Statistical Methods for Analysis of Genetic and Genomic Data in Population Science

Statistical Methods for Analysis of Genetic and Genomic Data in Population Science

A dissertation presented

by

Richard Thomas Barfield

to

The Department of Biostatistics

in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of
Biostatistics

Harvard University
Cambridge, Massachusetts

April 2017

Dissertation Advisors: Xihong Lin Richard Thomas Barfield

Statistical Methods for Analysis of Genetic and Genomic Data in Population Science

Abstract

In chapter 1, we develop a missing mediator analysis using the EM algorithm for studies where the mediator is a genomic marker. Typically measures such as DNA methylation or gene expression are collected on a subset of participants from a larger study. Under standard assumptions for mediation analysis and an additional assumption that the missing data mechanism is ignorable, we can estimate the causal direct and indirect effects using all individuals with exposure and outcome. We applied our method to Project Viva to assess whether cord blood DNA methylation mediates the effect of maternal prepregnancy BMI on childhood BMI.

In chapter 2, we develop a statistical method to estimate cell specific associations in whole blood DNA methylation data which is a mixture of several cell types using observed cell composition when cell-specific methylations are not observed. We use Generalized Estimating Equations to estimate cell specific exposure effects using observed whole blood methylation and cell type count data. We evaluated the performance of the proposed methods through simulation studies and analyzing data from the Normative Aging Study to assess for cell specific smoking associations on 49 probes established to be associated with smoking on the aggregate csale.

In chapter 3, we introduce a novel approach to help differentiate when multiple eQTL genes co-localize at disease loci (due to linkage disequilibrium, LD), to help in identifying the true susceptible gene. We developed LD aware MR-Egger regression, an extension of MR-Egger regression to when multiple SNPs in LD are associated with gene expression. This approach requires only summary GWAS and eQTL effects, along with LD from reference panels. Through simulations we show that when SNPs have direct (pleiotropic)

effects, our approach provides adequate control of type I error, high power, and less bias than previously proposed methods under certain conditions. We analyzed summary data from a GWAS on the risk of Breast Cancer with eQTL data from breast tissue from GTEx to demonstrate the usefulness of this method.

Contents

Acknowledgments

I would like to thank my advisors and committee members: Xihong Lin, Peter Kraft and Tyler VanderWeele. The expertise and advice you all provided was invaluable in this dissertation process. I would also like to thank Andrea Baccarelli for all of his advice.

I wish to thank all of my friends here in Boston that made the last five years worthwhile.

Most importantly, I wish to thank my parents for their unwavering support. I could not have done this without you both.

Mediation Analysis in the Presence of Partially Missing Data on the Mediator in Genomic Studies

Richard Thomas Barfield

Department of Biostatistics

Harvard Graduate School of Arts and Sciences

Xihong Lin

Department of Biostatistics

Harvard Chan School of Public Health

Department of Statistics

Harvard University

1.1 Introduction

Genomic studies are often part of a larger study, where measures such as DNA methylation or gene expression are collected on a subset of the original study participants. Which observations are selected to be genotyped is often based on factors such as funding, genomic consent or extreme phenotypes. When researchers want to test whether a genomic measure lies on a biological pathway from an exposure to an outcome via mediation analysis, the analysis is restricted to the individuals with genomic data collected. However, the participants without genomic data can still provide information. In this paper, we develop an EM algorithm to perform mediation analysis when there is missing data on the mediator thus, utilizing the participants without genomic data.

Mediation analysis is an informative statistical framework for assessing when an exposure's effect on an outcome is mediated by an intermediate variable ((VanderWeele and Vansteelandt, 2009; Baron and Kenny, 1986)). Given certain assumptions or if the study is performed in a randomized clinical trial, the effects from mediation analysis will have a causal interpretation. Researchers are often interested in the direct effect of the exposure on the outcome (natural direct effect), the effect through the mediator (natural indirect effect), and the natural total effect (sum of the direct and indirect effects).

One of the most widely used approaches for accounting for missing data is the Expectation Maximization (EM) algorithm (Dempster et al. (1977)). This algorithm estimates the parameters of interest by maximizing the likelihood where the unobserved data has been marginalized out. If the missing data mechanism is ignorable, it will converge to consistent estimates of the parameters of interest. The missing data mechanism is ignorable if the observations are missing at random (MAR) or completely at random (MCAR). Data is MAR, if conditional on observed variables, the missing data mechanism is independent of the unobserved data. MCAR implies that the mechanism is independent of any variable of interest. In genomic studies the data are often collected based on monetary restrictions or individuals signing genomic consent. It is then reasonable to assume that the missing data mechanism is ignorable. Mediation analysis specifies a model for the mediator and the outcome, making implementing the EM algorithm straightfor-

ward. The algorithm uses conditional expectation of the missing data moments, which are functions of the parameters of interest and the observed data.

In the psychology and sociology literature, there is a small missing mediation literature. There are Bayesian methods to account for missing data in outcome, mediator, exposure or confounders (Enders et al., 2013). However, given the computational time of an MCMC on the genomic scale, we sought to utilize an EM algorithm. Zhang and Wang examined several methods including multiple imputation and an EM algorithm that ignores the model parameters and finds an estimate for each unobserved mediator based on an unconstrained mean and covariance matrix, and upon convergence includes in the model as if it was observed (Zhang and Wang, 2013). Multiple imputation has been studied extensively by others as well (Wu and Jia, 2013) but has the same issue of computational time as the Bayesian MCMC. There is some work on Full Information Maximization Likelihood (FIML), but a detailed step by step algorithm is not provided, or details on potential independence between the outcome and mediator model parameters has not been addressed.

With the rise of "omic" data, mediation analysis can inform how genomic features such as DNA methylation (DNAm) can mediate the effect of an exposure on an outcome. DNAm is an epigenetic modification that occurs when a methyl group is added to a cytosine followed by a guanine on the genome and is a regulatory mechanism of gene expression. It is of particular interest to researchers as it can be affected by environmental exposures. DNAm has been shown to mediate the effects of arthritis (Liu et al., 2013), and smoking status (Kupers et al., 2015; Zhang et al., 2016). In this paper, we developed an algorithm to estimate causal effects in the presence of partially missing data on the mediator in the context of DNAm studies.

We applied our method to DNAm from Project Viva, a cohort of women enrolled pre-birth and their subsequent children from Easter Massachusetts (Oken et al., 2015). The study aims to better understand the relation between pre-pregnancy diet and other factors on maternal and child health. Women were recruited from 1999 to 2002 from obstetric offices of Atrius Harvard Vanguard Medical Associates. DNAm samples were collected from umbilical cord blood at time of delivery and were later assayed in 2014.

We tested for a natural indirect effect of maternal pre-pregnancy BMI through DNAm on childhood BMI measured at 6 months, 3 years, and 7 years.

The following paper is organized as follows. Section 2 discusses the mediation models and assumptions for causal effects in the absence or presence of missing data. Section 3 presents the estimation procedure for the parameters introduced in Section 2. Section 4 gives a theoretical result on the covariance of the estimated parameters and how to test the causal effects. Section 5 presents simulations studies and performance of our EM algorithm for various missing data scenarios. Section 6 is an application to Project Viva and is followed by discussion and conclusion.

1.2 The Model

1.2.1 No Missing Data

We collect information on n independent observations for an outcome (Y_i), a mediator (M_i), an exposure (A_i), and p confounders ($\boldsymbol{X}_i$). We have the following models for M_i and Y_i for individual i:

$$M_i = \gamma_0 + A_i\gamma_A + \boldsymbol{X}_i^T\boldsymbol{\gamma}_C + \epsilon_{M,i}; \epsilon_{M,i} \sim N(0, \sigma_M^2), \tag{1.1}$$

$$Y_i = \beta_0 + A_i\beta_A + M_i\beta_M + \boldsymbol{X}_i^T\boldsymbol{\beta}_C + \epsilon_{Y,i}; \epsilon_{Y,i} \sim N(0, \sigma_Y^2). \tag{1.2}$$

We assume the residuals ($\epsilon_{M,i}$ and $\epsilon_{Y,i}$) are independent. Define $\boldsymbol{\beta} = (\beta_0, \beta_A, \beta_M, \boldsymbol{\beta}_C^T)^T$, $\boldsymbol{\gamma} = (\gamma_0, \gamma_A, \boldsymbol{\gamma}_C^T)^T$, and $\boldsymbol{\Theta}^T = (\boldsymbol{\beta}^T, \boldsymbol{\gamma}^T, \sigma_Y^2, \sigma_M^2)$ as the parameters we seek to estimate. We have the following probability density functions for Y_i and M_i:

$$\begin{aligned}
f(M_i|A_i, \boldsymbol{X}_i, \boldsymbol{\Theta}) &= \phi(M_i, \gamma_0 + A_i\gamma_A + \boldsymbol{X}_i^T\boldsymbol{\gamma}_C, \sigma_M^2),\\
f(Y_i|M_i, A_i, \boldsymbol{X}_i, \boldsymbol{\Theta}) &= \phi(Y_i, \beta_0 + A_i\beta_A + M_i\beta_M + \boldsymbol{X}_i^T\boldsymbol{\beta}_C, \sigma_Y^2),\\
f(Y_i, M_i|A_i, \boldsymbol{X}_i, \boldsymbol{\Theta}) &= f(Y_i|M_i, A_i, \boldsymbol{X}_i, \boldsymbol{\Theta})f(M_i|A_i, \boldsymbol{X}_i, \boldsymbol{\Theta}),
\end{aligned}$$

where $\phi(Z_i, \mu, \sigma^2)$ is the p.d.f of a normal evaluated at Z_i with mean μ and variance σ^2. The likelihood of the data given $A_i, \boldsymbol{X}_i, \boldsymbol{\Theta}$ is then:

$$L_c = \prod_{i=1}^{n} f(Y_i, M_i|A_i, \boldsymbol{X}_i, \boldsymbol{\Theta}). \tag{1.3}$$

We now discuss when we have a causal interpretation of the parameters. Define the counter factual notation $Y_{(a,M)}$ as the value of Y that would be observed if A was set to a and M to m (VanderWeele and Vansteelandt (2009)). For the parameters to have causal interpretation, we require six assumptions: 1) there are no unmeasured exposure-outcome confounders given $\boldsymbol{X}\{Y_{(a,m)} \perp A|\boldsymbol{X}\}$, 2) there are no unmeasured mediator-outcome confounders given $\boldsymbol{X}$and A $\{Y_{(a,m)} \perp M|\boldsymbol{X}, A\}$, 3) there are no unmeasured exposure-mediator confounders given $\boldsymbol{X}\{M_{(a)} \perp A|\boldsymbol{X}\}$, and 4) there is no effect of the exposure that confounds the mediator-outcome relationship $\{Y_{(a,m)} \perp M_{(a*)}|\boldsymbol{X}\}$ (VanderWeele and Vansteelandt (2009)). We also require that the parameters occur in a sequential order and the assumption of consistency $\{Y_{(a,m)} = E(Y|A = a, M = m)\}$ be satisfied. Given these assumptions we can derive causal interpretation of the parameters in (1.1) and (1.2).

In this paper, we focus on three causal effects: the natural direct effect (NDE), natural indirect effect (NIE), and the natural total effect (NTE). The NDE is the effect of the exposure on the outcome at level a vs a^* if the mediator was set to $M_{(a^*)}$. The NIE is the effect of the exposure only through the mediator. Finally, the NTE is the sum of the NDE and NIE. Mathematically:

$$NDE = E(Y_{(a,M_{a^*})} - Y_{(a^*,M_{a^*})}|\boldsymbol{X}) = \beta_A(a - a^*), \tag{1.4}$$

$$NIE = E(Y_{(a,M_a)} - Y_{(a,M_{a^*})}|\boldsymbol{X}) = \beta_M\gamma_a(a - a^*), \tag{1.5}$$

$$NTE = E(Y_{(a,M_a)} - Y_{(a^*,M_{a^*})}|\boldsymbol{X}) = NIE + NDE = (\beta_a + \beta_m\gamma_a)(a - a^*). \tag{1.6}$$

1.2.2 Incorporating incomplete observations

We now only observe a subset of M_i. Our sample n has been split in two, with n_c observations having observed M_i, and n_m without. Define R_i as an indicator variable of whether M_i is observed ($\sum_{i=1}^{n} R_i = n_c$). In addition to the causal assumptions from the previous subsection, we make the additional assumption that the missing data mechanism is ignorable, thus removing the need to model R_i. If this assumption is violated, estimation

of $\boldsymbol{\Theta}$ may be biased. The likelihood of our observed data is now:

$$L_m = \prod_{i=1}^{n} f\left(Y_i, M_i | A_i, \boldsymbol{X}_i, \boldsymbol{\Theta}\right)^{R_i} \left\{ \int_{-\infty}^{\infty} f(Y_i, M_i | A_i, \boldsymbol{X}_i, \boldsymbol{\Theta}) \partial M_i \right\}^{1-R_i}. \tag{1.7}$$
$$= \prod_{i=1}^{n} f(Y_i, M_i | A_i, \boldsymbol{X}_i, \boldsymbol{\Theta})^{R_i} f(Y_i | A_i, \boldsymbol{X}_i, \boldsymbol{\Theta})^{1-R_i},$$

where we have the following moments of Y_i after integration

$$E(Y_i | A_i, \boldsymbol{X}_i, \boldsymbol{\Theta}) = \beta_0 + A_i \beta_A + (\gamma_0 + A_i \gamma_A + \boldsymbol{X}_i^T \boldsymbol{\gamma}_C)\beta_M + \boldsymbol{X}_i^T \boldsymbol{\beta}_C,$$
$$\text{var}(Y_i | A_i, \boldsymbol{X}_i, \boldsymbol{\Theta}) = \beta_M^2 \sigma_M^2 + \sigma_Y^2.$$

In the next section, we present estimation of $\boldsymbol{\Theta}$ with and without missing data.

1.3 Estimation

1.3.1 No Missing Data

If $R_i = 1$ for all observations, we can estimate $\boldsymbol{\Theta}$ using standard regression techniques. Define $\boldsymbol{B}_i^T = (1, A_i, M_i, \boldsymbol{X}_i^T)$ and $\boldsymbol{C}_i^T = (1, A_i, \boldsymbol{X}_i^T)$:

$$\hat{\boldsymbol{\gamma}} = \left(\sum_{i=1}^{n} \boldsymbol{C}_i \boldsymbol{C}_i^T \right)^{-1} \sum_{i=1}^{n} \boldsymbol{C}_i M_i,$$
$$\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{n} \boldsymbol{B}_i \boldsymbol{B}_i^T \right)^{-1} \sum_{i=1}^{n} \boldsymbol{B}_i Y_i,$$
$$\hat{\sigma}_M^2 = \frac{1}{n} \sum_{i=1}^{n} \left(M_i - \boldsymbol{C}_i^T \hat{\boldsymbol{\gamma}} \right)^2,$$
$$\hat{\sigma}_Y^2 = \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \boldsymbol{B}_i^T \hat{\boldsymbol{\beta}} \right)^2.$$

These estimates are derived by maximizing the log of the likelihood from (1.3). We estimate the covariance as -1 times the inverse of the observed information. The observed information is the second derivative of the log likelihood given in (1.3).

1.3.2 Incorporating individuals with missing data

We now have n_c observations with $R_i = 1$ and n_m with R_i=0. We utilize the EM algorithm to maximize the likelihood in (1.7). Briefly, the EM algorithm takes the score

equations derived from the likelihood assuming all the data is observed and then takes the expectation of the missing sufficient statistics conditional on the observed data (E-step) (Dempster et al. (1977)). These conditional sufficient statistics are then used to solve the score equations for $\boldsymbol{\Theta}$ (M-step). We then iterate between the E and M steps until convergence. In our application, for observations with $R_j = 0$, in the k^{th} E-step we calculate: $\hat{M}_j^{(k)} = E\{M_j|Y_j, A_j, \boldsymbol{X}_j, \hat{\boldsymbol{\Theta}}^{(k-1)}\}$, $\hat{M}_j^{2(k)} = E\{M_j^2|Y_j, A_j, \boldsymbol{X}_j, \hat{\boldsymbol{\Theta}}^{(k-1)}\}$, $\hat{\epsilon}_{Y,j}^{2(k)} = E\{(Y_j - \boldsymbol{B}_j\boldsymbol{\beta})^2|Y_j, A_j, \boldsymbol{X}_j, \hat{\boldsymbol{\Theta}}^{(k-1)}\}$, and $\hat{\epsilon}_{M,j}^{2(k)} = E\{(M_j - \boldsymbol{C}_j\boldsymbol{\gamma})^2|Y_j, A_j, \boldsymbol{X}_j, \hat{\boldsymbol{\Theta}}^{(k-1)}\}$, for observations $j = 1 \dots n_m$. We calculate these using the conditional normal:

$$
\begin{aligned}
M_j|Y_j, A_j, \boldsymbol{X}_j, \boldsymbol{\Theta} &\sim N(\mu_{M|Y}, \sigma^2_{M|Y}), \\
\mu_{M|Y} &= \frac{\sigma_Y^2(\gamma_0 + \gamma_A A_j + \boldsymbol{X}_j^T\boldsymbol{\gamma}_C) + \beta_M\sigma_M^2(Y_j - \beta_0 - \beta_A A_j - \boldsymbol{X}_j^T\boldsymbol{\beta}_C)}{\beta_M^2\sigma_M^2 + \sigma_Y^2}, \\
\sigma^2_{M|Y} &= \frac{\sigma_Y^2}{\beta_M^2\sigma_M^2 + \sigma_Y^2}.
\end{aligned}
$$

In the k^{th} M-step we estimate $\hat{\boldsymbol{\Theta}}^{(k)}$:

$$
\begin{aligned}
\hat{\boldsymbol{\gamma}}^{(k)} &= \left\{\sum_{i=1}^{n} \boldsymbol{C}_i^T\boldsymbol{C}_i\right\}^{-1}\left\{\sum_{i=1}^{n_c} \boldsymbol{C}_i^T M_i + \sum_{j=1}^{n_m} \boldsymbol{C}_j^T\hat{M}_j^{(k)}\right\}, \\
\hat{\boldsymbol{\beta}}^{(k)} &= \left[\sum_{i=1}^{n_c} \boldsymbol{B}_i^T\boldsymbol{B}_i + \sum_{j=1}^{n_m} E\{\boldsymbol{B}_j^T\boldsymbol{B}_j|Y_j, A_j\boldsymbol{X}_j, \hat{\boldsymbol{\Theta}}^{(k-1)}\}\right]^{-1} \\
&\qquad \left[\sum_{i=1}^{n_c} \boldsymbol{B}_i^T Y_i + \sum_{j=1}^{n_m} E\{\boldsymbol{B}_j^T Y_j|Y_j, A_j\boldsymbol{X}_j, \hat{\boldsymbol{\Theta}}^{(k-1)}\}\right], \\
\hat{\sigma}_M^{2(k)} &= \frac{1}{n}\left[\sum_{i=1}^{n_c}\left\{M_i - \boldsymbol{C}_i\hat{\boldsymbol{\gamma}}^{(k-1)}\right\}^2 + \sum_{j=1}^{n_m}\hat{\epsilon}_{M,j}^{2(k)}\right], \\
\hat{\sigma}_Y^{2(k)} &= \frac{1}{n}\left[\sum_{i=1}^{n_c}\left\{Y_i - \boldsymbol{B}_i\hat{\boldsymbol{\beta}}^{(k-1)}\right\}^2 + \sum_{j=1}^{n_m}\hat{\epsilon}_{Y,j}^{2(k)}\right].
\end{aligned}
$$

$\hat{M}_j^{2(k)}$ is used in the expectation of $E\{\boldsymbol{B}_j^T\boldsymbol{B}_j|\boldsymbol{Y}_j, A_j\boldsymbol{X}_j, \hat{\boldsymbol{\Theta}}^{(k-1)}\}$. We set our convergence tolerance as a Euclidean distance of 10^{-8}. The first E-step is informed by a base $\hat{\boldsymbol{\Theta}}^{(0)}$ from restricting to observations with $R_i = 1$. To estimate the covariance of $\hat{\boldsymbol{\Theta}}$, we take the negative inverse of the Expected Fisher Information of the EM algorithm (Louis (1982)):

$$
\boldsymbol{I}_{\Theta,m} = E_Y\left\{\frac{\partial^2\ell(\boldsymbol{Y}|\boldsymbol{A}, \boldsymbol{X}, \hat{\boldsymbol{\Theta}})}{\partial\boldsymbol{\Theta}\partial\boldsymbol{\Theta}^T}|\boldsymbol{A}, \boldsymbol{X}, \hat{\boldsymbol{\Theta}}\right\} -
$$

$$E_Y\left\{\frac{\partial \ell(\boldsymbol{Y}|\boldsymbol{A},\boldsymbol{X},\hat{\boldsymbol{\Theta}})}{\partial \boldsymbol{\Theta}}\frac{\partial \ell(\boldsymbol{Y}|\boldsymbol{A},\boldsymbol{X},\hat{\boldsymbol{\Theta}})}{\partial \boldsymbol{\Theta}^T}|\boldsymbol{A},\boldsymbol{X},\hat{\boldsymbol{\Theta}}\right\} \tag{1.8}$$

We estimate the causal effects by substituting in $\hat{\boldsymbol{\gamma}}$, $\hat{\boldsymbol{\beta}}$ of the final M step and construct confidence intervals or perform inference using the inverse of the above inversion. In practice, we use the observed information. The above Fisher information is equal to the expectation of the second order derivatives of (1.7).

