i(\hat{\boldsymbol{b}})}{\partial \boldsymbol{b}} \tag{93}$$

$$\hat{\boldsymbol{D}}_2 = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \tilde{\boldsymbol{s}}_i(\hat{\boldsymbol{\alpha}}')}{\partial \boldsymbol{\alpha}} \tag{94}$$

$$\hat{\boldsymbol{C}}_{11} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{s}_i(\hat{\boldsymbol{b}})\boldsymbol{s}_i(\hat{\boldsymbol{b}})' \tag{95}$$

$$\hat{\boldsymbol{C}}_{12} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{s}_i(\hat{\boldsymbol{b}})\tilde{\boldsymbol{s}}_i(\hat{\boldsymbol{\alpha}}')' \tag{96}$$

$$\hat{\boldsymbol{C}}_{22} = \frac{1}{n}\sum_{i=1}^{n} \tilde{\boldsymbol{s}}_i(\hat{\boldsymbol{\alpha}}')\tilde{\boldsymbol{s}}_i(\hat{\boldsymbol{\alpha}}')'. \tag{97}$$

This is suitable for scenario (1) through (4) in Table 4.1. For sibling data in scenario (5) and (6), estimating $\boldsymbol{C}_{11}$ becomes more challenging due to the correlation between siblings within families. Let $H_k = \{h_{k1}, h_{k2}, \ldots, h_{kn_k}\}$ be the set of families with k siblings in the data set for $k = 2, \ldots, K$. Then

$$\frac{1}{n}\mathrm{Var}\left(\sum_{i=1}^{n} \boldsymbol{s}_i(\boldsymbol{b}^*)\right) = \mathrm{Var}\left(\sum_{k=2}^{K} \frac{n_k}{n}\left(\frac{1}{n_k}\sum_{r=1}^{n_k}\sum_{i\in h_{kr}} \boldsymbol{s}_i(\boldsymbol{b}^*)\right)\right) =: \sum_{k=2}^{K}\frac{n_k}{n}\boldsymbol{V}_k \tag{98}$$

Due to the independence across families, $\boldsymbol{V}_k$ can be estimated by

$$\hat{\boldsymbol{V}}_k = \frac{1}{n_k}\sum_{r=1}^{n_k}\left(\sum_{i\in h_{kr}} \boldsymbol{s}_i(\hat{\boldsymbol{b}})\right)\left(\sum_{i\in h_{kr}} \boldsymbol{s}_i(\hat{\boldsymbol{b}})\right)'. \tag{99}$$

C.4 Derivation for theoretical variance reduction in trio data

We derive the theoretical variance reduction under linear regression model in scenario (1). We assume the correct model is:

$$Y = b_0 + b_1 T + b_2 NT + \epsilon,$$

where $(T, NT) \perp\!\!\!\perp \epsilon$, $\epsilon \sim \mathcal{N}(0, \sigma^2)$. In addition, we assume T and NT have zero expectation, unit variance and correlation r. Let $X := (1, T, NT)^\top$ be the covariates.

Consider the reduced(misspecified) model:

$$Y = \alpha_0 + \alpha_1 T + \epsilon'.$$

Let $\tilde{X} := (1, T)^\top$ be the covariates. The estimating equation under the reduced model is:

$$\tilde{s}(\boldsymbol{\alpha}) = (Y - \alpha_0 - \alpha_1 T) \begin{pmatrix} 1 \\ T \end{pmatrix} = (Y - \tilde{X}^\top \boldsymbol{\alpha}) \tilde{X} \tag{100}$$

Then

$$\begin{aligned} & \mathbb{E}_{(X,Y)} [\tilde{s}(\boldsymbol{\alpha})] = \mathbf{0} \\ \Longrightarrow & \mathbb{E}_{(X,Y)} \left[(Y - \tilde{X}^T \boldsymbol{\alpha}) \tilde{X} \right] = \mathbf{0} \\ \Longrightarrow & \mathbb{E}_X \left[\mathbb{E}_{Y|X} (Y - \tilde{X}^T \boldsymbol{\alpha}) \tilde{X} \right] = \mathbf{0} \\ \Longrightarrow & \mathbb{E}_X [\tilde{X} (X^T \boldsymbol{b} - \tilde{X}^T \boldsymbol{\alpha})] = \mathbf{0} \end{aligned}$$

From the first equation, we can get $\alpha_0 = b_0$ and $\alpha_1 = b_1 + b_2 r$. It follows that

$$\mathrm{Var}(\epsilon') = \mathrm{Var}((b_1 - \alpha_1) T + b_2 NT + \epsilon) = \mathrm{Var}(-r b_2 T + b_2 NT + \epsilon) = \sigma^2 + (1 - r^2) b_2^2$$

.