1.4 Theory and Testing

1.4.1 Independence of β_M and γ_A

Using (1.8), we can prove that the cells in the inverse of $\boldsymbol{I}_{\Theta,m}$ corresponding to $\partial\boldsymbol{\beta}$ and $\partial\boldsymbol{\gamma}$ are 0. The proof is as follows, (assuming correctly specified mean model). We first expand the cells of the matrix corresponding to β and γ:

$$\begin{aligned}E_Y\left\{\frac{\partial^2 \ell(Y|A,\boldsymbol{X},\boldsymbol{\Theta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\gamma}^T}|A,\boldsymbol{X},\boldsymbol{\Theta}\right\} &= E_Y\left[E_M\left\{\frac{\partial^2 \ell(Y,M|A,\boldsymbol{X},\boldsymbol{\Theta})}{\partial\boldsymbol{\beta}\boldsymbol{\gamma}^T}|Y,A,\boldsymbol{X},\boldsymbol{\Theta}\right\}|A,\boldsymbol{X},\boldsymbol{\Theta}\right]\\ -E_Y\left[E_M\left\{\frac{\partial \ell(Y|M,A,\boldsymbol{X},\boldsymbol{\Theta})}{\partial\boldsymbol{\beta}}\frac{\partial \ell(M|A,\boldsymbol{X},\boldsymbol{\Theta})}{\partial\boldsymbol{\gamma}^T}|Y,A,\boldsymbol{X},\boldsymbol{\Theta}\right\}|A,\boldsymbol{X},\boldsymbol{\Theta}\right].\end{aligned}$$

The first term is 0 as it is based on the information when all data was observed. The second term involves the score equations.

$$\begin{aligned}\boldsymbol{S}_\beta &= \frac{\partial \ell(Y|M,A,\boldsymbol{X})}{\partial\boldsymbol{\beta}},\\ \boldsymbol{S}_\gamma &= \frac{\partial \ell(M|A,\boldsymbol{X})}{\partial\boldsymbol{\gamma}},\\ E_Y\left\{\frac{\partial^2 \ell(Y|A,X)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\gamma}^T}\right\} &= E_Y\left\{E_M\left(\boldsymbol{S}_\beta\boldsymbol{S}_\gamma^T|Y,A,\boldsymbol{X}\right)\right\}\\ &= E_M\left\{E_Y\left(S_\beta|M,A,\boldsymbol{X}\right)S_\gamma^T|A,\boldsymbol{X}\right\} = \boldsymbol{0},\end{aligned}$$

which is the desired result. A similar approach can then be used to show that the cells of $\boldsymbol{I}_{\Theta,m}$ corresponding to $\partial\sigma_Y^2\partial\boldsymbol{\gamma}$ and $\partial\sigma_M^2\partial\boldsymbol{\gamma}$ are 0. Using this and blockwise inversion gives that the block corresponding to the covariance between $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\beta}}$ in the inverse of $\boldsymbol{I}_{\Theta,m}$ is 0. This result will be utilized in the testing of the NIE. A more detailed proof showing the independence between $\hat{\gamma}_A$ and $\hat{\beta}_M$ only is provided in supplement (Appendix 4.1.3).

1.4.2 Testing

Let σ^2_{NDE} represent our estimate of the variance of $\hat{\beta}_A$. Our test statistic of $H_0 : \beta_A = 0$ is $Q_{NDE} = \hat{\beta}_A/\sigma_{NDE}$. In practice, we use the inverse of the observed information as opposed to the Fisher information. To test for the NDE, we compare to the quantiles of the standard normal distribution.

Next for the NIE, the null hypothesis is $H_0 : \gamma_A\beta_M = 0$, which occurs when $\gamma_A = 0$ or $\beta_M = 0$. This is a composite null that occurs under three cases. Case 1) $\gamma_a = 0, \beta_M \neq 0$; Case 2) $\gamma_A \neq 0, \beta_M = 0$, and Case 3) $\gamma_A = \beta_M = 0$. We use the joint significance method to test for the NIE, where we reject the null hypothesis when we reject the null for both β_M and γ_A. As $\hat{\beta}_M \perp \hat{\gamma}_A$, we can take the maximum of the p-values for $H_0 : \beta_M = 0$ and $H_0 : \gamma_A = 0$. Define the association specific p-values as $p_1 = 2P(Z \geq |\hat{\gamma}_A/\sigma_A|)$ and, $p_2 = 2P(Z \geq |\hat{\beta}_M/\sigma_M|)$, where $Z \sim N(0, 1)$ and σ_A and σ_M are the estimated standard errors of $\hat{\gamma}_A$ and $\hat{\beta}_M$ respectively. Our p-value for the NIE is then: $p_{NIE} = \max(p_1, p_2)$ (MacKinnon et al., 2002).

In a genome-wide test we expect that a majority of our mediators will be in Case 3. Each p-value follows a uniform under this scenario ($\beta_M = \gamma_A = 0$) making the type I error α^2 as p_{NIE} is now a Beta(2,1). Under Case 1 or 2, p_{NIE} converges to a Beta(1,1) (a Uniform(0,1)). This has been reported previously in the literature (MacKinnon et al. (2002)). We therefore use the approach from Liu and Lin to derive a new NIE p-value (Liu and Lin, 2017). Briefly, we estimate the proportion of Case 1, Case 2, and Case 3 in the genome wide sample and then construct a weighted average of p_1, p_2, and p^2_{NIE} by the estimated probability of being in Case 1, 2 or 3 given we are in the null. Each of those three p-values (p_1, p_2, and p^2_{NIE}) will have the correct type I error under their respective null case. This weighted p-value is denoted as $p_{average}$. If it still does not follow a uniform, an empirical correction can be applied where a reference p_{u1} and p_{u2} are calculated from uniforms and the new MCMC $p_{average}$ becomes the mean of the weighted reference p-values (p_{u1} and p_{u2}) less than the observed $p_{average}$.

Finally, for the natural total effect our test statistic is $Q_{NTE} = (\hat{\beta}_A + \hat{\gamma}_A\hat{\beta}_M)/\sigma_{NTE}$. σ_{NTE} is calculated using the multivariate delta method on the covariance of $\hat{\beta}_A$, $\hat{\gamma}_A$ and

$\hat{\beta}_M$. While the above theoretical results shows that the expected covariance between $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ is 0, in practice it may not be exactly zero, so we use the negative inverse of the observed information.

1.5 Simulations

1.5.1 Bias and Variance

We performed simulation studies to examine the finite sample performance of our method in the presence of varying levels of missingness. We set $\boldsymbol{X}_i = (1, X_{i,1}, X_{i,2})^T$, with $X_{i,1}$ generated from a standard normal and $X_{i,2}$ from a Bernoulli with probability 0.4. We did not generate A_i, but instead treated either $X_{i,1}$ or $X_{i,2}$ as the exposure to assess our performance for both continuous and binary exposures. For each simulation we set n=1000. We generate the variables in the following order (for $i = 1 \dots 1000$):

$$M_i = \gamma_0 + \gamma_1 X_{i,1} + \gamma_2 X_{i,2} + \epsilon_{M,i}; \epsilon_{M,i} \sim N(0,1), \tag{1.9}$$

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \beta_3 M_i + \epsilon_{Y,i}; \epsilon_{Y,i} \sim N(0,1). \tag{1.10}$$

with $\epsilon_{M,i} \perp \epsilon_{Y,i}$. The missing data mechanism was generated by the selection model:

$$logit(P(R_i = 1)) = \lambda_0 + \lambda_1 X_{i,2}. \tag{1.11}$$

We varied the λ parameters for different missing scenarios. When MCAR, we fix $\lambda_1 = 0$ and vary λ_0 such that the mean missingness is 20, 50, or 70%. For MAR, we set $\lambda_1 = -\log(1.5)$ so that observations with $X_2 = 1$ have odds of missing data 1.5 times those of observations with $X_2 = 0$. λ_0 was set such that the overall level of missingness in the sample was again 20, 50, or 70%. We did not simulate when the missing data mechanism is non-ignorable, as our approach is not valid under that scenario.

We fixed $\gamma_0 = \beta_0 = 0.2$ and $\gamma_1 = \gamma_2 = \beta_1 = \beta_2 = \beta_3 = 0.14$. We compared the bias and variance of the estimates to when we restrict the analysis to individuals with $R_i = 1$ (complete data-CD). We have six different scenarios to examine: MCAR and MAR each with 20, 50 or 70% missing overall. We perform 10K replications of each parameter combination. We denote when we incorporate all available data as the AD when abbreviation is necessary.

Our results for MAR are displayed in Table 1.1. Both approaches provide unbiased estimates of the parameters of interest. For β_1 and β_2, using all available data has less variable estimates compared to using just complete data. Unsurprisingly, the variance of the parameters when we use just the complete data observations increases with increasing levels of missingess (Table 1.1). For $\boldsymbol{\gamma}$, β_3 and σ^2_M, there is little difference in the empirical variance between using all available or just complete data. We also have the desired coverage for all mean parameters at all levels of missingness (Table 1.1). There are similar results when the data is MCAR (Table 4.1). If we compare the efficiency of the estimates (Figure 1.1 top panel), we see that the parameters for β_1 and β_2 when we use all available data are much more efficient compared to when use just complete data while there is only marginal gains for γ_1, γ_2 and β_3.

For the causal effects (Table 1.2), we have unbiased estimates of the NIE and NTE. Using all available data led to estimates of the NTE with smaller variance compared to just using complete data. There is no difference between the empirical variances of the NIE, and using all available data or just complete data leads to increasing variance with increasing levels of missingness. These results hold whether the exposure is continuous or categorical. The results for the NDE (β_1 and β_2) are in Table 1.1. Examining the efficiency, (Figure 1.1 row 2), we see that the causal estimates for the natural total effects using the all available data are much more efficient, while there is only slight gains for the natural indirect effect. Efficiency results when MCAR are provided in Supplementary Figure 4.1.

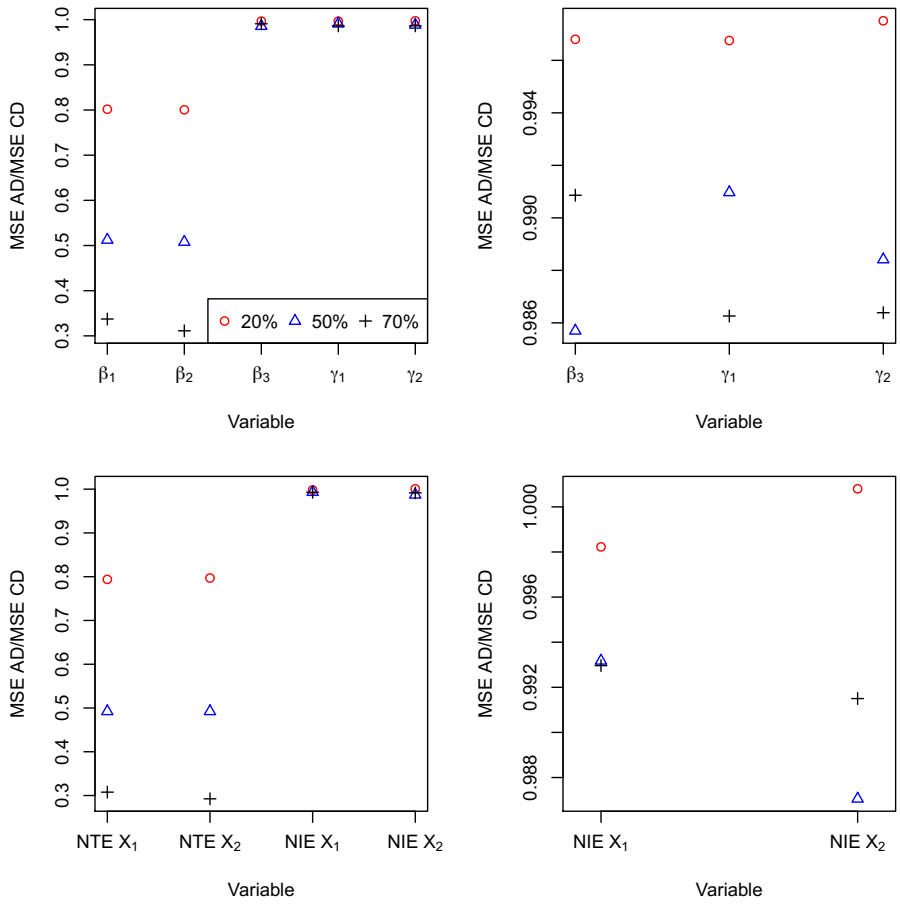

Figure 1.1: Efficiency when MAR over 10^4 simulations comparing when use all available data vs just complete data.Top panel shows the efficiency for the parameters $\beta_1, \beta_2, \beta_3, \gamma_1$ and γ_2. Bottom panel for NTE and NIE. Right column of plots shows efficiency for β_3,γ_1,γ_2, and NIE's. X_1 is continuous and X_2 is categorical.

Table 1.1: Mean estimate, empirical variance, and 95% coverage when MAR for 10^4 simulations.

		All Available Data			Just Complete Data		
Parameter	% Missing	Mean	Var	95% CI	Mean	Var	95% CI
$\underline{\beta_0}$	20%	0.200	0.002	0.952	0.200	0.002	0.950
	50%	0.200	0.002	0.949	0.199	0.003	0.948
	70%	0.200	0.002	0.953	0.202	0.006	0.948
$\underline{\beta_1}$	20%	0.140	0.001	0.950	0.140	0.001	0.950
	50%	0.140	0.001	0.952	0.140	0.002	0.950
	70%	0.139	0.001	0.948	0.140	0.003	0.953
$\underline{\beta_2}$	20%	0.140	0.004	0.951	0.140	0.005	0.949
	50%	0.140	0.004	0.949	0.140	0.009	0.950
	70%	0.139	0.004	0.953	0.137	0.014	0.948
$\underline{\beta_3}$	20%	0.140	0.001	0.951	0.140	0.001	0.951
	50%	0.141	0.002	0.947	0.141	0.002	0.949
	70%	0.140	0.003	0.942	0.139	0.003	0.947
$\underline{\gamma_0}$	20%	0.200	0.002	0.951	0.200	0.002	0.952
	50%	0.200	0.003	0.950	0.200	0.003	0.952
	70%	0.199	0.006	0.946	0.199	0.006	0.950
$\underline{\gamma_1}$	20%	0.140	0.001	0.946	0.140	0.001	0.947
	50%	0.141	0.002	0.947	0.141	0.002	0.949
	70%	0.141	0.003	0.943	0.141	0.003	0.946
$\underline{\gamma_2}$	20%	0.140	0.005	0.948	0.140	0.005	0.950
	50%	0.140	0.008	0.951	0.140	0.008	0.951
	70%	0.142	0.014	0.947	0.142	0.014	0.947
$\underline{\sigma_Y^2}$	20%	0.996	0.002	—	1.000	0.002	—
	50%	0.995	0.002	—	1.001	0.004	—
	70%	0.993	0.002	—	1.000	0.007	—
$\underline{\sigma_M^2}$	20%	0.996	0.002	—	0.999	0.002	—
	50%	0.995	0.004	—	1.001	0.004	—
	70%	0.990	0.007	—	1.000	0.007	—

Table 1.2: Mean estimate, empirical variance, and 95% coverage when MAR for 10^4 simulations examining causal effects.

		All Available Data			Just Complete Data		
Parameter	% Missing	Mean	Var	95 CI	Mean	Var	95 CI
NIE Continuous	20%	0.0196	5.20e-05	0.929	0.0196	5.21e-05	0.930
	50%	0.0198	8.39e-05	0.925	0.0198	8.44e-05	0.928
	70%	0.0198	1.48e-04	0.913	0.0196	1.49e-04	0.911
NIE Categorical	20%	0.0196	1.33e-04	0.938	0.0196	1.33e-04	0.937
	50%	0.0198	2.18e-04	0.935	0.0197	2.21e-04	0.933
	70%	0.0198	3.78e-04	0.924	0.0197	3.80e-04	0.925
NTE Continuous	20%	0.1600	1.03e-03	0.950	0.1600	1.29e-03	0.950
	50%	0.1598	1.01e-03	0.952	0.1594	2.05e-03	0.951
	70%	0.1592	1.05e-03	0.946	0.1592	3.40e-03	0.954
NTE Categorical	20%	0.1596	4.27e-03	0.950	0.1595	5.36e-03	0.952
	50%	0.1593	4.25e-03	0.950	0.1596	8.64e-03	0.951
	70%	0.1585	4.20e-03	0.954	0.1568	1.44e-02	0.950

1.5.2 Hypothesis Testing

We next assessed the type I error and power. The data were again generated from (1.9), (1.10), and (1.11), with $X_{i,1}$ and $X_{i,2}$ generated as before. We used the same level of missingness as in the point estimation section. We set the direct effect to null by fixing $\beta_1 = \beta_2 = 0$. We examined the performance under the null for the NIE, NDE, and NTE under three scenarios: 1) $\gamma_1 = \gamma_2 = 0$, $\beta_3 = 0.39$, 2) $\gamma_1 = \gamma_2 = 0.39$, $\beta_3 = 0$, and 3) $\gamma_1 = \gamma_2 = \beta_3 = 0$. These correspond to the three cases detailed in the testing section for NIE. We fixed $\gamma_0 = \beta_0 = 0.2$ for all scenarios. We generated 5x10^4 datasets and examined the type I error at α of 0.05 and 0.01.

The type I error results for MAR are provided in Tables 1.3 and 1.4. Briefly, for all levels of missingness, the type I error rate is controlled at the desired α level for 0.05 or 0.01. This holds for both approaches under all three null scenarios considered. The results are identical for both continuous and categorical and identical to when MCAR (Supplementary Table 4.3). The TIE for the NIE is close to α^2 for Case 3 as expected.

Table 1.3: Type I error evaluated at $\alpha = 0.05$ when MAR over 5x10^4 simulations with n=1000. Examined under different TIE scenarios. Case 1: $\gamma_1 = \gamma_2 = \beta_1 = \beta_2 = 0$, $\beta_3 = 0.39$, Case 2: $\gamma_1 = \gamma_2 = 0.39$, Case 3: $\gamma_1 = \gamma_2 = \beta_1 = \beta_2 = \beta_3 = 0$.

Exposure	TIE Case	% Missing	All Available Data			Just Complete Data		
			NDE	NIE	NTE	NDE	NIE	NTE
Continuous	Case 1	20%	0.050	0.050	0.051	0.051	0.050	0.051
		50%	0.050	0.050	0.050	0.048	0.050	0.049
		70%	0.050	0.050	0.050	0.051	0.053	0.049
	Case 2	20%	0.051	0.049	0.052	0.052	0.050	0.026
		50%	0.051	0.050	0.052	0.050	0.052	0.025
		70%	0.052	0.052	0.051	0.050	0.055	0.026
	Case 3	20%	0.049	0.003	0.048	0.049	0.003	0.048
		50%	0.052	0.003	0.052	0.051	0.003	0.052
		70%	0.050	0.003	0.053	0.050	0.003	0.051
Categorical	Case 1	20%	0.049	0.049	0.049	0.050	0.049	0.049
		50%	0.052	0.052	0.050	0.050	0.052	0.050
		70%	0.052	0.051	0.052	0.052	0.055	0.050
	Case 2	20%	0.053	0.049	0.051	0.053	0.050	0.043
		50%	0.052	0.050	0.052	0.050	0.052	0.043
		70%	0.051	0.047	0.051	0.051	0.050	0.041
	Case 3	20%	0.050	0.002	0.049	0.051	0.002	0.049
		50%	0.051	0.003	0.053	0.051	0.003	0.052
		70%	0.049	0.002	0.051	0.050	0.003	0.049

Table 1.4: Type I error evaluated at $\alpha = 0.01$ when MAR over 5x10^4 simulations with n=1000. Examined under different TIE scenarios. Case 1: $\gamma_1 = \gamma_2 = \beta_1 = \beta_2 = 0$, $\beta_3 = 0.39$, Case 2: $\gamma_1 = \gamma_2 = 0.39$, Case 3: $\gamma_1 = \gamma_2 = \beta_1 = \beta_2 = \beta_3 = 0$.

Exposure	TIE Case	% Missing	All Available Data			Just Complete Data		
			NDE	NIE	NTE	NDE	NIE	NTE
Continuous	Case 1	20%	0.010	0.010	0.011	0.010	0.010	0.010
		50%	0.010	0.010	0.010	0.010	0.011	0.010
		70%	0.010	0.011	0.011	0.010	0.011	0.010
	Case 2	20%	0.011	0.009	0.010	0.011	0.010	0.004
		50%	0.010	0.010	0.010	0.011	0.011	0.003
		70%	0.010	0.010	0.011	0.010	0.012	0.004
	Case 3	20%	0.010	0.000	0.010	0.010	0.000	0.010
		50%	0.010	0.000	0.010	0.010	0.000	0.010
		70%	0.010	0.000	0.011	0.010	0.000	0.011
Categorical	Case 1	20%	0.009	0.010	0.011	0.010	0.010	0.010
		50%	0.010	0.011	0.010	0.010	0.012	0.010
		70%	0.010	0.011	0.010	0.011	0.011	0.010
	Case 2	20%	0.011	0.009	0.011	0.011	0.010	0.008
		50%	0.010	0.010	0.011	0.010	0.011	0.008
		70%	0.010	0.008	0.011	0.010	0.009	0.008
	Case 3	20%	0.010	0.000	0.010	0.010	0.000	0.010
		50%	0.010	0.000	0.010	0.010	0.000	0.010
		70%	0.010	0.000	0.011	0.010	0.000	0.010

We next examined the type I error rate on the genome wide scale, where a majority of the mediators are in Case 3. We generated 4x10^5 potential mediators, where 0.00035 each were in Case 1 and Case 2, and the remaining were in Case 3. For the simulations in Case 3) $\gamma_1 = \gamma_2 = \beta_3 = 0$. For the .035% in Case 1, $\gamma_1 = \gamma_2 = 0$ and $\beta_3 = 0.39$, and reverse for the .035% in Case 2. We expect that p_{NIE} will be severely deflated, but that upon correction, we will have the desired size (Liu and Lin, 2017).

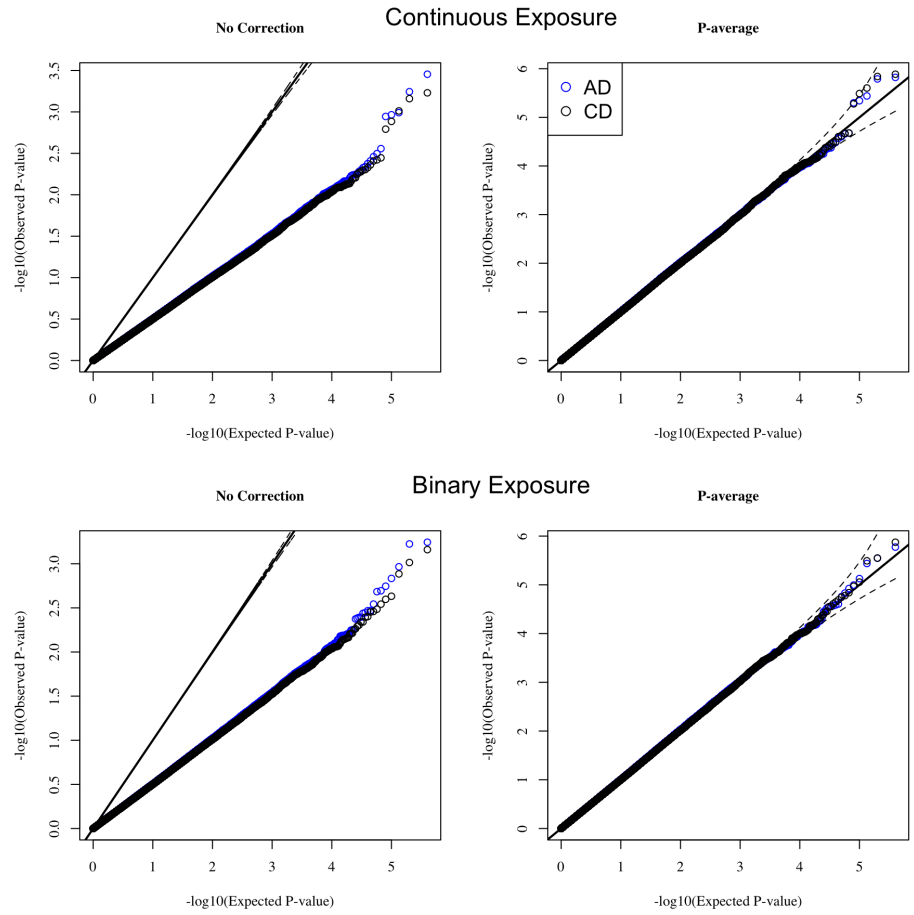

Figure 1.2: Results from genome wide type I error simulations when 50% missing and MAR. Done over 4x10^5 replications where .035% are in Case 1, .035% in Case 2 and the remaining in Case 3. Plots to the left shows p_{NIE}, plots to the right shows $p_{average}$. AD: All available data. CD: Just complete data.

For brevity, we only display the results when there was on average 50% missing (Figure 1.2). The other levels of missingness and MCAR results are provided in the supplement (Figures 4.2 to 4.6). As expected we see deflation of the joint p-values due to the majority of the simulations being in Case 3 (Figure 1.2 first column). Upon performing the p-average correction, the p-values for the NIE following a uniform as expected (Figure 1.2 second column) (Liu and Lin, 2017).

We next investigated the empirical power. For each level of missing we varied three sets of variables (β_1, β_2), (γ_1, γ_2), and β_3. We set $\beta_1, \beta_2 \in (0.05, 0.1)$, $\gamma_1, \gamma_2 \in (0.05, 0.1)$, and $\beta_3 \in (0.05, 0.1)$ and examined the combination of these parameter values (8 different sets). We performed 10^4 simulations of each set and evaluated the power at $\alpha = 0.05$.

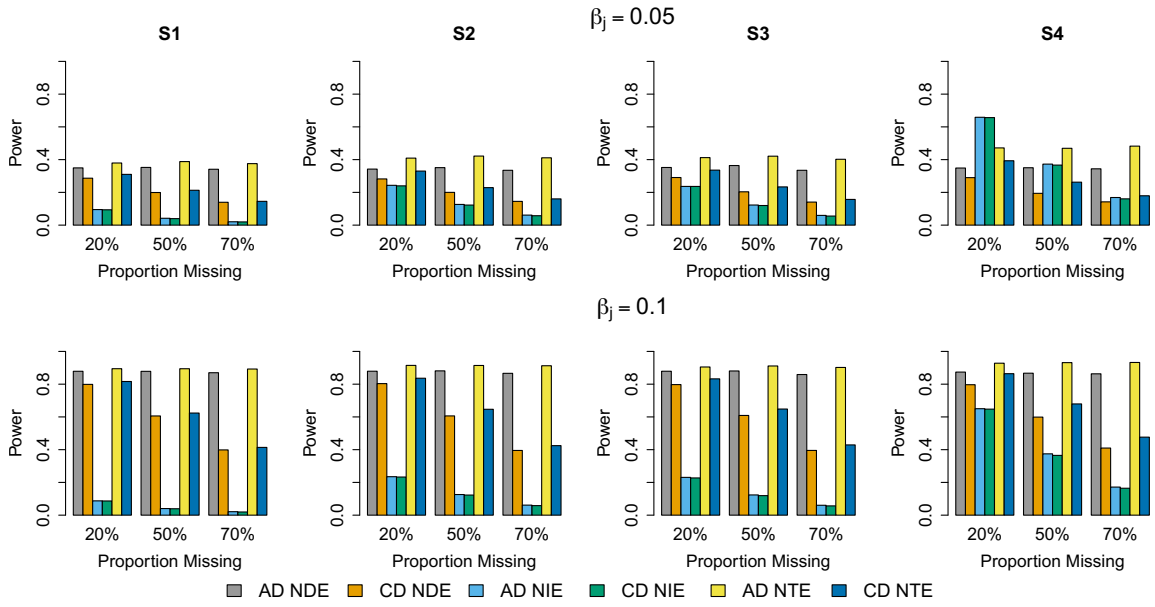

Figure 1.3: Power for continuous variable when MAR, evaluated at $\alpha = 0.05$. Top panel shows when $\beta_1 = 0.05$, bottom panel when $\beta_1 = 0.1$. Moving from left to right: S1) $\gamma_1 = \beta_3 = 0.05$, S2) $\gamma_1 = 0.1, \beta_3 = 0.05$, S3) $\gamma_1 = 0.05, \beta_3 = 0.1$ and S4) $\gamma_1 = \beta_3 = 0.1$. AD: All available data. CD: Just complete data.

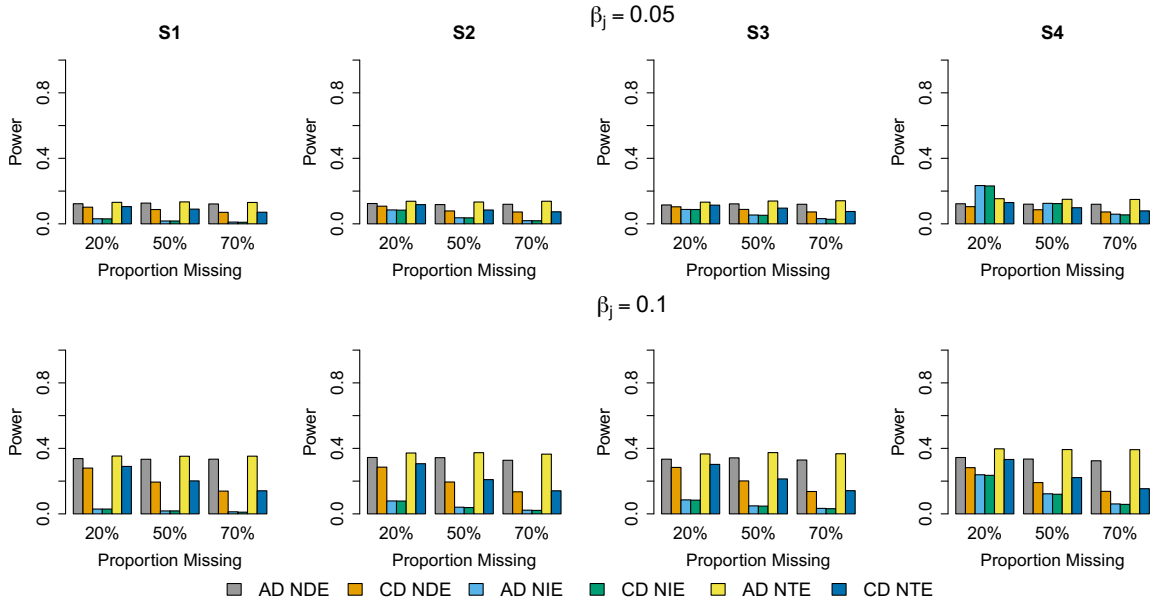

Figure 1.4: Power for categorical variable when MAR, evaluated at $\alpha = 0.05$. Top panel shows when $\beta_2 = 0.05$, bottom panel when $\beta_2 = 0.1$. Moving from left to right: S1) $\gamma_1 = \beta_3 = 0.05$, S2) $\gamma_1 = 0.1, \beta_3 = 0.05$, S3) $\gamma_1 = 0.05, \beta_3 = 0.1$ and S4) $\gamma_1 = \beta_3 = 0.1$. AD: All available data. CD: Just complete data.

Figure 1.3 displays the power when MAR for a continuous variable. When there is a low level of missingness, using all available data has slightly higher power than using just the complete data for the NDE and NTE. With higher level of missingness, we see power using just complete data decreases for all three causal effects. The power of our algorithm for testing NDE and NTE hardly changes. We have the most power for NTE, except when $\beta_1 = 0.05$ and $\gamma_1 = \beta_3 = 0.1$. For that scenario, the NIE has more power when 20% missing. For the NIE test, using all available data has only marginally more power than using just complete data regardless of the proportion missing. There are similar patterns for dichotomous variables (Figure 1.4). The categorical has less power for these parameter values. The result is the same when the missingess is MCAR (Supplementary Figures 4.7 and 4.8).

1.6 Application to Project Viva

We tested for the indirect effect of maternal prepregnancy BMI through DNAm on childhood BMI z-score at three different time points (6 months, 3 years, and 7 years). Observations were included if the outcome (child BMI z-score), exposure (maternal pre-pregnancy BMI), and covariates were observed. The covariates were mothers income, maternal race, first time parent, married or cohabiting, age at enrollment, father's BMI, and child's gender. Study characteristics are provided in Tables 1.5 and 1.6. More information on the study design can be found in the literature (Agha et al., 2016; Oken et al., 2015). BMI z-scores at 6 months were calculated using the WHO reference (WHO, 2010) and BMI z-scores at 3 years and 7 years were calculated using the CDC reference (Kuczmarski et al., 2000). We removed any observations where the gestational age at delivery was less than 34 weeks. The analysis was done in two groups: one restricted to parents and children who were white and another that included every parent-child pair.

Table 1.5: Study characteristics restricted to white parents-child pairs.

Parameter	6 months	3 years	7 years
N	746	798	655
DNAm observed	247 (33%)	289 (36%)	274 (42%)
Mean BMI Z-score	0.686	0.461	0.300
Mean Maternal BMI	24.126	24.023	24.068
Age at enrollment	33.116	33.393	33.431
Mean Paternal BMI	26.337	26.393	26.469
Married or Cohabitation	728 (0.98)	782 (0.98)	642 (0.98)
First Child	374 (0.5)	380 (0.48)	319 (0.49)
Attended College	588 (0.79)	637 (0.8)	528 (0.81)
Smoking Status			
Former Smoker	177 (0.237)	195 (0.244)	152 (0.232)
Smoked During pregnancy	71 (0.095)	73 (0.091)	52 (0.079)
Never Smoker	498 (0.668)	530 (0.664)	451 (0.689)

Table 1.6: Study characteristics of analysis using every mother-child pair. Maternal race other groups individuals who wrote Other, or identified as Asian, or more than one race.

Parameter	6 months	3 years	7 years
N	1090	1174	1027
DNAm observed	356 (33%)	430 (37%)	402 (39%)
Mean BMI Z-score	0.6771	0.461	0.39
Mean Maternal BMI	24.582	24.656	24.714
Age at enrollment	32.461	32.471	32.248
Mean Paternal BMI	26.405	26.443	26.474
Married or Cohabitation	1031 (0.95)	1098 (0.94)	941 (0.92)
First Child	534 (0.49)	555 (0.47)	488 (0.48)
Attended College	777 (0.71)	834 (0.71)	702 (0.68)
Smoking Status			
Former Smoker	222 (0.2)	237 (0.20)	195 (0.190)
Smoked During pregnancy	109 (0.1)	123 (0.10)	99 (0.096)
Never Smoker	759 (0.7)	814 (0.69)	733 (0.714)
Maternal race			
Black	130 (0.119)	139 (0.118)	158 (0.154)
Hispanic	64 (0.059)	74 (0.063)	67 (0.065)
Other	97 (0.089)	109 (0.093)	101 (0.098)
White	799 (0.733)	852 (0.726)	701 (0.683)

We analyzed DNAm from cord blood collected from participants who had given genetic consent at one recruitment center. Testing between observed covariates and having observed DNAm are provided in Supplementary Tables 4.5 and 4.6. DNAm was collected from white blood cells and arrayed using the Infinium Human Methylation 450 BeadChip array. DNAm was assessed on the logit scale to better approximate normality (Du et al. (2010)). Non CpG probes, probes near SNPs, cross-reactive probes (Chen et al., 2013) and probes on the sex chromosomes were removed prior to analysis leaving us with 372,563 probes. We used a q-value less than 0.10 as our significance threshold (Storey, 2002) .

We first accounted for technical covariates and cell type composition which were not observed in the individuals without observed DNAm. We performed PCA on the subset of individuals with DNAm and then regressed out the top ten PC's. Principal com-

ponents in DNAm are often correlated with technical artifacts and cell type composition (Rahmani et al., 2016; Barfield et al., 2014), and therefore the residuals should represent a new set of "DNAm" markers independent of cell type and technical variables. We performed this for each set of observations at 6 months, 3 years, and 7 years and each group (every mother-child pair or just white mother-child pairs.).

Examining the qq-plots of our NIE tests, we see the expected deflation as the majority of the probes are likely in Case 3 (Figure 1.5). Upon correcting the p-values via the $p_{average}$ and MCMC $p_{average}$, the p-values have the desired distribution (Figure 1.6). Individual level association p-values are provided in Supplementary Figures 4.9 and 4.10. There was no DNAm probe that passed the q-value cutoff of 0.1 for any time point of any group analyzed for a significant NIE.

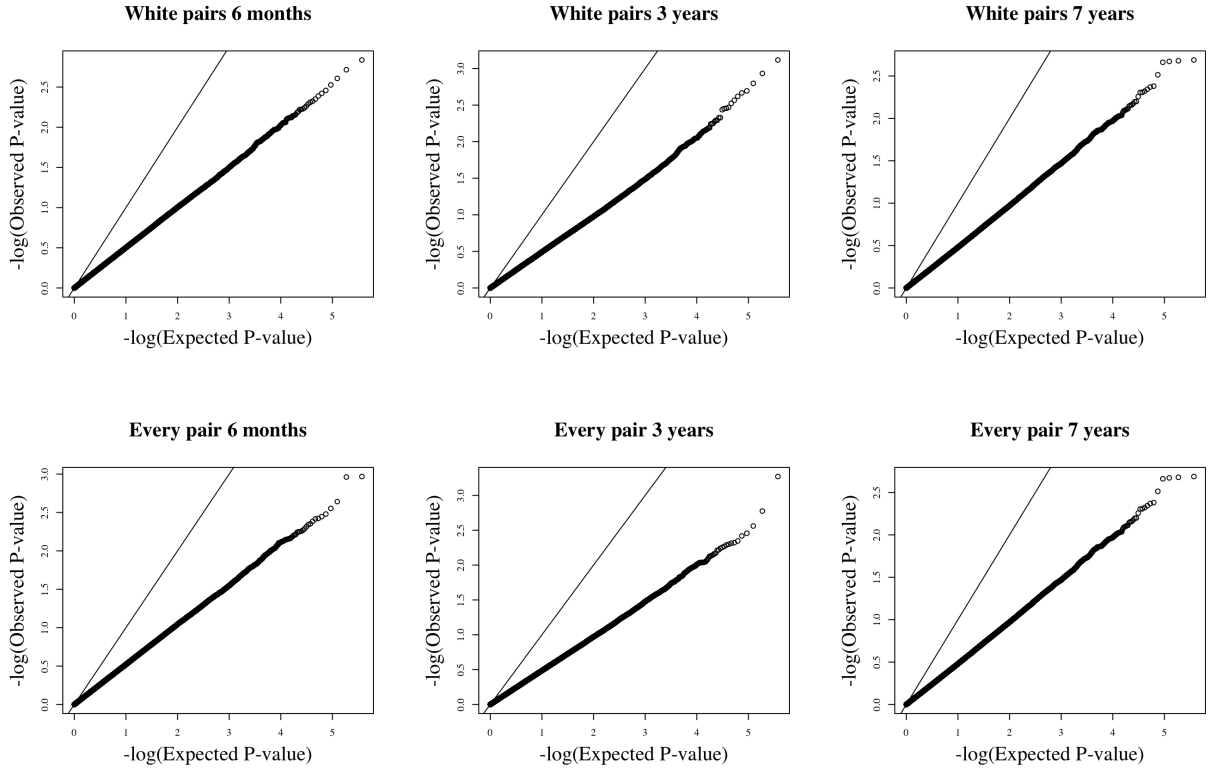

Figure 1.5: Results from analysis in Project Viva pre-adjusted NIE p-values using all available data. Top panel shows analysis in whites, bottom when include every mother-child pair. From left to right, analysis at 6 months, 3years, 7years.

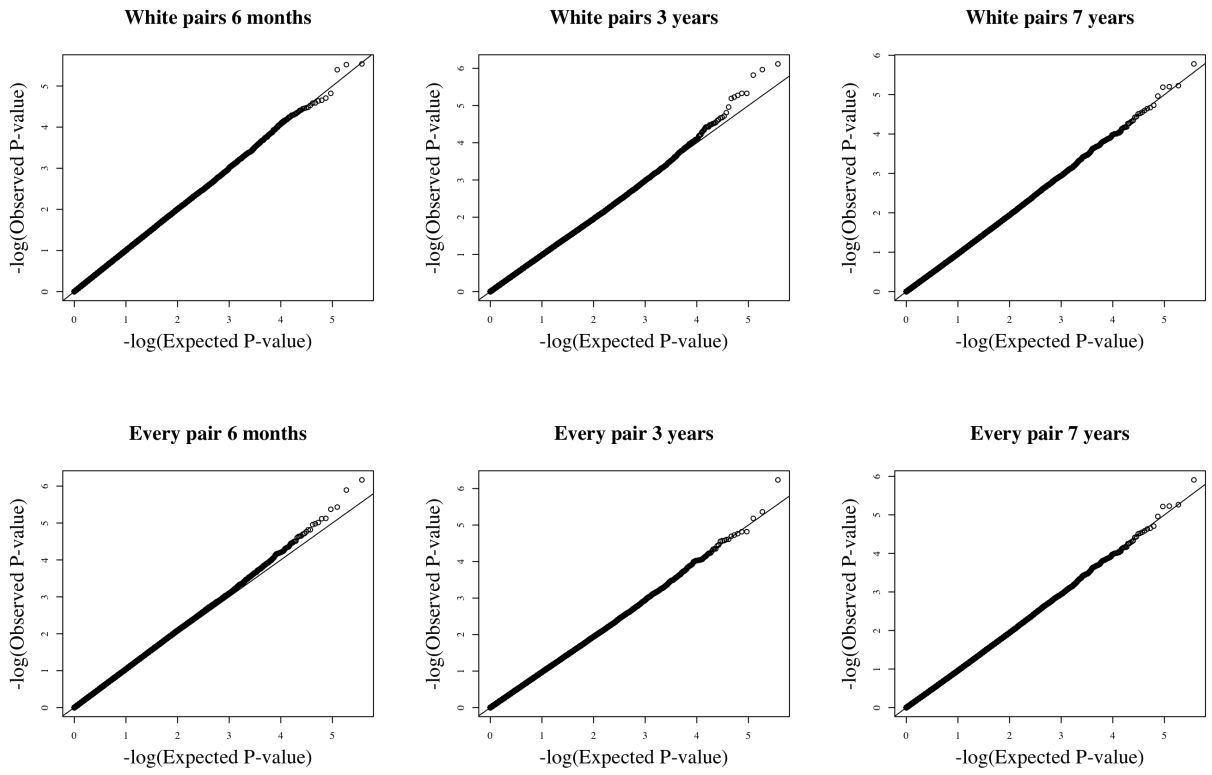

Figure 1.6: Results from analysis in Project Viva post-adjusted MCMC $p_{average}$ NIE p-values using all available data. Top panel shows analysis in whites, bottom when include every mother-child pair. From left to right, analysis at 6 months, 3years, 7years.