Finding $\boldsymbol{D}$ and $\boldsymbol{C}$

The score equation for the correct model is given by

$$s_i(\boldsymbol{\alpha}) = (Y_i - \tilde{X}_i^T \boldsymbol{\alpha})\tilde{X}_i, \quad s_i(\boldsymbol{b}) = (Y_i - X_i^T \boldsymbol{b})X_i$$

Therefore

- $\boldsymbol{D}_1 = \mathbb{E}\left[\frac{\partial s(\boldsymbol{b})}{\partial \boldsymbol{b}}\right] = \mathbb{E}\left[X_i^T X_i\right] = -\Sigma_X$
- $\boldsymbol{D}_2 = \mathbb{E}\left[\frac{\partial \tilde{s}(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}}\right] = \mathbb{E}\left[\tilde{X}_i^T \tilde{X}_i\right] = -\Sigma_{\tilde{X}}$
- $\boldsymbol{C}_{11} = \mathbb{E}\left[s(\boldsymbol{b})s(\boldsymbol{b})^T\right] = \mathbb{E}\left[(Y_i - X_i^T \boldsymbol{b})^2 X_i X_i^\top\right]/\sigma^4 = \sigma^2 \Sigma_X$
- $\boldsymbol{C}_{22} = \mathbb{E}\left[s(\boldsymbol{\alpha})s(\boldsymbol{\alpha})^T\right] = \mathbb{E}\left[(Y_i - \tilde{X}_i^T \boldsymbol{\alpha})^2 \tilde{X}_i \tilde{X}_i^\top\right]/(\sigma')^4 = (\sigma')^2 \Sigma_{\tilde{X}}$
- $\boldsymbol{C}_{12} = \mathbb{E}\left[s(\boldsymbol{b})s(\boldsymbol{b})^T\right] = \mathbb{E}\left[(Y_i - X_i^T \boldsymbol{b})(Y_i - X_i^T \boldsymbol{\alpha})X_i \tilde{X}_i^\top\right] = \sigma^2 \Sigma_{X,\tilde{X}}$,

where

$$\Sigma_X = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & r \\ 0 & r & 1 \end{pmatrix}, \quad \Sigma_{\tilde{X}} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}, \quad \Sigma_{X,\tilde{X}} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \\ 0 & r \end{pmatrix}$$

Computation of variation reduction

From previous results, we know that

$$\sqrt{n}\begin{pmatrix} \hat{\boldsymbol{b}} - \boldsymbol{b} \\ \hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha} \\ \hat{\boldsymbol{\alpha}}' - \boldsymbol{\alpha} \end{pmatrix} \to \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \boldsymbol{D}_1^{-1}\boldsymbol{C}_{11}\boldsymbol{D}_1^{-1} & \boldsymbol{D}_1^{-1}\boldsymbol{C}_{12}\boldsymbol{D}_2^{-1} & \rho\boldsymbol{D}_1^{-1}\boldsymbol{C}_{12}\boldsymbol{D}_2^{-1} \\ \boldsymbol{D}_2^{-1}\boldsymbol{C}_{21}\boldsymbol{D}_1^{-1} & \boldsymbol{D}_2^{-1}\boldsymbol{C}_{22}\boldsymbol{D}_2^{-1} & \rho\boldsymbol{D}_2^{-1}\boldsymbol{C}_{22}\boldsymbol{D}_2^{-1} \\ \rho\boldsymbol{D}_2^{-1}\boldsymbol{C}_{21}\boldsymbol{D}_1^{-1} & \rho\boldsymbol{D}_2^{-1}\boldsymbol{C}_{22}\boldsymbol{D}_2^{-1} & \frac{n}{N}\boldsymbol{D}_2^{-1}\boldsymbol{C}_{22}\boldsymbol{D}_2^{-1} \end{pmatrix}\right)$$
$$\equiv \mathcal{N}(0, \boldsymbol{V}) ,$$

where n is the sample size of the internal data, N is the sample size of the external data, and ρ is the sample overlap ratio between internal data and external data.

Substituting the values of $\boldsymbol{D}$ and $\boldsymbol{C}$, we get

$$\boldsymbol{V} = \begin{bmatrix} \sigma^2 & 0 & 0 & \sigma^2 & 0 & \rho\sigma^2 & 0 \\ 0 & \frac{\sigma^2}{1-r^2} & \frac{-r\sigma^2}{1-r^2} & 0 & \sigma^2 & 0 & \rho\sigma^2 \\ 0 & \frac{-r\sigma^2}{1-r^2} & \frac{\sigma^2}{1-r^2} & 0 & 0 & 0 & 0 \\ \sigma^2 & 0 & 0 & (\sigma')^2 & 0 & \rho(\sigma')^2 & 0 \\ 0 & \sigma^2 & 0 & 0 & (\sigma')^2 & 0 & \rho(\sigma')^2 \\ \rho\sigma^2 & 0 & 0 & \rho(\sigma')^2 & 0 & \frac{n}{N}(\sigma')^2 & 0 \\ 0 & \rho\sigma^2 & 0 & 0 & \rho(\sigma')^2 & 0 & \frac{n}{N}(\sigma')^2 \end{bmatrix}$$

By simple linear algebra we have

$$\sqrt{n}\begin{pmatrix} \hat{b}_1 - \hat{b}_2 \\ \hat{\alpha}_1' - \hat{\alpha}_1 \end{pmatrix} \to \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} v_{11} & v_{12} \\ v_{21} & v_{22} \end{bmatrix}\right) = \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} \frac{2\sigma^2}{1-r} & \sigma^2(1-\rho) \\ \sigma^2(1-\rho) & (\sigma')^2\left(1+\frac{n}{N}-2\rho\right) \end{bmatrix}\right)$$

Thus, using (4.7) , the variance reduction is given by

$$\mathrm{VR} = \frac{v_{12}^2}{v_{11}v_{22}} = \frac{\sigma^2(1-\rho)^2(1-r)}{2(\sigma')^2(1+\frac{n}{N}-2\rho)} = \frac{\sigma^2(1-\rho)^2(1-r)}{2\left(\sigma^2+b_2^2(1-r^2)\right)\left(1+\frac{n}{N}-2\rho\right)}.$$

When $\rho = 0$, we get the equation in the main text.