Just for completeness, we examined the top ten significant NIE probes for each time point and study group analyzed to see if there was any overlap with reported hits. One probe (cg19268652 on chromosome 6) overlapped with the HIVEP1 gene which had a nearby SNP found to be significant in a GWAS of BMI (rs2228213, Locke et al. (2015)).This SNP was 113K BP away from our probe cg19268652. This probe was in the top ten associations when analyzing BMI z-scores at 3 years in the analysis including every mother-child pair. The probe had a p-value of 4.51e-03 for the association with BMI z-scores and a p-value of 2.88e-03 with maternal pre-pregnancy BMI. This probe was not significant by the q-value threshold.

We therefore scaled back and looked at p-values for the association between maternal pre-pregnancy BMI and DNAm (γ_A) and DNAm to BMI z-score (β_M) to see if anything was significant for these associations. There were 16 unique probes that had an associa-

tion in these analyses (Tables 1.7 and 1.8). Seven probes were found in the analyses restricted to white mother-child pairs, with the remaining nine in the analysis that included every mother-child pair irrelevant of race. There was no overlap between these sets of probes (Tables 1.7 and 1.8). There was one probe found significant with pre-pregnancy BMI in white pairs when we restricted to the set of individuals with observed BMI z-scores at 6 months. This probe was only significant when included all available white mother-child pairs, and not when restrcitred to just complete data white mother-child pairs. There were six probes found significant at this time point when we analyzed every mother-child pair, three significant when restricting to just mother-child pairs with complete data. Of these six, four were associated with pre-pregnancy BMI, and two with child's BMI z-score. In the set of individuals with BMI at three years, there was one probe found significant in white pairs by analyzing all available data with child's BMI z-score. Finally at seven years, there were five probes significant in white pairs by incorporating all data, of which only one was significant if restricted to just complete data in white mother child pairs. All five probes had significant association with BMI z-scores. There was three probes found significant in the analysis of every mother-child pairs at this time point, one with pre-pregnancy BMI, and two with BMI z-scores. Two of these probes were called significant if used just pairs with complete data. We did not see any overlap with the probes reported in the Mendelson et al paper from earlier this year (Mendelson et al. (2017)). One possibility for the lack of overlap is the difference in study design (younger population) and the sample tissue considered (cord blood vs whole blood Leukocytes).

Interestingly, if we compare the p-values of these 16 probes between incorporating all available data and just using complete data, there is not much difference (Table 1.7 and 1.8). It appears that incorporating all available data leading to a slight decrease in the MSE (via simulations) can lead to more probes being called significant at the genome wide scale. Only seven probes in total were found significant if we restricted the analysis to just pairs with observed DNAm data. None of these probes had a significant mediation effect however

We focus on one probe with a significant β_M, cg26283921 on chromosome 21 near COL18A1 (Collagen XVIII α 1), which was significant when using all available data, but

Table 1.7: Probes significant when include every mother-child pair in Project Viva. "*" denotes a probe only significant by analyzing all available data. "**" significant via incoprorating all available data and just restricting to pairs with complete data. "A" denotes maternal prepregnancy BMI, "M" is the DNAm M-value at that probe, and "Y" is the child's BMI z-score at the time point. AD: All available data. CD: Just complete data.

Probe	Cytoband	Assoc.	Time	Gene	AD P-val	CD P-val
cg01903305*	5p15.33	M to Y	6 M	SLC12A7	1.64e-07	8.19e-07
cg20707409**	10q26.13	A to M	6 M	FAM24B-CUZD1	4.16e-07	6.57e-07
cg01056242**	12q24.33	A to M	6 M	LINC01257	6.32e-08	1.10e-07
cg26922917**	12q24.33	A to M	6 M	LINC01257	1.95e-07	3.24e-07
cg03284642*	13q12.3	A to M	6 M	SLC7A1	1.11e-06	1.73e-06
cg13543355*	14q24.3	M to Y	6 M	LRRC74A	5.06e-07	2.10e-06
cg26240433*	14q32.12	M to Y	7 Y	IFI27L1	4.92e-07	1.67e-06
cg07524348**	16q24.1	A to M	7 Y	MIR5093	1.57e-07	2.58e-07
cg00416475**	19p13.11	M to Y	7 Y	LPAR2/GMIP	1.74e-07	6.87e-07

not if restricted to just observations with complete data, in white mother-child pairs. It was significantly associated with BMI z-score at 7 years of age. The DNAm M-values at this probe were associated with increasing BMI z-score. A study published in 2014 found a low frequency coding variant near this gene that was significantly associated with fasting blood triglycerides (TG) in African Americans and was associated with a decrease in HDL-C levels (Peloso et al., 2014). This reported variant was approximately 50KB away from cg26283921. In addition, a specific isoform of COL18A1 was also found in a mouse model to be associated with a decrease in adiposity and hypertriglyceridemia levels (Aikio et al., 2014). In the analysis including every mother-child pair this probe did not reach the genome wide threshold but was still fairly significant (using all available data p-value of 4.16E-05, just complete data of 8.14E-05). This probe did not have a significant mediation effect.

1.7 Discussion

In this paper we detailed an EM algorithm to incorporate individuals with missing mediator observations into a mediation analysis. This approach can provide more efficient

Table 1.8: Probes significant when restrict to white mother-child pair in Project Viva. "*" denotes a probe only significant by analyzing all available data. "**" significant via incoprorating all available data and just restricting to pairs with complete data. "A" denotes maternal prepregnancy BMI, "M" is the DNAm M-value at that probe, and "Y" is the child's BMI z-score at the time point. AD: All available data. CD: Just complete data.

Probe	Cytoband	Assoc.	Time	Gene	AD P-val	CD P-val
cg27541317*	1q21.2	A to M	6 M	BOLA1	1.80e-07	3.15e-07
cg07574267**	1p36.32	M to Y	7 Y	LOC100129534/SKI	3.14e-09	5.08e-08
cg07896438**	2q34	M to Y	7 Y	IKZF2	3.19e-07	1.90e-06
cg27359472*	6p22.1	M to Y	7 Y	OR12D3	5.94e-07	3.14e-06
cg07256732*	16p13.3	M to Y	3 Y	PIGQ	1.14e-07	5.47e-07
cg11558551*	19p13.11	M to Y	7 Y	BST2	3.97e-07	2.27e-06
cg26283921*	21q22.3	M to Y	7 Y	COL18A1	9.67e-07	4.65e-06

estimators and increased power to detect the NDE and NTE compared to restricting the analysis to just individuals with observed data. Under the necessary causal assumptions and the additional assumption that the missing data mechanism is ignorable, we can place a causal interpretation on the estimates.

While in our simulations we did not see a large increase in efficiency for the parameters associated with the mediator, in real data applications we did see some more significant results for *parameter specific* association. This is looking at the associations from A to M or M to Y. If we had just used complete data, only seven probes had significant A to M or M to Y associations. Incorporating all available data led to an additional nine probes having either a significant A to M or M to Y association. We just did not see any intersection of those hits, something that was significant A to M and M to Y. Not detecting any NIE could simply be due to there not being a shared biological mechanism through DNAm in cord blood and child's BMI. We did not test for the natural direct effect, as it was likely to be the same for every probe. Our method however would be beneficial for detecting a NDE with an exposure that varied with probe such as nearby SNPs.

One of the major assumptions we make is that the missing data mechanism is ignorable. In genomic studies this is a reasonable assumption. Individuals do not have observed genomic data for two major reasons: either not signing genomic consent or a

limitation or funds on the part of the researcher. If the decision on who to collect genomics is done randomly due to budget, then the assumption is not violated. If the selection is based on certain covariates (say extreme phenotypes) than the missingness is at random. Either way the missingess is ignorable. The other major reason is not signing genomic consent. This involves individuals personal beliefs and can be influenced by a variety of potential confounders. It however seems unlikely that DNAm in cord blood is associated with personal beliefs. For it to be missing not at random, this would imply that DNAm causes personal beliefs that then cause a decision to not give genomic consent. A final potential reason not mentioned above is technical errors in the lab, which is of no fault of the unobserved genomic data.

In the Project Viva analysis, we could also analyze maternal BMI as a categorical variable with levels for the mother being underweight, normal, overweight, or obese based on the WHO criteria. A recent study found that analyzing maternal pre-pregnancy in these four categories led to more significant results compared to when maternal BMI was considered as a continuous variable (Sharp et al., 2015).

Further work could be done to potentially incorporate multiple mediators and perform a gene set analysis. A problem with detecting probe specific natural indirect effects on the genome wide scale is that each association (exposure to DNAm and DNAm to outcome) needs to pass the stringent genome wide threshold. Testing gene sets could reduce this multiple testing burden. Other future work could be done to expand the mediation algorithm for different classes of outcomes and mediators.

In conclusion, we have provided an EM algorithm that utilizes all observations when there is missingness on the mediator. We provided a proof of independence between the β and γ terms allowing us to use the traditional joint significant p-value. This EM algorithm is ideal for genomic studies as it is computationally efficient. While in simulations the modest gain in efficiency seen for the parameters relating to the missing mediator were small, this still led to more probes having significant parameter specific associations than if we had restricted the analysis to just observations with complete data.

Estimating Cell Type Specific Associations from Whole Blood Methylation

Richard Thomas Barfield

Department of Biostatistics

Harvard Graduate School of Arts and Sciences

Andrea Baccarelli

Departments of Environmental Health Sciences and Epidemiology

Columbia University Mailman School of Public Health

Xihong Lin

Department of Biostatistics

Harvard Chan School of Public Health

Department of Statistics

Harvard University

2.1 Introduction

DNA methylation (DNAm) is an epigenetic mechanism that occurs when a methyl group is attached to a cytosine base pair followed by a guanine (CpG Site) on the genome. It is a biological tool associated with differentiating cells and tissue types leading to different DNAm patterns by cell (Smith and Meissner, 2013; Meissner et al., 2008). Studies that examine how DNAm changes by cell are often conducted on small sample groups (Horvath, 2013; Lokk et al., 2014; Ma et al., 2014; Schultz et al., 2015). Most DNAm studies use peripheral blood, which is a collection of different cells, with cell type composition varying between study participants and estimate exposures effects on the aggregate DNAm. This type of data is easy to collect and is primarily composed of white blood cells. Currently, there is no approach to examine cell specific DNAm profiles for multiple phenotypes without fractionation and sending multiple samples from the same participants to be arrayed. In this paper, we develop a statistical method to study the effects of exposures on cell-type specific methylation using DNAm measured from peripheral blood, thus avoiding additional costs.

Epigenome Wide Association Studies (EWAS) measure DNAm at hundreds of thousands of probes at CpG sites across the genome. As mentioned above, studies of an exposures effect on DNAm are usually run on whole blood and measure aggregate methylation, a weighted sum of methylations at different white blood cells. Traditional EWAS analyses regresses a probe-specific DNAm proportion, defined as the number of methylated signals over the total signal (Beta-value) on an exposure across the genome. There are also more complicated methods, such as a beta-regression, bump hunting or probe-set based analysis (Peters et al., 2015; Jaffe et al., 2012; Siegmund, 2011).

No matter what analysis is run, cell type composition is a major source of confounding that needs to be adjusted for in the analysis (Jaffe and Irizarry, 2014). If cell type composition is not collected in a study, it can be estimated with a reference panel (Houseman et al., 2012) and included as additional covariates. Several other methods have been proposed to adjust for confounding due to differential cell type composition between study participants, such as sparse principal components analysis, surrogate vari-

able analysis, and mixed models (Leek and Storey, 2007; Zou et al., 2014; Houseman et al., 2014; McGregor et al., 2016; Rahmani et al., 2016). Regardless, failure to adjust for cell type confounding can lead to spurious associations (Jaffe and Irizarry, 2014).

Assuming the mean model is correctly specified, adjusting for cell type composition allows for valid interpretation of the exposure effects on the aggregate whole blood DNAm, but it does not address whether there are cell type specific associations. It is currently cost prohibitive to examine cell specific DNAm profiles. If there are m cell types of interest, we have multiplied our cost by m as there are now m observations for each study participant. There is also the additional cost of fractionation. As costs fall, this may no longer be an issue, but if the sample was frozen, fractionation can not be performed at a later date. Developing methods to circumvent this would therefore be beneficial.

In this paper, we develop a model to estimate the effects of exposures on an unobserved cell type specific DNAm Beta-values using aggregate whole blood methylation, when subject-specific cell composition is measured. We analyze and estimate these effects using Generalized Estimating Equations (GEEs) (Liang and Zeger, 1986). We allow unobserved cell type specific DNAm proportions to be correlated and make robust inference by allowing misspecification of the correlation using the Bias-corrected Sandwich variance estimate (Mancl and DeRouen, 2001) and a modification of the scaled χ^2 approach put forth by Sun et al (Sun et al., 2017).

Our work was motivated by the Normative Aging Study, a closed cohort based in the greater Boston metropolitan area (Bell et al., 1972). Originally started in 1963 by the Department of Veteran Affairs, the study aims to better understand the mortality risks in American men. DNA samples were collected from 2004 to 2007 in whole blood, with DNAm assessed via the Infinium Human Methylation450 BeadChip array. Here, we studied potential cell type specific associations between smoking status and DNAm. We tested for this association in 49 probes that have previously been established as being associated with smoking in the literature (Tsaprouni et al., 2014; Zeilinger et al., 2013).

The rest of the paper is organized as follows. Section 2 presents the model. Section 3 proposes the GEE method for estimation. Section 4 presents the testing procedure. Section 5 investigates the relationship between the regression coefficients in cell type spe-

cific DNAm models and the regression coefficients in standard aggregate DNAm models. Section 6 performs a simulation study and the analysis of the Normative Aging Study DNAm data, followed by discussions.

2.2 The Model

We collect the DNAm proportion at a CpG probe of n subjects in whole blood that is composed of m different cell types. We do not observe the DNAm profiles for these cell types but do observe cell type composition. For subject i ($i = 1, \cdots, n$), we assume that the observed overall whole blood methylation Y_i (Beta-value) at a probe is a weighted sum of unobserved cell type specific DNAm Beta-values $\boldsymbol{y}_i = (y_{i1}, \cdots, y_{im})^T$, with the weights equal to the observed cell type composition $\boldsymbol{\pi}_i = (\pi_{i1}, \cdots, \pi_{im})^T$:

$$Y_i = \sum_{j=1}^{m} \pi_{ij} y_{ij} = \boldsymbol{\pi}_i^T \boldsymbol{y}_i. \tag{2.1}$$

To study the effects of a $(p+1) \times 1$ vector of covariates $\boldsymbol{X}_i$ on cell-specific methylations y_{ij}, we assume cell-specific linear regression models

$$y_{ij} \quad = \boldsymbol{X}_i^T \boldsymbol{\beta}_j + \epsilon_{ij}, \tag{2.2}$$

where $\boldsymbol{\beta}_j$ is the regression coefficient vector measuring the effects of $\boldsymbol{X}_i$ on the j^{th} cell type methylation y_{ij}, and ϵ_{ij} is the residual. We assume that $\boldsymbol{X}_i$ has an intercept term. As the unobserved m cell type specific methylation $\boldsymbol{y}_i$ are likely to be correlated, we assume the residuals $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \epsilon_{i2}, \cdots, \epsilon_{im})^T$ have mean $\mathbf{0}$ and covariance $\boldsymbol{\Sigma}$.

Define $\widetilde{\boldsymbol{X}}_i^T = \boldsymbol{\pi}_i^T \otimes \boldsymbol{X}_i^T = (\pi_{i1}\boldsymbol{X}_i^T, \pi_{i2}\boldsymbol{X}_i^T, \cdots, \pi_{im}\boldsymbol{X}_i^T)$, where the $\otimes$ symbol represents the Kronecker product and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \cdots, \boldsymbol{\beta}_m^T)^T$. Both $\widetilde{\boldsymbol{X}}_i$ and $\boldsymbol{\beta}$ are $m(p+1) \times 1$ vectors. Substituting (2.2) into (2.1) yields:

$$Y_i = \widetilde{\boldsymbol{X}}_i^T \boldsymbol{\beta} + \boldsymbol{\pi}_i^T \boldsymbol{\epsilon}_i. \tag{2.3}$$

Therefore the observed aggregate whole blood methylation (Y_i) follows a linear regression subject to heteroskedasticity, i.e, the residual variance of Y_i varies with subjects and is a quadratic function of cell composition $\{\text{var}(Y_i|\boldsymbol{X}_i, \boldsymbol{\pi}_i) = \boldsymbol{\pi}_i^T \boldsymbol{\Sigma} \boldsymbol{\pi}_i\}$. We propose in the next

section estimation of $\boldsymbol{\beta}$ using Generalized Estimating Equations (GEE) (Liang and Zeger, 1986; Liang et al., 1992).

2.3 Estimation using Generalized Estimating Equations

2.3.1 Estimating Equation for the Regression Coefficients $\boldsymbol{\beta}$

As the true covariance matrix $\boldsymbol{\Sigma}$ of $\boldsymbol{\epsilon}_i$ contains $m(m-1)/2$ parameters, we consider a working covariance matrix $\boldsymbol{D}(\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a vector of variance components that is potentially misspecified. Define v_i as the working variance of Y_i given $\boldsymbol{X}_i$, $\boldsymbol{\pi}_i$ and using $\boldsymbol{D}(\boldsymbol{\theta})$: $v_i = \boldsymbol{\pi}_i^T \boldsymbol{D}(\boldsymbol{\theta}) \boldsymbol{\pi}_i$. As $E(Y_i|\boldsymbol{X}_i, \boldsymbol{\pi}_i) = \widetilde{\boldsymbol{X}}_i^T \boldsymbol{\beta}$, our unbiased estimating equation for $\boldsymbol{\beta}$ is

$$\boldsymbol{U}_\beta = \sum_{i=1}^{n} \widetilde{\boldsymbol{X}}_i v_i^{-1} \left(Y_i - \widetilde{\boldsymbol{X}}_i^T \boldsymbol{\beta} \right) = \widetilde{\boldsymbol{X}}^T \boldsymbol{V}^{-1} (\boldsymbol{Y} - \widetilde{\boldsymbol{X}} \boldsymbol{\beta}), \tag{2.4}$$

where $\widetilde{\boldsymbol{X}} = (\widetilde{\boldsymbol{X}}_1, \cdots, \widetilde{\boldsymbol{X}}_n)^T$, $\boldsymbol{Y} = (Y_1, \cdots, Y_n)^T$, and $\boldsymbol{V} = diag(v_1, \cdots, v_n)$.

The simplest choice of the working covariance $\boldsymbol{D}(\boldsymbol{\theta})$ is to assume working independence and a homogeneous variance among the cell-specific methylations, i.e., $\boldsymbol{D} = \sigma^2 \boldsymbol{I}_m$, where $\boldsymbol{I}_m$ is the identity matrix. This choice can be relaxed by a working independence covariance with cell-specific variances, i.e., $\boldsymbol{D} = diag(\sigma_1^2, \cdots, \sigma_m^2)$, where $\sigma_j^2 = var(e_{ij})$; or by an exchangeable covariance with cell-specific variances, i.e. $\boldsymbol{D} = \boldsymbol{\Sigma}_0^{1/2} \{(1-\rho)\boldsymbol{I} + \rho \boldsymbol{J}\} \boldsymbol{\Sigma}_0^{1/2}$, where $\boldsymbol{\Sigma}_0 = diag(\sigma_1^2, \cdots, \sigma_m^2)$ and $\boldsymbol{J}$ is a m by 1 matrix of ones. The most complex working covariance matrix is unstructured by setting $\boldsymbol{D} = \boldsymbol{\Sigma}$.

One can easily see that (2.4) is an unbiased estimating equation of $\boldsymbol{\beta}$ and the solution $\widehat{\boldsymbol{\beta}}$ is consistent even when $\boldsymbol{D}$ is misspecified. Given $\boldsymbol{\theta}$, the estimator $\widehat{\boldsymbol{\beta}}$ has a closed form solution as $\widehat{\boldsymbol{\beta}} = \left(\widetilde{\boldsymbol{X}}^T \boldsymbol{V}^{-1} \widetilde{\boldsymbol{X}} \right)^{-1} \widetilde{\boldsymbol{X}}^T \boldsymbol{V}^{-1} \boldsymbol{Y}$. Given $\boldsymbol{\theta}$, the covariance of $\widehat{\boldsymbol{\beta}}$ can be estimated using a sandwich estimator (Liang and Zeger, 1986) as

$$\boldsymbol{V}_S = cov(\widehat{\boldsymbol{\beta}}) = \boldsymbol{V}_m \left\{ \sum_{i=1}^{n} \widetilde{\boldsymbol{X}}_i \frac{\left(Y_i - \widetilde{\boldsymbol{X}}_i^T \widehat{\boldsymbol{\beta}} \right)^2}{v_i^2} \widetilde{\boldsymbol{X}}_i^T \right\} \boldsymbol{V}_m, \tag{2.5}$$

where $\boldsymbol{V}_m = (\widetilde{\boldsymbol{X}}^T \boldsymbol{V}^{-1} \widetilde{\boldsymbol{X}})^{-1}$. We can use other modifications of the robust variance true estimators and will introduce one later. The most efficient estimator of $\boldsymbol{\beta}$ is obtained

when $\boldsymbol{D} = \boldsymbol{\Sigma}$. Under the homogeneous working independence $\boldsymbol{D} = \sigma^2 \boldsymbol{I}_m$, the working variance of Y_i becomes $v_i = \sigma^2 \sum_j \pi_{i,j}^2$. The linear regression in (2.3) reduces to a simple weighted linear regression with individual weights $w_i = (\sum_j \pi_{i,j}^2)^{-1}$.

2.3.2 Estimating Equation for the Variance Components $\boldsymbol{\theta}$

To construct estimating equations for the $\boldsymbol{\theta}$, assuming $cov(\boldsymbol{\epsilon}_i) = \boldsymbol{D}(\boldsymbol{\theta})$, we have:

$$v_i(\boldsymbol{\pi}_i, \boldsymbol{\theta}) = v_i = E\{(Y_i - \widetilde{\boldsymbol{X}}_i^T \boldsymbol{\beta})^2 | \boldsymbol{X}_i, \boldsymbol{\pi}_i\} = \boldsymbol{\pi}_i^T \boldsymbol{D}(\boldsymbol{\theta}) \boldsymbol{\pi}_i.$$

We can construct quadratic estimating equations for $\boldsymbol{\theta}$ as follows (Liang et al., 1992)

$$\boldsymbol{U}_\theta \quad = \sum_{i=1}^n \frac{\partial v_i(\boldsymbol{\pi}_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \left\{ \left(Y_i - \widetilde{\boldsymbol{X}}_i^T \boldsymbol{\beta} \right)^2 - v_i(\boldsymbol{\pi}_i, \boldsymbol{\theta}) \right\}. \tag{2.6}$$

If $\boldsymbol{D}(\boldsymbol{\theta})$ is nonlinear in $\boldsymbol{\theta}$, (2.6) does not have a closed form solution and can be solved iteratively using the Newton-Raphson algorithm. The estimators of the regression coefficient $\boldsymbol{\beta}$ and the variance components $\boldsymbol{\theta}$ can be obtained by solving (2.4) and (2.6) back and forth until convergence.

If the elements of $\boldsymbol{D}(\boldsymbol{\theta})$ are linear in $\boldsymbol{\theta}$, then v_i is a linear function of $\boldsymbol{\theta}$ and can be written as $v_i = \boldsymbol{Z}_i^T \boldsymbol{\theta}$. For example, if $\boldsymbol{D}(\boldsymbol{\theta})$ is a working independence covariance with cell-specific variances $\boldsymbol{D} = diag(\sigma_1^2, \cdots, \sigma_m^2)$, then $v_i = \sum_{j=1}^m \pi_{ij}^2 \sigma_j^2$, $\boldsymbol{Z}_i = (\pi_{1i}^2, \cdots, \pi_{im}^2)^T$ and $\boldsymbol{\theta} = (\sigma_1^2, \cdots, \sigma_m^2)^T$. If $\boldsymbol{D} = \boldsymbol{\Sigma}$, then

$$v_i = \left(\sum_{j=1}^m \pi_{ij}^2 \sigma_j^2 \right) + \sum_{j=1}^{m-1} \sum_{k=j+1}^m 2\pi_{ij}\pi_{ik}\sigma_{jk} = \boldsymbol{Z}_i^T \boldsymbol{\theta},$$

where

$$\boldsymbol{Z}_i = (\pi_{i1}^2, \cdots, \pi_{im}^2, 2\pi_{i1}\pi_{i2}, \ldots, 2\pi_{i,m-1}\pi_{im})^T,$$
$$\boldsymbol{\theta} = (\sigma_1^2, \cdots, \sigma_m^2, \sigma_{12}, \ldots, \sigma_{m-1,m})^T,$$

then (2.6) becomes

$$\boldsymbol{U}_\theta = \sum_{i=1}^n \boldsymbol{Z}_i \left\{ (Y_i - \widetilde{\boldsymbol{X}}_i^T \boldsymbol{\beta})^2 - \boldsymbol{Z}_i^T \boldsymbol{\theta} \right\}.$$

The solution $\widehat{\boldsymbol{\theta}}$ has a closed form as

$$\widehat{\boldsymbol{\theta}} = \left(\sum_{i=1}^n \boldsymbol{Z}_i \boldsymbol{Z}_i^T \right)^{-1} \left\{ \sum_{i=1}^n \boldsymbol{Z}_i (Y_i - \widetilde{\boldsymbol{X}}_i^T \boldsymbol{\beta})^2 \right\}.$$

For $\boldsymbol{\theta}$ to be identifiable, the design matrix $\boldsymbol{Z} = (\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n)^T$ needs to be full rank and the mean values of $\boldsymbol{\pi}$ need to be reasonably large. It is necessary that there is sufficient variation in $\boldsymbol{\pi}$ across subjects. If there is insufficient variation in the $\boldsymbol{\pi}_i$'s or the structure of $\boldsymbol{D}(\boldsymbol{\theta})$ is complex and involves many parameters, the eigenvalues of $\sum_{i=1}^{n} \boldsymbol{Z}_i^T \boldsymbol{Z}_i$ are likely to be close to 0 and the estimator $\widehat{\boldsymbol{\theta}}$ will be unstable. Even if $\boldsymbol{Z}$ is full rank, there is no guarantee that the projection $\left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}^T$ will map to a $\hat{\boldsymbol{\theta}}$ such that $\boldsymbol{D}(\hat{\boldsymbol{\theta}})$ is positive definite.

Given cell-specific methylations ($y_{i,j}$) are all missing and only the overall whole blood methylations are observed, if $\boldsymbol{D}(\boldsymbol{\theta})$ is unstructured, there is more often than not insufficient information in the data to stably estimate the many variance and covariance parameters. We recommend in practice to assume simpler structured working covariance matrix such as the homogeneous working independence covariance, $\boldsymbol{D} = \sigma^2 \boldsymbol{I}_m$. Assuming working independence covariance ($\boldsymbol{D} = diag(\sigma_1^2, \cdots, \sigma_m^2)$) can still result in estimates outside of the parameter space. Note that misspecification of $\boldsymbol{D}$ still gives consistent estimators of the regression coefficients $\boldsymbol{\beta}$, although it is subject to loss of efficiency. We evaluate the degree of efficiency loss in the simulation studies of $\boldsymbol{D} = \sigma^2 \boldsymbol{I}_m$ vs $\boldsymbol{D} = diag(\sigma_1^2, \cdots, \sigma_m^2)$. To ensure that the estimators of σ^2's are positive, we can fix σ_j^2 to zero if it converges outside of the parameter space or use a non negative least square approach to solve (2.6) when v_i is linear in $\boldsymbol{\theta}$.

2.4 Testing

The traditional robust variance of $\widehat{\boldsymbol{\beta}}$ was given in equation (2.5) with $\boldsymbol{\theta}$ estimated by $\hat{\boldsymbol{\theta}}$. Empirical evidence has shown that the traditional sandwich estimate is subject to large bias in finite sample unless the sample size is sufficiently large (Mancl and DeRouen, 2001; MacKinnon and White, 1985). To overcome this problem, Mancl and DeRouen (2001) proposed the bias corrected sandwich variance as

$$\boldsymbol{V}_{BC} = \boldsymbol{V}_m \left\{ \sum_{i=1}^{n} \widetilde{\boldsymbol{X}}_i \frac{\left(Y_i - \widetilde{\boldsymbol{X}}_i^T \widehat{\boldsymbol{\beta}}\right)^2}{(1-H_i)^2 \hat{v}_i^2} \widetilde{\boldsymbol{X}}_i^T \right\} \boldsymbol{V}_m, \tag{2.7}$$

where $H_i = \widetilde{\boldsymbol{X}}_i^T \boldsymbol{V}_m \widetilde{\boldsymbol{X}}_i v_i^{-1}$. This Bias Corrected is similar in the univariate case to the heteroskadicity variance estimator HC3 (Long and Ervin, 2000; MacKinnon and White, 1985). We wish to test for a $\boldsymbol{C}$ contrast matrix ($d_1 \times (m * (p+1))$ matrix whether $\boldsymbol{C}\hat{\boldsymbol{\beta}} = 0$. Mancl and DeRouen (2001) proposed to use the test statistic $Q = (\boldsymbol{C}\hat{\boldsymbol{\beta}})^T (\boldsymbol{C}\hat{\boldsymbol{V}}_{BC}\boldsymbol{C}^T)^{-1}\boldsymbol{C}\hat{\boldsymbol{\beta}}$ to test for $H_0 : \boldsymbol{C}\boldsymbol{\beta} = 0$ by comparing Q with a chi-square distribution with degrees of freedom equal to the length of $\boldsymbol{C}\hat{\boldsymbol{\beta}}$, d_1. We do not, as in simulations (not shown here) with this type of data, this approach did not control type I error.

Pan and Wall (2002) state that for small samples if $\boldsymbol{V}_S$ or $\boldsymbol{V}_{BC}$ is highly variable, the test statistic for $\boldsymbol{\beta}_{(k)}$ approximately follows a Hotelling T^2, which is a scaled F distribution. Pan & Wall propose that if $Q \sim$ Hotelling $T^2(d_1, v)$, due to $\boldsymbol{V}_S \sim Wishart(v)$ then $\frac{v-d_1+1}{vp} Q \sim F(d_1, v - d_1 + 1)$. The term v is estimated from the covariance of $\boldsymbol{V}_S$ (or $\boldsymbol{V}_{BC}$). We do not use this method as in real data applications it estimated v such that $v - d_1 + 1 < 0$. We propose to estimate the scale parameter in the F distribution instead via bootstrap. Specifically, we relax the assumption that $(v - d_1 + 1)/vp\, Q \sim F(d_1, v - d_1 + 1)$ and instead propose that $Q \sim cF(d_1, d_2)$ and estimate c and d_2. We generate B bootstrap samples and calculate the bootstrap test statistics $Q_{(k),l}$ of the form:

$$Q_{(k),l} = \{\boldsymbol{C}\hat{\boldsymbol{\beta}}_l - \boldsymbol{C}\hat{\boldsymbol{\beta}}\}^T (\boldsymbol{C}\hat{\boldsymbol{V}}_{BC,l}\boldsymbol{C}^T)^{-1}\{\boldsymbol{C}\hat{\boldsymbol{\beta}}_l - \boldsymbol{C}\hat{\boldsymbol{\beta}}\},$$

where $l (l = 1 \cdots B)$ represents the l^{th} bootstrap sample. Briefly, we estimate c and d_2 from the first and second moments of the log of Q_l, which follows a Fisher Z-distribution plus a constant using the R-package *limma* (Phipson et al., 2016; Smyth, 2004). Upon estimating c and d_2, we compare $Q/\hat{c}$ to the quantiles of an F-distribution with d_1 numerator degrees of freedom and $\hat{d}_2$ denominator degrees of freedom.

2.5 Relationship between Aggregate and Cell Model

There is a simple linear relationship between our proposed model and the traditional aggregate analyses. The aggregate model specifies that $E(Y_i|\boldsymbol{X}_i, \boldsymbol{\pi}_i) = \boldsymbol{\pi}_i^T \boldsymbol{\gamma}_\pi + \boldsymbol{X}_i^T \boldsymbol{\gamma}_X = \boldsymbol{S}_i^T \boldsymbol{\gamma}$, with $\boldsymbol{S}_i^T = (\boldsymbol{\pi}_i^T, \boldsymbol{X}_i^T)$ and $\boldsymbol{\gamma}^T = (\boldsymbol{\gamma}_\pi^T, \boldsymbol{\gamma}_X^T)$. We assume that $\boldsymbol{X}_i$ has been re-

parametrized to no longer include the intercept, making (2.1):

$$Y_i = \sum_{j=1}^{m}(\pi_{i,j}\beta_{0,j} + \pi_{i,j}\boldsymbol{X}_i^T\boldsymbol{\beta}_j) + \boldsymbol{\pi}_i^T\epsilon_i = \boldsymbol{\pi}_i^T\boldsymbol{\beta}_0 + \widetilde{\boldsymbol{X}}_i^T\boldsymbol{\beta}_X + \boldsymbol{\pi}_i^T\boldsymbol{\epsilon}_i,$$

where $\boldsymbol{\beta}_0 = (\beta_{0,1}, \beta_{0,2}, \cdots, \beta_{0,m})^T$, $\boldsymbol{\beta}_X = (\boldsymbol{\beta}_1^T, \cdots, \boldsymbol{\beta}_m^T)^T$, and $\widetilde{\boldsymbol{X}}_i = (\pi_{i,1}\boldsymbol{X}_i^T, \cdots, \pi_{i,m}\boldsymbol{X}_i)^T$. Let $\mathbf{0}_{p\times m}$ be a p by m matrix of zeros. Assuming that the mean model is correctly specified, we have the following relationship between $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$:

$$\boldsymbol{\gamma} = \begin{bmatrix}\boldsymbol{I}_m \\ \mathbf{0}_{p\times m}\end{bmatrix}\boldsymbol{\beta}_\pi + E\left(\boldsymbol{S}_i\boldsymbol{S}_i^T\right)^{-1}\begin{bmatrix}E\left\{\boldsymbol{\pi}_i\boldsymbol{\pi}_i^T(\boldsymbol{I}_m \otimes \boldsymbol{X}_i^T)\right\} \\ E\left\{\boldsymbol{X}_i\boldsymbol{X}_i^T(\boldsymbol{\pi}_i^T \otimes \boldsymbol{I}_p)\right\}\end{bmatrix}\boldsymbol{\beta}_X. \tag{2.8}$$

Define $\boldsymbol{\mu}_x = E(\boldsymbol{X}_i)$ and $\boldsymbol{\mu}_\pi = E(\boldsymbol{\pi})$. If $\boldsymbol{\pi}_i$ is independent of $\boldsymbol{X}_i$ (2.8) simplifies to:

$$\boldsymbol{\gamma}_\pi = \boldsymbol{\beta}_0 + (\boldsymbol{I}_m - \boldsymbol{J}_m\boldsymbol{\mu}_\pi^T)(\boldsymbol{I}_m \otimes \boldsymbol{\mu}_X^T)\boldsymbol{\beta}_X, \tag{2.9}$$

$$\boldsymbol{\gamma}_X = (\boldsymbol{\mu}_\pi^T \otimes \boldsymbol{I}_p)\boldsymbol{\beta}_X = \sum_{j=1}^{m}\mu_{\pi,j}\boldsymbol{\beta}_j. \tag{2.10}$$

The proof is given in Appendix 2.8.1. These indicates that the regression coefficient in the aggregate analysis are a weighted sum of the cell-specific regression coefficients. Therefore if there is no aggregate association, it does not necessarily imply that the cell specific associations are also null. There could be probes where the cell specific effects are in opposite directions. Using the above result, we can construct an estimate of the aggregate regression coefficients using $\hat{\boldsymbol{\beta}}$ and the sample means and covariances of the observed $\boldsymbol{\pi}_i$'s and $\boldsymbol{X}_i$'s.

2.6 Simulation Study and Real Data Application

2.6.1 Point Estimation

We performed simulation studies to examine the estimation properties of our method. We set $\boldsymbol{X}_i$ =$(1, X_{i,1}, X_{i,2})^T$, with $X_{i,1}$ generated from a normal with mean 45 and variance 5 and $X_{i,2}$ from a Bernoulli with probability 0.3. In all simulations, we set n=1000. To mimic the Normative Aging Study, we considered three cell types: Granulocytes, Lymphocytes, and Monocytes (Figure 2.1). To generate $\boldsymbol{\pi}_i$, we first simulate $\boldsymbol{a}_i$ from a half normal ($N_{\boldsymbol{a}>0}$):

$$\boldsymbol{a}_i \sim N_{\boldsymbol{a}>0}(\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_a), \tag{2.11}$$

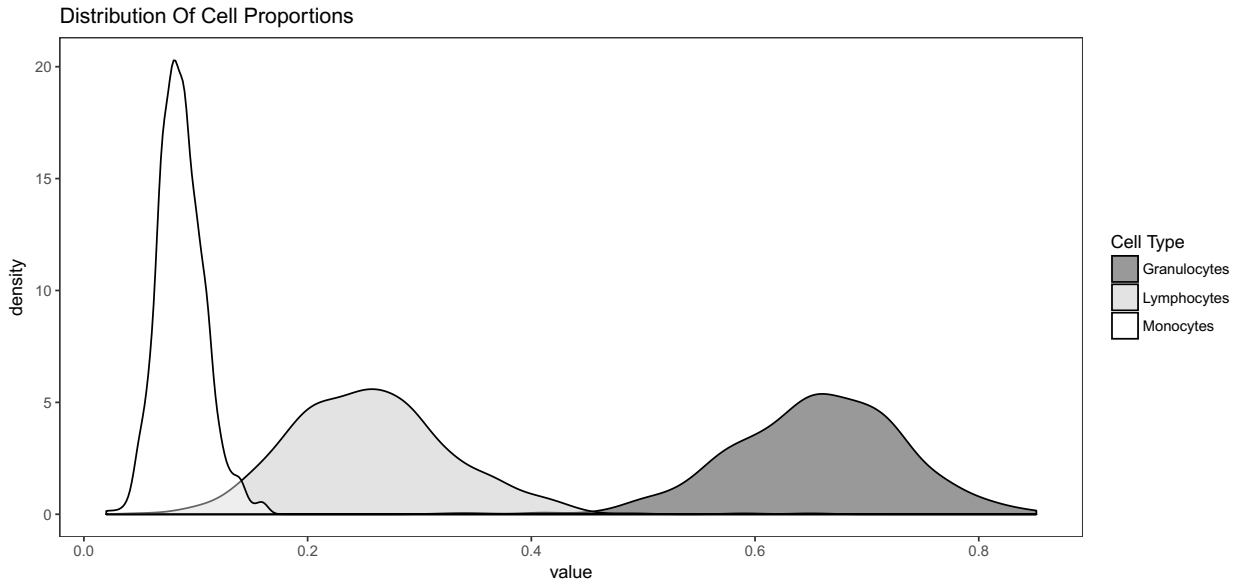

Figure 2.1: Distribution of $\boldsymbol{\pi}$ in Normative Aging Study of 618 individuals.

$$\boldsymbol{\mu}_a = \begin{bmatrix} 65 \\ 25 \\ 10 \end{bmatrix}, \boldsymbol{\Sigma}_a = \begin{bmatrix} 56.57 & -51.9 & -4.22 \\ -51.9 & 51.88 & -0.038 \\ -4.22 & -0.038 & 4.32 \end{bmatrix},$$

and then set $\boldsymbol{\pi}_i$=$\boldsymbol{a}_i / \sum_{j=1}^{3} \boldsymbol{a}_{i,j}$. Each unobserved $\boldsymbol{y}_i$ is:

$$y_{i,j} = \beta_{j,0} + \beta_{j,1} X_{i,1} + \beta_{j,2} X_{i,2} + \epsilon_{i,j}, \tag{2.12}$$

where $\boldsymbol{\epsilon}_i \sim N_3(\mathbf{0}, \boldsymbol{\Sigma})$. $\boldsymbol{\Sigma}$ is of the form:

$$\boldsymbol{\Sigma} = \begin{bmatrix} 0.5 & \rho_{1,2}\sqrt{0.15} & \rho_{1,3}\sqrt{0.075} \\ \rho_{1,2}\sqrt{0.15} & 0.3 & \rho_{2,3}\sqrt{0.045} \\ \rho_{1,3}\sqrt{0.075} & \rho_{2,3}\sqrt{0.045} & 0.15 \end{bmatrix}.$$

The off-diagonal terms were varied to generate exchangeable, unstructured or an independent covariance structure. For exchangeable, $\rho_{1,2} = \rho_{1,3} = \rho_{2,3} = 0.7$; for unstructured, $\rho_{1,2} = 0.7, \rho_{1,3} = -0.2$, and $\rho_{2,3} = 0.1$ and for independent all ρ's were set to 0. For each replication, we examined the performance when the working covariance was homogeneous or when working independence was heterogeneous.

We assessed the bias and variance of our estimate of $\boldsymbol{\beta}$ by generating 10^4 data sets for each of the $\boldsymbol{\Sigma}$'s. We set $\boldsymbol{\beta}_1 = (0.2, 0, 0.2)^T$, $\boldsymbol{\beta}_2 = (0.2, 0.02, 0)^T$, and $\boldsymbol{\beta}_3 = (0.2, 0, 0.1)^T$. We also calculated the aggregate estimates $\boldsymbol{\gamma}$ when $\boldsymbol{\pi}_i$ is included as a covariate in the model to assess the validity of () and ().

Parameter	True Value	Mean D_1	Mean D_2	Empirical Var D_1	Empirical Var D_2	Mean $\boldsymbol{V}_{BC}$ D_1	Mean $\boldsymbol{V}_{BC}$ D_2
$\beta_{1,1}$	0.00	0.0001	0.0001	0.0024	0.0024	0.0025	0.0024
$\beta_{1,2}$	0.20	0.2007	0.2009	0.0576	0.0576	0.0578	0.0574
$\beta_{2,1}$	0.02	0.0186	0.0187	0.0086	0.0086	0.0087	0.0087
$\beta_{2,2}$	0.00	-0.0014	-0.0019	0.2023	0.2025	0.2054	0.2039
$\beta_{3,1}$	0.00	0.0035	0.0035	0.1242	0.1245	0.1240	0.1233
$\beta_{3,2}$	0.10	0.1041	0.1041	2.9320	2.9320	2.9297	2.9123

Table 2.1: Mean, empirical variance, and mean variance estimate of 10^4 simulations comparing when assume $D(\boldsymbol{\theta}) = \sigma^2 \boldsymbol{I}_m$ (D_1) vs $D(\boldsymbol{\theta}) = diag(\sigma_1^2, \cdots, \sigma_m^2)$ (D_2). For D_2, 4 of the simulations set $\sigma_2^2 = 0$ and 5048 set $\sigma_3^2 = 0$. True covariance was unstructured.

Table 2.1 displays the results when the true covariance structure is unstructured. We have relatively unbiased estimates of the $\boldsymbol{\beta}$ parameters. We see an increase in the variance of the parameter estimates for effects in smaller cell types. The variance estimated from $\boldsymbol{V}_{BC}$, while slightly larger than the empirical variance, are close (Table 2.1). When we assume $\boldsymbol{D}(\boldsymbol{\theta}) = diag(\sigma_1^2, \cdots, \sigma_m^2)$ (D_2 in Table 2.1), half of the σ_3's are held at zero as are some σ_2 (5048 and 4 respectively). As expected since the true covariance is unstructured, the mean estimates of these parameters are biased (mean values $\hat{\sigma}_1^2 = 0.579, \hat{\sigma}_2^2 = 0.833, \hat{\sigma}_3^2 = 2.40$). As mentioned previously, there is no guarantee that the projection to estimate $\boldsymbol{\theta}$ will be such that $\boldsymbol{D}(\hat{\boldsymbol{\theta}})$ is positive definite. There is however little loss in efficiency by using the simpler homogenous independent working variance ($\boldsymbol{D} = \sigma^2 \boldsymbol{I}_m$, D_1 in Table 2.1) as opposed to the working heterogeneous independence (D_2). There perhaps would be a gain in efficiency if half of the $\boldsymbol{\Theta}$ had not converged at the parameter boundary. Therefore, for the rest of the paper we only report the results from this simpler working covariance. We see similar results when the true covariance structure is independent or exchangeable (Supplementary Tables 4.7 and 4.8).

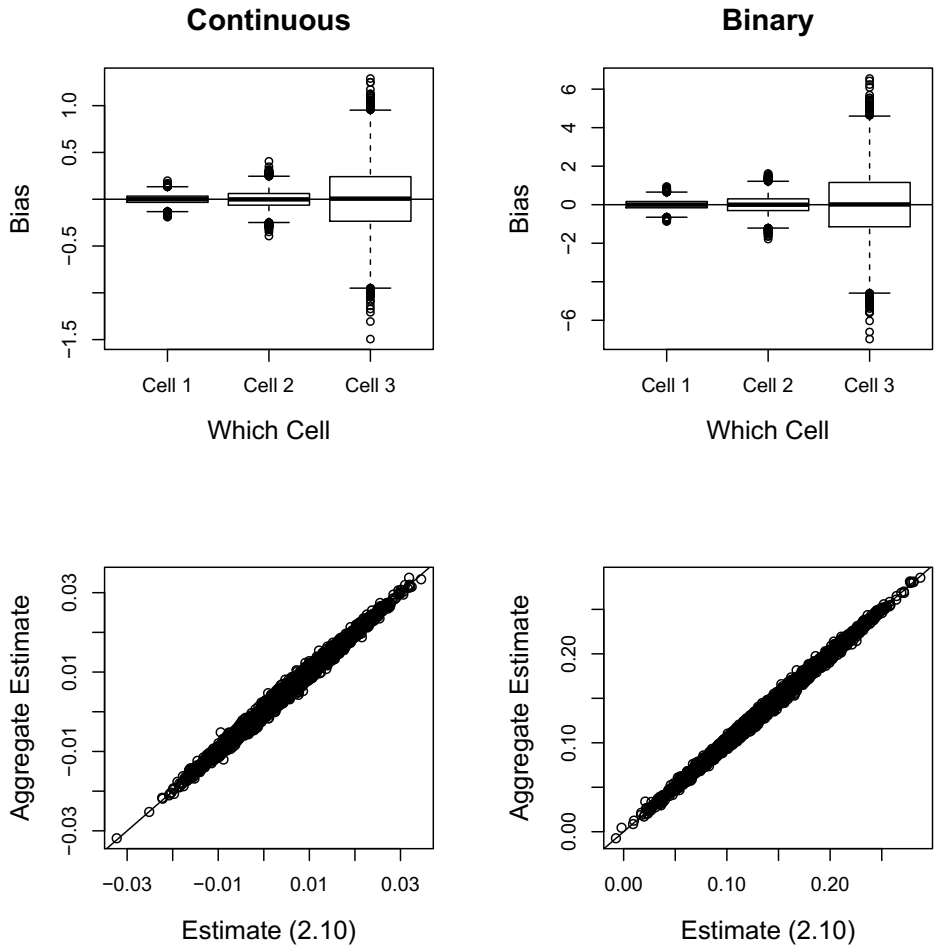

Figure 2.2: Bias and relationship between γ_X and (2.6.1). 10^4 replications, true variance is unstructured, assuming $\boldsymbol{D} = \sigma^2 \boldsymbol{I}_m$.

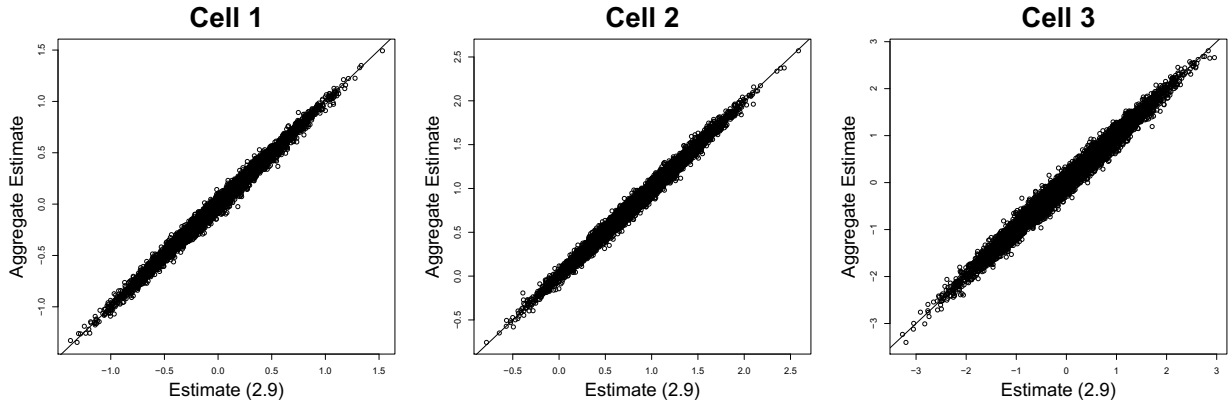

Figure 2.3: Comparing aggregate estimates of γ_π and (2.6.1) from 10^4. True variance is unstructured, assuming $\boldsymbol{D} = \sigma^2 \boldsymbol{I}_m$.

We next compared the aggregate estimates ($\hat{\gamma}$) to (2.6.1) substituting for $\boldsymbol{\beta}_X$, $\hat{\boldsymbol{\beta}}$ and the sample mean of $\boldsymbol{\pi}$ for $\boldsymbol{\mu}_\pi$. As expected, we see strong agreement between the two (Figure 2.2). In Figure 2.2, we also display the bias with unbiased parameters on average and an increasing variance for parameters corresponding to smaller cell type. In Figure 2.3, we compare the terms from (2.6.1) and see agreement between $\hat{\boldsymbol{\gamma}}_\pi$ and the transformation of $\hat{\boldsymbol{\beta}}_\pi$ and $\hat{\boldsymbol{\beta}}_x$. We had identical results when the true covariance structure was exchangeable or independent (Supplement Figures 4.11 to 4.14).

2.6.2 Hypothesis testing

We next assessed the models type I error and power. We set n=1000 for each simulation. We only assessed the homogenous independent working covariance (D_1) as we did not see a loss in efficiency compared to the heterogenous independent structure (D_1 vs D_2 Table 2.1). To consider the size of the test on a stringent genome wide scale, we generated 2x10^5 data sets for each possible $\boldsymbol{\Sigma}$. We generated $\boldsymbol{y}_i$ from (2.12) and set $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \boldsymbol{\beta}_3 = (0.2, 0, 0)^T$. We examined the TIE of four different tests: the overall (testing for association across all three cell types), and all individual tests for an association in each cell type.

In Table 2.2, we display the type I error evaluated at four different α's when the true covariance structure is unstructured. Our testing procedure controls the type I error rates for all tests (overall and in each cell) for both continuous and binary variables. Results when the unobserved structure is independent or exchangeable are provided in Supplementary Tables 4.9 and 4.10. We see slight inflation in type I error when $\alpha = 10^{-5}$, but this maybe due to only performing 2x10^5 simulations. QQ-plots comparing the expected and observed p-values are provided in the supplement and show good agreement with the expected distribution under the null (Supplementary Figures 4.15, 4.16, and 4.17).

We next examined the empirical power. The $\boldsymbol{\beta}$'s were set such that $X_{i,1}$ or $X_{i,2}$ explained some proportion of the variation in $y_{i,j}$. We examined the power under four different scenarios. One where there was an association in every cell type and then three where there was an association in only one cell type ($y_{i,1}, y_{i,2}$ or $y_{i,3}$). We simulated each scenario twice, once where just $X_{i,1}$ was associated with $\boldsymbol{y}_i$ and another where it was just $X_{i,2}$. When the association was in all cell types, the parameter of interest ($X_{i,1}$ or $X_{i,2}$)

	Continuous				Categorical			
Level α	Overall	Cell 1	Cell 2	Cell 3	Overall	Cell 1	Cell 2	Cell 3
0.05	0.04962	0.04879	0.04832	0.04850	0.04863	0.04927	0.04807	0.04913
0.01	0.00987	0.00938	0.00926	0.00938	0.00966	0.00946	0.00918	0.00970
10^{-3}	0.00094	0.00103	0.00089	0.00104	0.00101	0.00092	0.00088	0.00105
10^{-5}	0.00001	0.00002	0.00002	0.00001	0.00001	0.00003	0.00000	0.00003

Table 2.2: Type I error results for continuous and binary variable when the true covariance is unstructured. Each value represents the proportion of 2x10^5 p-values smaller than α.

explained the same proportion of variance in each DNAm cell type ($y_{i,j}$). We compared to the standard aggregate analysis which includes $\boldsymbol{\pi}_i$ as an additional covariate in the model and tests for $\boldsymbol{\gamma}$. $\beta_{j,0}$ was set to 0.2 for each cell type. We then performed 10^4 replications for each level of association and true covariance structure. Power was assessed at $\alpha = 0.05$.

Figure 2.4 displays the power of the overall test (across cell types) and the aggregate when the true covariance is unstructured. Independent and exchangeable are provided in Supplementary Figures 4.19 to 4.24. The overall test has less power than the aggregate test in all scenarios. The decrease in power is likely due to our test being equivalent to a three df test, while the aggregate is a one df test of a weighted sum of the $\boldsymbol{\beta}$ terms being 0 (2.6.1). In Scenario 1, when the exposure is associated with all cell types, we have adequate power to detect an association when $X_{i.1}$ or $X_{i,2}$ explains a minimum of 1% of the variation in each $y_{i,j}$. When there is an association only in $y_{i,1}$ (Scenario 2), the overall and aggregate have smaller power to detect an association. For Scenarios 3 and 4, there is no power to detect any association with this low of effect sizes. In Supplement Figure 4.18, we display the results for large effect sizes and we briefly summarize here. Under Scenario 2, once $X_{i,1}$ or $X_{i,2}$ explains 5% of the variation in $y_{i,1}$, the overall test and the aggregate have maximum power to detect an association. Under Scenario 3, when X explains 15% of the variation in $y_{i,2}$ both tests have power of approximately 80%. Finally for Scenario 4, when there is only an association in the smallest cell type $y_{i,3}$, X needs to explain upward of 70% of the variation in $y_{i,3}$ to have adequate power.

In Figure 2.5 ,we display the power results for testing for a cell specific effect. For $y_{i,1}$, we have over 80% power when X explains 20% of its variance. For the second most

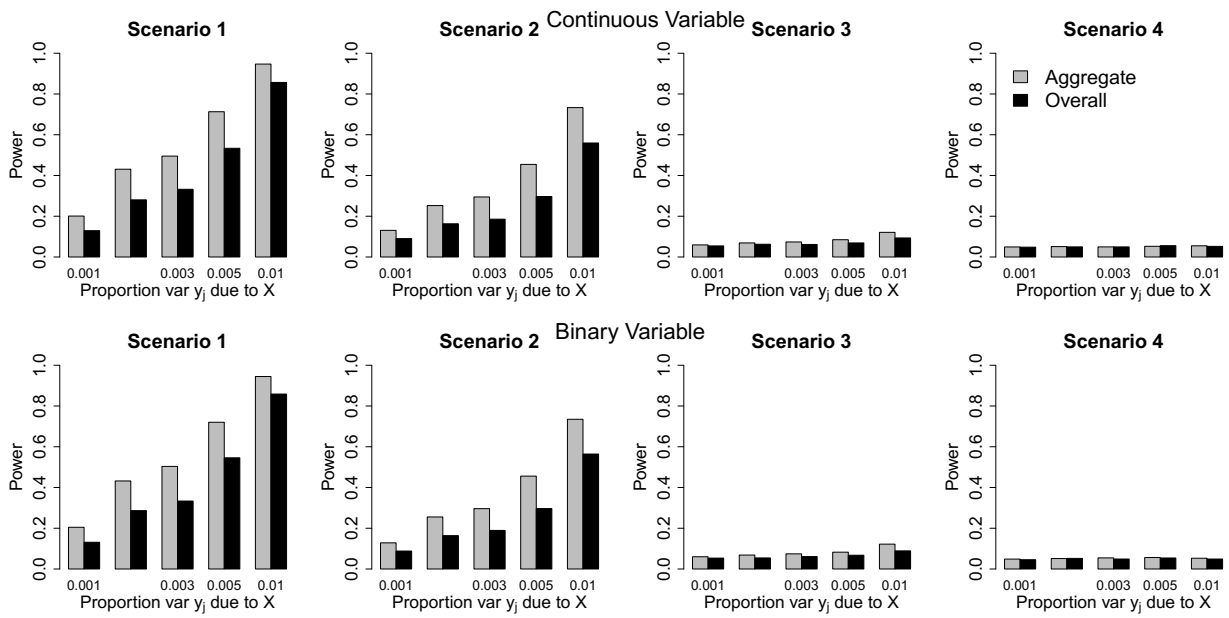

Figure 2.4: Power results of overall and aggregate, the true covariance is unstructured, $\alpha = 0.05$. Top panel corresponds to continuous while bottom corresponds to binary variable. From left to right, association in all cell types (Scenario 1), just cell 1(Scenario 2), just cell 2 (Scenario 3), just cell 3 (Scenario 4). Variables were set to explain .1%, .25%,.3%,.5%,1% of the variation in the $y_{i,j}$'s.

prevalent cell type, the variables need to explain 60% of the variation in $y_{i,2}$ before we have adequate power. Finally for cell type 3, even when the variables explains 90% of the variation in $y_{i,3}$ we only have between 30% and 40% power (Figure 2.5).

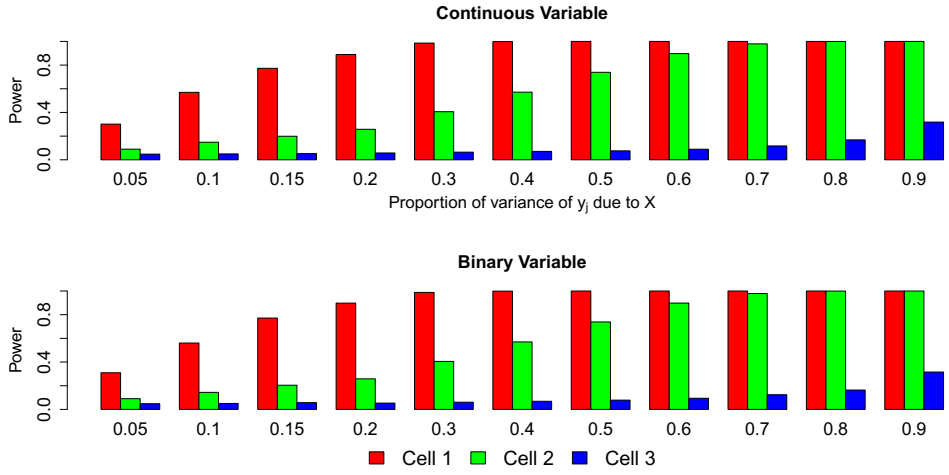

Figure 2.5: Power to detect individual cell specific effects. True covariance is unstructured, at $\alpha = 0.05$. Top panel is continuous variable, bottom panel is binary.

We next evaluated when the aggregate has no power to detect an association. We set $\beta_{21} = -E(\pi_{i,1})(\beta_{11})/E(\pi_{i,2}) * 1.05$, thus making the aggregate effect approximately 0 (2.6.1). $\beta_{1,1}$ was set such that $X_{i,1}$ explained 5% of the variation in $y_{i,1}$, and $\beta_{3,1} = 0$. $X_{i,2}$ was not associated with $\boldsymbol{y}_i$. We performed 10^4 simulations of this situation

In Figure 2.6, we see that the aggregate test has no power to detect an association, due to the effects being in opposite direction making $\gamma_1 \approx 0$. The overall test has power to detect an association in the cells, while we have largest power for testing $\beta_{2,1} = 0$ (Cell 2 in Figure 2.6). The overall test having less power is due to the unobserved cell specific effects being in opposite direction. The high power of cell type 2 is due to $\beta_{2,1}$ being equal to $\beta_{1,1}$ multiplied by $E(\pi_{i,1})/E(\pi_{i,2}) * 1.05 = 2.73$. This was to guarantee that the effects cancel out on the aggregate level.

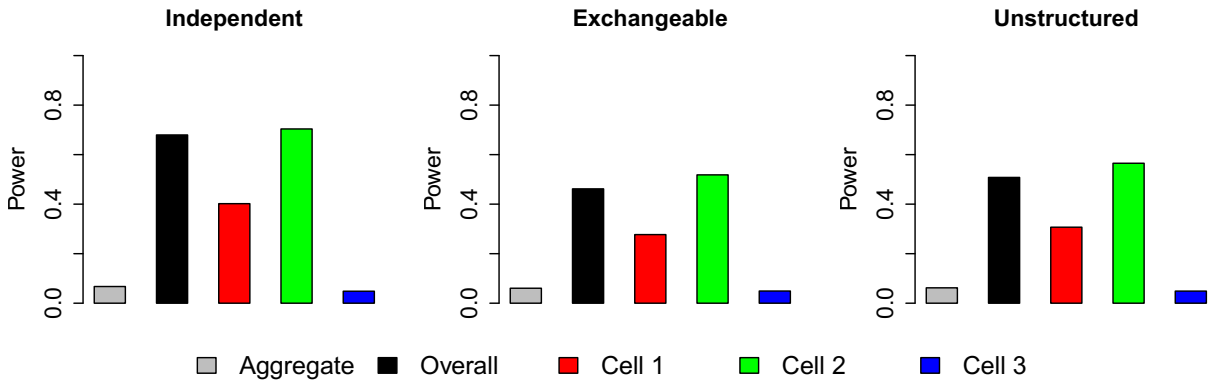

Figure 2.6: Power when $\boldsymbol{\gamma}_X \approx 0$, due to $\boldsymbol{\beta}_X$ being in opposite direction, α=0.05. Left to right, when true covariance is independent, exchangeable, or unstructured.

2.6.3 Covariates associated with cell type composition

We next examined the effect of correlation between $\boldsymbol{\pi}$ and $\boldsymbol{X}$. $\boldsymbol{\pi}_i$ was generated such that its distribution was determined by $X_{i,2}$. If $X_{i,2}$=1, $\boldsymbol{a}_i$ was generated as before, while if $X_{i,2} = 0$, $\boldsymbol{\mu}_a = (10, 25, 65)^T$ and we used $\boldsymbol{I}_{a,3}\boldsymbol{\Sigma}\boldsymbol{I}_{a,3}$ as the covariance, where $\boldsymbol{I}_{a,3}$ is the anti-identity matrix. This rotates $\boldsymbol{\Sigma}$, making the variance of $a_{i,1}$ 4.32 and so on (2.11). We set $\boldsymbol{\beta}$ equal to the values under the point estimation simulations. and did 10^4 replications.

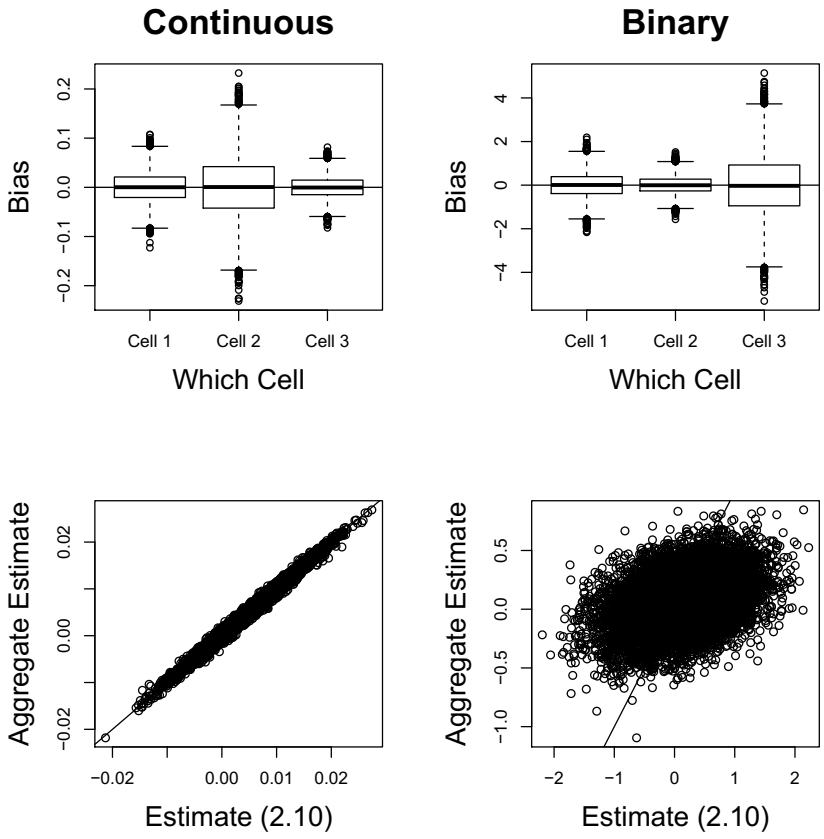

Figure 2.7: Results when X_2 confounds $\boldsymbol{\pi}$ distribution. Bias and relationship between aggregate estimates (2.6.1), 10^4 replications, true covariance unstructured, assuming $\boldsymbol{D} = \sigma^2 \boldsymbol{I}_m$.

Parameter	True Value	Mean D_1	Empirical Var D_1
$\beta_{1,1}$	0.00	0.0002	0.0010
$\beta_{1,2}$	0.20	0.2032	0.3294
$\beta_{2,1}$	0.02	0.0201	0.0038
$\beta_{2,2}$	0.00	0.0027	0.1628
$\beta_{3,1}$	0.00	-0.0002	0.0005
$\beta_{3,2}$	0.10	0.0845	1.9514

Table 2.3: Mean estimates and empirical variance when X_2 confounds $\boldsymbol{\pi}$. 10^4 simulations, assuming $D(\boldsymbol{\theta}) = \sigma^2 \boldsymbol{I}_m$ (D_1). True covariance unstructured

Figure 2.7 shows the results when the true covariance is unstructured. The first column shows results for continuous while the second shows for categorical. The continuous variable has smaller variance for all cells compared to when there was no confound-

ing, and we see agreement between the aggregate and the estimate from (2.6.1). This decrease in variance is due to there being more variation and a larger composition from each cell compared to the original analysis (Tables 2.1 and 2.3). We see a decrease in the variance of the parameter associated with cell type 3, as the mean $\pi_{i,3}$ has increased. For X_2, which was deterministic of the $\boldsymbol{\pi}$ distribution, we no longer see agreement between the aggregate estimate and the cell type estimate from (2.6.1), as X_2 and $\boldsymbol{\pi}$ are no longer independent. Results when the true covariance are independent or exchangeable are in Supplementary Figures 4.25, 4.26 and Tables 4.11, 4.12. The variance of the parameters for X_2 have increased for cell type 1, and decreased for the other cells (Table 2.3). This is in comparison to Table 2.1. We also see that the estimate of the effect of X_2 in cell type 3 is slightly bias compared to when no confounding (Table 2.3 vs Table 2.1).

2.6.4 Application to Normative Aging Study

We applied our method to the Normative Aging Study to test for cell specific associations between DNAm and smoking status. After removing individuals with missing data and restricting to Non-Hispanic Whites, we were left with 618 participants. The mean age of our study was 72 with 26 current smokers, 418 former smokers, and 174 never smokers. Smoking status was modeled as a categorical variable with three levels: never, former, or current smoker, with never smokers as the reference group.

We had observed cell type composition available on Neutrophils (61.6%), Monocytes (8.7%), Lymphocytes (25.7%), Eosinophils (3.4%), and Basophils (0.6%). Due to the low number of Eosinophils and Basophils, we collapsed these cells with Neutrophils to denote Granulocytes. On average, Granulocytes accounted for 65.6% of the sample composition. Cell type composition was measured in lab from whole blood.

DNAm was collected using the Infinium Human Methylation450 BeadChip array. More information on the study design has been published previously (Panni et al., 2016). We examined the association between smoking status and DNAm at 49 probes found to be associated with smoking (Tsaprouni et al., 2014; Zeilinger et al., 2013). Additional technical covariates such as plate, position on plate, and position on chip were adjusted for prior to analysis via ComBat (Johnson et al., 2007). We transformed the Beta-values

to the M-value scale (Du et al., 2010), ran Combat, and then transformed back to the Beta-value scale on the whole 450K data and subsetted to our 49 probes of interest. We assumed a working covariance matrix of $\boldsymbol{D}(\boldsymbol{\theta}) = \sigma^2 \boldsymbol{I}_m$. Significance was assessed at the FDR value of 0.1 after adjusting for multiple testing (Benjamini and Hochberg, 1995).

Of the 49 probes, 46 came back as significant via the traditional aggregate approach, and 40 came back as significant via the overall test. This smaller number of significant probes is expected based on our simulation results. Comparing the aggregate estimates to the weighted averages via (2.6.1) and (2.6.1) we see excellant agreement (Figure 2.8).

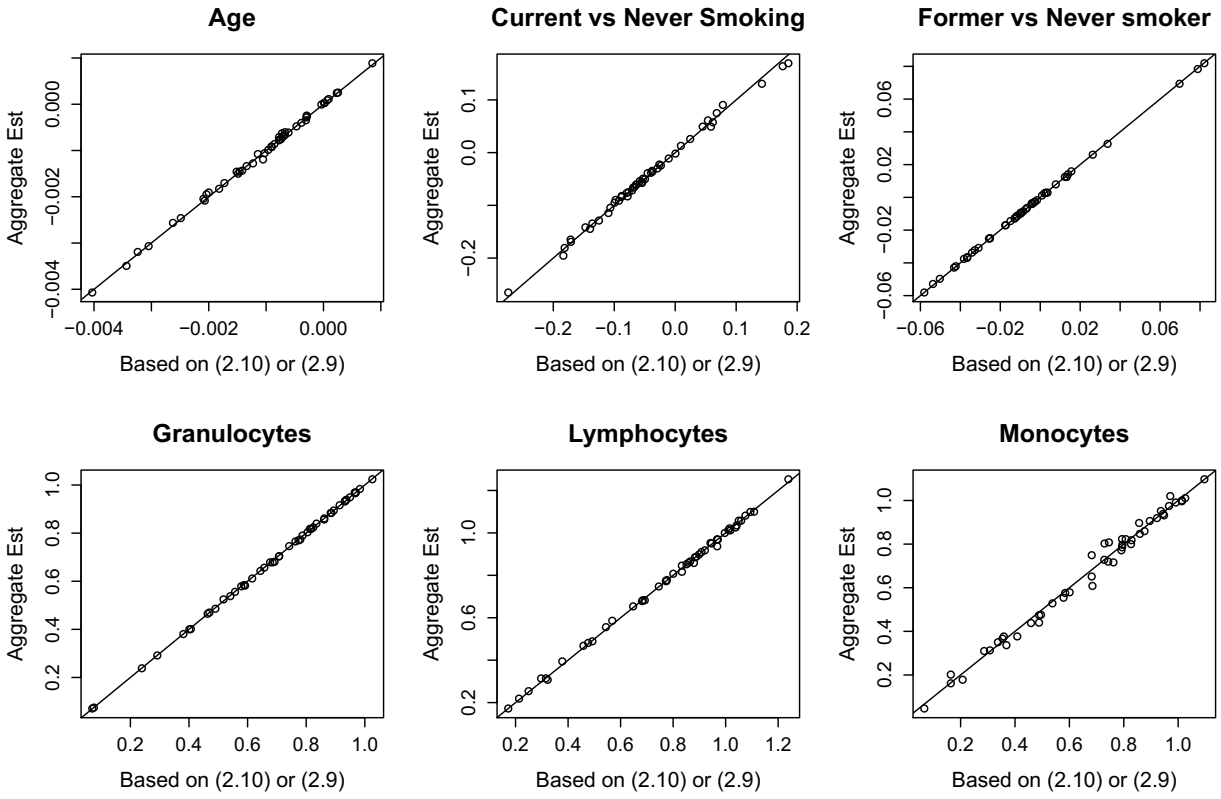

Figure 2.8: Comparing the estimates from aggregate analysis of 49 probes from (2.6.1) and (2.6.1). For all six parameters in the aggregate analysis

Testing for cell-specific effects, there were eight probes that had a significant association at the FDR level of 0.1. Seven of these probes were associated in Granulocytes, and one was associated in Lymphocytes (Table 2.4). Of the eight, they mapped to five genes: ALPPL2 (cg03329539, cg05951221, cg21566642), IER3 (cg06126421, cg14753356),

RARA (cg19572487), AHRR (cg26703534), and GNG12 (cg25189904). Two of these probes found in Granulocytes (cg21566642 and cg06126421), were reported in a paper published last year to have an association with smoking status in observed Granulocytes DNAm compared to other cells (Su et al., 2016). This paper also found an increase in methylation differences between smokers at Granulocytes for the AHRR probe cg05575921, but not for cg26703534. The other two probes at ALPPL2, are both close on the genome to cg21566642 and were also only associated in Granulocytes as well (Table 2.4). The other probe at IER3 (cg14753356) is also very close to cg06126421 and was only associated in Granulocytes. Interestingly, this paper also found an association at these probes with Monocytes and B-cells but not T-cells. B-cells and T-cells are a subset of Lymphocytes that we did not have observed information on. We therefore examined the Houseman estimated cell types, which provides estimates of Lymphocytes (Houseman et al., 2012). In the Lymphocytes, the mean estimated CD8T proportions were 2.29%, CD4T 13.24%, NK cells 6.94%, and B-cells 3.24%. Given that it appears the majority of the Lymphocytes in our sample are non B-cells and B-cells are such a small proportion, it is not surprising that we did not observe any association in the overall Lymphocytes for these probes.

Probe	Signal	Gene	CHR	Location	P-value
cg25189904	Lymphocytes	GNG12	1	68299493	0.001789
cg03329539	Granulocytes	ALPPL2	2	233283329	0.007266
cg05951221	Granulocytes	ALPPL2	2	233284402	0.005974
cg21566642	Granulocytes	ALPPL2	2	233284661	0.001371
cg26703534	Granulocytes	AHRR	5	377358	0.006065
cg06126421	Granulocytes	IER3	6	30720080	0.005748
cg14753356	Granulocytes	IER3	6	30720108	0.013869
cg19572487	Granulocytes	RARA	17	38476024	0.007617

Table 2.4: CpG probes found to have a significant cell specific association for smoking status in the Normative Aging Study.

The probes at RARA (cg19572487) and GNG12 (cg25189904), were both found to be associated with all-cause mortality in a paper published recenetly (Zhang et al., 2017). Methylation at RARA is associated with breast cancer and hepatocellular/thyroid carcinomas (Zhang et al., 2017). The gene GNG12 has been found to be associated with En-

dometrial cancer, but there has not been an examination of methylation at this gene and Endometrial cancer (Zhang et al., 2017). There has been a reported association between Lymphocyte count and Lymphocyte to Neutrophil ratio however being an indicator of Endometrial cancer survival (Cummings et al., 2015; de Jong et al., 2009). One probe at IER3 (cg06126421) was also found to be associated with all cause mortality (Zhang et al., 2017).

2.7 Conclusion and Discussion

In this paper, we proposed a novel approach to test for an association between unobserved cell specific DNAm and exposure when only cell type composition and aggregate DNAm is observed. We provided an equation relating the cell specific effects ($\boldsymbol{\beta}$) and the traditional aggregate estimates ($\boldsymbol{\gamma}$). We also extended the work by Sun (Sun et al. (2017)) by comparing our test statistic to a scaled F-distribution to account for the residual heterogeneity in our model.

There are several limitations to our real data analysis and our method. First, we have only a small number of current smokers in our sample. While this is a limitation, our ability to reproduce a number of previously established results is promising. Also, we did not address the possibility of cell type composition being associated with unobserved cell specific DNAm ($\boldsymbol{y_{i,j}} = \boldsymbol{\pi}_i\boldsymbol{\eta}_j + \boldsymbol{X}_i\boldsymbol{\beta}_j$). Further work on this area should be pursued. Also, we had the luxury of having observed cell composition, which most researchers do not. Incorporating an estimated cell compositions and how that affects the analysis should be explored.

With the exception of an analysis on heterogeneity in brain tissue DNA methylation (Montaño et al., 2013), we know of no other statistical methods on individual level data for detecting cell type specific associations. That previous method however was based on simply estimating differences between neurons and non-neurons and did not allow for multiple covariates. In our simulations we showed that we have unbiased estimates, proper type I error, and can allow for multiple covariates. The power to detect an association was correlated with the abundance of that cell type. In addition, a re-

cent method developed provides information on potential enrichment of signal in cells or tissue, but requires uploading aggregate results online which are then compared to an external reference set (Breeze et al. (2016)).

Here we note that the computational cost of bootstrapping to perform the testing procedure can be circumvented via coding in C++ and paralyzation. We also note that there is not a large amount of power to detect cell specific associations in the smallest cell type. Our approach could potentially be used as a secondary analysis after an initial aggregate EWAS captures a set of biologically relevant probes. This may miss probes where the cell specific effects are in opposite directions but it is not known how prevalent that occurs in the genome.

In summary, we have developed a model to estimate cell-specific DNAm associations using only whole-blood methylation and the cell type composition. Though not performed here, we believe this model could be extendable to other admixed samples from DNAm or even gene expression drawn from sources other than whole blood. All that is required is information about the mixing component for each individual. Our approach provides a potential way to circumvent the high price of collecting individual cell type DNAm or when the sample has been frozen and cell type can no longer be determined.

2.8 Appendix

2.8.1 Relation between Aggregate analysis and Cell Specific

The aggregate effects are a function of the cell specific effects. Let $\boldsymbol{\gamma}$ represent the terms from the aggregate model that fits $E(Y_i|\boldsymbol{X}_i, \boldsymbol{\pi}_i) = \boldsymbol{\pi}_i^T\boldsymbol{\gamma}_\pi + \boldsymbol{X}_i^T\boldsymbol{\gamma}_X = \boldsymbol{S}_i^T\boldsymbol{\gamma}$, with $\boldsymbol{S}_i^T = (\boldsymbol{\pi}_i^T, \boldsymbol{X}_i^T)$ and $\boldsymbol{\gamma}^T = (\boldsymbol{\gamma}_\pi^T, \boldsymbol{\gamma}_X^T)$. We reparametrize $\boldsymbol{X}_i$ to not include an intercept. Making our model for Y_i:

$$Y_i = \sum_{j=1}^{m}(\pi_{i,j}\beta_{0,j} + \pi_{i,j}\boldsymbol{X}_i^T\boldsymbol{\beta}_j) + \boldsymbol{\pi}_i^T\boldsymbol{\epsilon}_i = \boldsymbol{\pi}_i^T\boldsymbol{\beta}_0 + \widetilde{\boldsymbol{X}}_i\boldsymbol{\beta}_X + \boldsymbol{\pi}_i^T\boldsymbol{\epsilon}_i, \tag{2.13}$$

where $\boldsymbol{\beta}_0$ is all the cell specific intercepts and $\boldsymbol{\beta}_X$ are all the covariate specific cell effects for $\boldsymbol{X}$. Note that $\boldsymbol{\pi}_i^T$=$\boldsymbol{S}_i^T\boldsymbol{R}$, with $\boldsymbol{R} = (\boldsymbol{I}_m, \boldsymbol{0}_{p\times m}^T)^T$. $\boldsymbol{0}_{p\times m}$ is a p by m matrix of 0's. We can

rewrite (2.13) as:

$$Y_i = \boldsymbol{\pi}_i^T \boldsymbol{\beta}_0 + \widetilde{\boldsymbol{X}}_i^T \boldsymbol{\beta}_X + \boldsymbol{\pi}_i^T \epsilon_i = \boldsymbol{S}_i^T \boldsymbol{R} \boldsymbol{\beta}_0 + \widetilde{\boldsymbol{X}}_i^T \boldsymbol{\beta}_X + \boldsymbol{\pi}_i^T \epsilon_i = \begin{bmatrix} \boldsymbol{S}_i^T \boldsymbol{R} & \widetilde{\boldsymbol{X}}_i^T \end{bmatrix} \boldsymbol{\beta} + \boldsymbol{\pi}^T \boldsymbol{\epsilon}_i.$$

$\boldsymbol{\gamma}$ is a solution to the following score equation:

$$E\left\{\boldsymbol{S}_i(Y_i - \boldsymbol{S}_i^T \boldsymbol{\gamma})\right\} = 0,$$

therefore:

$$\begin{aligned} E\left(\boldsymbol{S}_i \begin{bmatrix} \boldsymbol{S}_i^T \boldsymbol{R} & \widetilde{\boldsymbol{X}}_i^T \end{bmatrix}\right) \boldsymbol{\beta} &= E\left(\boldsymbol{S}_i \boldsymbol{S}_i^T\right) \gamma, \\ \begin{bmatrix} E\left(\boldsymbol{S}_i \boldsymbol{S}_i^T\right) \boldsymbol{R} & E\left(\boldsymbol{S}_i \widetilde{\boldsymbol{X}}_i^T\right) \end{bmatrix} \boldsymbol{\beta} &= E\left(\boldsymbol{S}_i \boldsymbol{S}_i^T\right) \gamma, \\ \begin{bmatrix} \boldsymbol{R} & E\left(\boldsymbol{S}_i \boldsymbol{S}_i^T\right)^{-1} E\left(\boldsymbol{S}_i \widetilde{\boldsymbol{X}}_i^T\right) \end{bmatrix} \boldsymbol{\beta} &= \gamma. \end{aligned}$$

Therefore giving us:

$$\boldsymbol{\gamma} = \begin{bmatrix} \boldsymbol{I}_m \\ \boldsymbol{0}_{pxm} \end{bmatrix} \boldsymbol{\beta}_\pi + E\left(\boldsymbol{S}_i \boldsymbol{S}_i^T\right)^{-1} \begin{bmatrix} E\left(\boldsymbol{\pi}_i \widetilde{\boldsymbol{X}}_i^T\right) \\ E\left(\boldsymbol{X}_i \widetilde{\boldsymbol{X}}_i^T\right) \end{bmatrix} \boldsymbol{\beta}_X. \tag{2.14}$$

Now we assume that $\boldsymbol{\pi}$ and $\boldsymbol{X}$ are independent. Define $\boldsymbol{\mu}_X = E(\boldsymbol{X}_i)$ and $\boldsymbol{\mu}_\pi = E(\boldsymbol{\pi})$. Before we continue the proof note that:

$$\widetilde{\boldsymbol{X}}_i^T = \boldsymbol{\pi}_i^T \otimes \boldsymbol{X}_i^T = \boldsymbol{X}_i^T (\boldsymbol{\pi}_i^T \otimes \boldsymbol{I}_p) = \boldsymbol{\pi}_i^T (\boldsymbol{I}_m \otimes \boldsymbol{X}_i^T).$$

We now focus on the expectation left over in 2.14:

$$\begin{aligned} E\left(\boldsymbol{S}_i \boldsymbol{S}_i^T\right) &= \begin{bmatrix} E(\boldsymbol{\pi}_i \boldsymbol{\pi}_i^T) & \boldsymbol{\mu}_\pi \boldsymbol{\mu}_X^T \\ \boldsymbol{\mu}_X \boldsymbol{\mu}_\pi^T & E(\boldsymbol{X}_i \boldsymbol{X}_i^T) \end{bmatrix}, \\ \begin{bmatrix} E\left(\boldsymbol{\pi}_i \widetilde{\boldsymbol{X}}_i^T\right) \\ E\left(\boldsymbol{X}_i \widetilde{\boldsymbol{X}}_i^T\right) \end{bmatrix} &= \begin{bmatrix} E\left\{\boldsymbol{\pi}_i \boldsymbol{\pi}_i^T (\boldsymbol{I}_m \otimes \boldsymbol{X}_i^T)\right\} \\ E\left\{\boldsymbol{X}_i \boldsymbol{X}_i^T (\boldsymbol{\pi}_i^T \otimes \boldsymbol{I}_p)\right\} \end{bmatrix} = \begin{bmatrix} E(\boldsymbol{\pi}_i \boldsymbol{\pi}_i^T)(\boldsymbol{I}_m \otimes \boldsymbol{\mu}_X^T) \\ E(\boldsymbol{X}_i \boldsymbol{X}_i^T) E(\boldsymbol{\pi}_i^T \otimes \boldsymbol{I}_p) \end{bmatrix} \\ &= \begin{bmatrix} E(\boldsymbol{\pi}_i \boldsymbol{\pi}_i^T) & \boldsymbol{0}_{pxm} \\ \boldsymbol{0}_{mxp} & E(\boldsymbol{X}_i \boldsymbol{X}_i^T) \end{bmatrix} \begin{bmatrix} \boldsymbol{I}_m \otimes \boldsymbol{\mu}_X^T \\ \boldsymbol{\mu}_\pi^T \otimes \boldsymbol{I}_p \end{bmatrix}. \end{aligned}$$

We take the inverse of $E\left(\boldsymbol{S}_i \boldsymbol{S}_i^T\right)$:

$$E\left(\boldsymbol{S}_i \boldsymbol{S}_i^T\right)^{-1} = \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{bmatrix},$$

$$
\begin{aligned}
\boldsymbol{A}_{11} &= E(\boldsymbol{\pi}_i\boldsymbol{\pi}_i^T)^{-1} + \\
&\quad E(\boldsymbol{\pi}_i\boldsymbol{\pi}_i^T)^{-1}\boldsymbol{\mu}_\pi\boldsymbol{\mu}_X^T\left\{E(\boldsymbol{X}_i\boldsymbol{X}_i^T) - \boldsymbol{\mu}_X\boldsymbol{\mu}_\pi^T E(\boldsymbol{\pi}_i\boldsymbol{\pi}_i^T)^{-1}\boldsymbol{\mu}_\pi\boldsymbol{\mu}_X^T\right\}^{-1}\boldsymbol{\mu}_X\boldsymbol{\mu}_\pi^T E(\boldsymbol{\pi}_i\boldsymbol{\pi}_i^T)^{-1}, \\
\boldsymbol{A}_{12} &= -E(\boldsymbol{\pi}_i\boldsymbol{\pi}_i^T)^{-1}\boldsymbol{\mu}_\pi\boldsymbol{\mu}_X^T\left\{E(\boldsymbol{X}_i\boldsymbol{X}_i^T) - \boldsymbol{\mu}_X\boldsymbol{\mu}_\pi^T E(\boldsymbol{\pi}_i\boldsymbol{\pi}_i^T)^{-1}\boldsymbol{\mu}_\pi\boldsymbol{\mu}_X^T\right\}^{-1}, \\
\boldsymbol{A}_{21} &= -\left\{E(\boldsymbol{X}_i\boldsymbol{X}_i^T) - \boldsymbol{\mu}_X\boldsymbol{\mu}_\pi^T E(\boldsymbol{\pi}_i\boldsymbol{\pi}_i^T)^{-1}\boldsymbol{\mu}_\pi\boldsymbol{\mu}_X^T\right\}^{-1}\boldsymbol{\mu}_X\boldsymbol{\mu}_\pi^T E(\boldsymbol{\pi}_i\boldsymbol{\pi}_i^T)^{-1}, \\
\boldsymbol{A}_{22} &= \left\{E(\boldsymbol{X}_i\boldsymbol{X}_i^T) - \boldsymbol{\mu}_X\boldsymbol{\mu}_\pi^T E(\boldsymbol{\pi}_i\boldsymbol{\pi}_i^T)^{-1}\boldsymbol{\mu}_\pi\boldsymbol{\mu}_X^T\right\}^{-1}.
\end{aligned}
$$

$\boldsymbol{\pi}_i$ is a normalized vector (i.e. $\boldsymbol{\pi}_i^T\boldsymbol{J}_m = 1$, where $\boldsymbol{J}_m$ is a $m \times 1$ vector of 1's). Thus:

$$
\begin{aligned}
E(\boldsymbol{\pi}_i\boldsymbol{\pi}_i^T)\boldsymbol{J}_m = \boldsymbol{\mu}_\pi \rightarrow E(\boldsymbol{\pi}_i\boldsymbol{\pi}_i^T)^{-1}\boldsymbol{\mu}_\pi = \boldsymbol{J}_m, \\
\boldsymbol{\mu}_\pi^T\boldsymbol{J}_m = 1 \rightarrow \boldsymbol{\mu}_\pi^T E(\boldsymbol{\pi}_i\boldsymbol{\pi}_i^T)^{-1}\boldsymbol{\mu}_\pi = 1.
\end{aligned}
$$

We use these results and $E(\boldsymbol{X}_i\boldsymbol{X}_i^T) - \boldsymbol{\mu}_X\boldsymbol{\mu}_X^T = \boldsymbol{\Sigma}_X = \text{cov}(X_i)$ to simplify the inverse of $E\left(\boldsymbol{S}_i\boldsymbol{S}_i^T\right)$.

$$
\begin{aligned}
\boldsymbol{A}_{11} &= E(\boldsymbol{\pi}_i\boldsymbol{\pi}_i^T)^{-1} + \boldsymbol{J}_m\boldsymbol{\mu}_X^T\boldsymbol{\Sigma}_X^{-1}\boldsymbol{\mu}_X\boldsymbol{\mu}_\pi^T E(\boldsymbol{\pi}_i\boldsymbol{\pi}_i^T)^{-1}, \\
\boldsymbol{A}_{12} &= -\boldsymbol{J}_m\boldsymbol{\mu}_X^T\boldsymbol{\Sigma}_X^{-1}, \\
\boldsymbol{A}_{21} &= -\boldsymbol{\Sigma}_X^{-1}\boldsymbol{\mu}_X\boldsymbol{\mu}_\pi^T E(\boldsymbol{\pi}_i\boldsymbol{\pi}_i^T)^{-1}, \\
\boldsymbol{A}_{22} &= \boldsymbol{\Sigma}_X^{-1}.
\end{aligned}
$$

Going back, we use that $\boldsymbol{\mu}_\pi^T(\boldsymbol{I}_m \otimes \boldsymbol{\mu}_X^T) = \boldsymbol{\mu}_X^T(\boldsymbol{\mu}_\pi^T \otimes \boldsymbol{I}_p)$ and $\boldsymbol{\Sigma}_x = E(\boldsymbol{X}_i\boldsymbol{X}_i^T) - \boldsymbol{\mu}_X\boldsymbol{\mu}_X^T$:

$$
E\left(\boldsymbol{S}_i\boldsymbol{S}_i^T\right)^{-1}\begin{bmatrix} E\left(\boldsymbol{\pi}_i\widetilde{\boldsymbol{X}}_i^T\right) \\ E\left(\boldsymbol{X}_i\widetilde{\boldsymbol{X}}_i^T\right) \end{bmatrix} = \begin{bmatrix} (\boldsymbol{I}_m - \boldsymbol{J}_m\boldsymbol{\mu}_\pi^T)(\boldsymbol{I}_m \otimes \boldsymbol{\mu}_X^T) \\ \boldsymbol{\mu}_\pi^T \otimes \boldsymbol{I}_p \end{bmatrix}.
$$

This then gives us the final result that:

$$
\begin{aligned}
\boldsymbol{\gamma} &= \begin{bmatrix} \boldsymbol{I}_m \\ \boldsymbol{0}_{pxm} \end{bmatrix}\boldsymbol{\beta}_\pi + \begin{bmatrix} (\boldsymbol{I}_m - \boldsymbol{J}_m\boldsymbol{\mu}_\pi^T)(\boldsymbol{I}_m \otimes \boldsymbol{\mu}_X^T) \\ \boldsymbol{\mu}_\pi^T \otimes \boldsymbol{I}_p \end{bmatrix}\boldsymbol{\beta}_X, \\
\boldsymbol{\gamma}_\pi &= \boldsymbol{\beta}_\pi + (\boldsymbol{I}_m - \boldsymbol{J}_m\boldsymbol{\mu}_\pi^T)(\boldsymbol{I}_m \otimes \boldsymbol{\mu}_X^T)\boldsymbol{\beta}_X, \\
\boldsymbol{\gamma}_X &= (\boldsymbol{\mu}_\pi^T \otimes \boldsymbol{I}_p)\boldsymbol{\beta}_X.
\end{aligned}
$$

Assessing the genetic effect mediated through gene expression from summary eQTL and GWAS data

Richard Barfield

Department of Biostatistics

Harvard Graduate School of Arts and Sciences

Peter Kraft

Department of Biostatistics

Harvard Chan School of Public Health

Program in Genetic Epidemiology and Statistical Genetics

Harvard T.H. Chan School of Public Health

3.1 Introduction

Integrating genome-wide association study (GWAS) and expression quantitative trait loci (eQTL) data can help in detecting novel disease loci and pinpointing genes of interest. This is done by aggregating association signals across multiple SNPs associated with transcript levels or by establishing if the association between SNPs and disease are mediated through the expression of particular genes. However, there are often not sufficiently large amounts of data on the disease of interest, SNPs, and gene expression in the relevant tissue of interest all available on the same set of individuals. In contrast, there are large amounts of publicly available summary statistics from separate studies; and there has been a rise of statistical methods to utilize and jointly analyze these summary data (Barbeira et al., 2016; Gamazon et al., 2015; Gusev et al., 2016a; Zhu et al., 2016).

While the cited methods can identify association between genetically-predicted expression levels and disease, they cannot distinguish between mediation (the SNPs affect disease risk through their effects on a particular gene's expression) and co-localization (e.g. the eQTLs and causal disease SNPs are distinct but in linkage disequilibrium causing pleiotropy). This can potentially lead to spurious inference regarding the association between gene expression and disease. The Transcriptome Wide Association Study (TWAS), PrediXcan and MetaXcan test statistics for example, cannot distinguish between co-localization and mediation (Barbeira et al., 2016; Gamazon et al., 2015; Gusev et al., 2016a). This may lead to significant test statistics, even though the tested gene expression levels do not affect the outcome. We were motivated by this current problem to develop a potential correction for when there is a direct effect of SNPs on an outcome that is not through gene of interest expression.

As mentioned previously, the association from gene to outcome could be due to disease SNPs not being associated with gene expression but in linkage with the eQTL SNPs. To distinguish between true susceptibility genes (i.e. when the genetic effect on phenotype is mediated through expression) and spurious co-localizations, we developed LD-aware Mendelian Randomization Egger regression (LDA MR-Egger): an extension of MR-Egger regression (Bowden et al., 2015) to multiple SNPs in linkage disequilibrium.

LDA MR-Egger requires only summary GWAS, eQTL statistics, and LD information gathered from a reference panel. The current MR-Egger was motivated for an intervening variable that was a phenotypic trait as opposed to gene expression and was an extension of traditional summary Mendelian Randomization (MR) (Bowden et al., 2015). The SNPs of interest were picked from across the genome making LD not an issue. The MR has also been extended to incorporate correlated SNPs (LDA MR) (Burgess et al., 2016). We combine these two approaches for our LDA MR-Egger. This method can help in differentiating loci where there is co-localization of disease and eQTL SNPs.

In this paper, we first introduce the model relating the outcome and gene expression to the SNPs. We next discuss four existing approaches for testing and/or estimating the association between the outcome and the gene of interest, and introduce our new LDA MR-Egger regression. The statistical properties of these estimates and their performance are then examined in presence or absence of co-localization. We perform an empirical study to assess the type I error, power, and bias of the estimates. Finally, we apply the various approaches to summary statistics from a GWAS on Breast Cancer (Michailidou et al., 2017) with eQTL data from a breast tissue panel in GTEx (Lonsdale et al., 2013).

3.2 Methods

3.2.1 The Models

Let $\boldsymbol{Y}$ denote our outcome (nx1), $\boldsymbol{M}$ the mediator (nx1), and $\boldsymbol{G}$ the SNP matrix (nxJ) of interest. We assume that the columns of $\boldsymbol{G}$ have been standardized to have mean zero and variance one. If $\boldsymbol{G}$ has not been standardized and the GWAS effect estimates are on the minor allele counts, we can transform the effects to what would have been observed if the SNPs had been standardized (Appendix 3.10.1). We denote the LD structure of the SNPs $\boldsymbol{G}$ as $\boldsymbol{\Sigma}$, a JxJ symmetric positive definite matrix. For a link function g, we have the following models relating $\boldsymbol{M}$, $\boldsymbol{G}$, and $\boldsymbol{Y}$:

$$g(E(\boldsymbol{Y}|\boldsymbol{M},\boldsymbol{G})) = \gamma_0 + \boldsymbol{M}\gamma + \boldsymbol{G}\boldsymbol{\theta}, \tag{3.1}$$

$$g(E(\boldsymbol{Y}|\boldsymbol{G})) = \gamma_0^* + \boldsymbol{G}\boldsymbol{\beta}_G, \tag{3.2}$$

$$\boldsymbol{M} = \beta_0 + \boldsymbol{G}\boldsymbol{\beta}_E + \boldsymbol{\epsilon}_M; \boldsymbol{\epsilon}_M \sim N(\boldsymbol{0}, \boldsymbol{I}_n\sigma^2). \tag{3.3}$$

In the above models, $\boldsymbol{\theta}$ is a J-column vector of $\boldsymbol{G}$ effects on $\boldsymbol{Y}$ conditional on $\boldsymbol{M}$, γ is the effect of $\boldsymbol{M}$ on $\boldsymbol{Y}$ conditional on $\boldsymbol{G}$, $\boldsymbol{\beta}_E$ is the J-column vector of SNP effects on the mediator, and $\boldsymbol{\beta}_G$ is the J-vector of the $\boldsymbol{G}$ effects marginal over $\boldsymbol{M}$. $\boldsymbol{\beta}_G$ and $\boldsymbol{\beta}_E$ represent mutually-conditioned SNP effects on outcome and gene expression respectively. $\boldsymbol{\epsilon}_M$ represents the residual variance in $\boldsymbol{M}$ and $\boldsymbol{I}_n$ is the n x n identity matrix. We are interested in the situation where we cannot directly estimate the parameters in model (3.1), as we do not have complete data on $\boldsymbol{Y}$, $\boldsymbol{M}$ and $\boldsymbol{G}$ from (sufficiently many) individuals. We want to derive inference on γ, because if $\gamma \neq 0$ the gene affects the trait. To relate the parameters in the marginal model of $\boldsymbol{Y}$ and the marginal model of $\boldsymbol{M}$ we assume one of the following for the remainder of the paper:

- g is either the log or identity link function.
- $\boldsymbol{Y}$ is a sufficiently rare binary trait and g is the logit link.

If either of the two conditions above hold, we will have the following:

$$\boldsymbol{\beta}_G \approx \boldsymbol{\beta}_E\gamma + \boldsymbol{\theta} \tag{3.4}$$

If g is the log or linear link, the approximation will be exact. (3.4) suggests that the effects of $\boldsymbol{G}$ marginal over $\boldsymbol{M}$ (GWAS effects) are a function of the eQTL statistics ($\boldsymbol{\beta}_E$), the effect of the gene expression on the outcome conditional on $\boldsymbol{G}$ (γ), and the effect of $\boldsymbol{G}$ on the outcome conditional on gene expression ($\boldsymbol{\theta}$). We will call $\boldsymbol{\theta}$ the direct effect and $\boldsymbol{\beta}_E\gamma$ the mediated effect. If the necessary causal assumptions are met, these parameters could be interpreted in the causal/counterfactual framework, but for this article it is not necessary (Valeri and VanderWeele, 2013).

In practice, we do not have an overlap of data on $\boldsymbol{M}$, $\boldsymbol{Y}$, and $\boldsymbol{G}$. Instead, we have a sample of size N that the GWAS was run on to estimate $\boldsymbol{\beta}_G$ and an independent sample of size N_E that was used to estimate $\boldsymbol{\beta}_E$. We therefore cannot estimate γ directly. Moreover, we typically only have estimates of individual SNP effects marginal over the other SNPs, $\hat{\boldsymbol{\beta}}^*_E$ and $\hat{\boldsymbol{\beta}}^*_G$. For our purposes, $\hat{\beta}^*_{G,j}$ and $\hat{\beta}^*_{E,j}$ were estimated with the same reference allele

for SNP j. If they were not, the sign of the effect can be changed so as to refer to the same reference allele.

The formulas given above relating the mean of $\boldsymbol{Y}$ and $\boldsymbol{M}$ to $\boldsymbol{G}$ were given on the conditional level. We therefore need to transform our marginal estimates ($\hat{\boldsymbol{\beta}}_E^*$ and $\hat{\boldsymbol{\beta}}_G^*$) to the conditional scale. Given an estimate of the LD matrix ($\boldsymbol{\Sigma}$), we estimate the conditional eQTL and GWAS effects as $\hat{\boldsymbol{\beta}}_E = \boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\beta}}_E^*$ and $\hat{\boldsymbol{\beta}}_G = \boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\beta}}_G^*$ (Pasaniuc et al., 2014). If the marginal effect estimates were not calculated on the standardized genotypes they can be transformed (Appendix 3.10.1). We assume that $\hat{\boldsymbol{\beta}}_E$ and $\hat{\boldsymbol{\beta}}_G$ are unbiased for $\boldsymbol{\beta}_E$ and $\boldsymbol{\beta}_G$ respectively.

We do not know the form of $\boldsymbol{\theta}$; it could be a constant or vary by SNP. All we know is that it is a vector of length J. Our estimated GWAS effects, (given our assumptions above) are a function of $\boldsymbol{\theta}$, $\boldsymbol{\beta}_E$, γ and sampling error:

$$\hat{\boldsymbol{\beta}}_G \approx \boldsymbol{\beta}_E\gamma + \boldsymbol{\theta} + \boldsymbol{\epsilon}_G; \boldsymbol{\epsilon}_G \sim N(\mathbf{0}, \boldsymbol{\Sigma}_G).$$

If the SNPs are not in LD, then the marginal and the conditional will be equal ($\hat{\boldsymbol{\beta}}_G \approx \hat{\boldsymbol{\beta}}_G^*$).

We next derive the covariance of $\hat{\boldsymbol{\beta}}_G$ ($\boldsymbol{\Sigma}_G$). Let $\boldsymbol{\sigma}_G^*$ denote a Jx1 vector of the marginal standard errors of $\hat{\boldsymbol{\beta}}_E^*$. Let "$\cdot$" denote element wise multiplication between two matrices. As $\boldsymbol{G}$ has been standardized, $\text{cov}(\hat{\boldsymbol{\beta}}_G^*) = \boldsymbol{\Sigma} \cdot \boldsymbol{\sigma}_G^*\boldsymbol{\sigma}_G^{*T}$. This gives the covariance of our conditional GWAS estimates:

$$\boldsymbol{\Sigma}_G = \text{cov}(\hat{\boldsymbol{\beta}}_G) = \text{cov}(\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\beta}}_G^*) = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma} \cdot \boldsymbol{\sigma}_G^*\boldsymbol{\sigma}_G^{*T})\boldsymbol{\Sigma}^{-1}.$$

If $\boldsymbol{\sigma}_G^* = v\mathbf{1}_J$, where $\mathbf{1}_J$ is a column vector of ones, then $\boldsymbol{\Sigma}_G = v^2\boldsymbol{\Sigma}^{-1}$. Our goal is to derive a valid test and also estimate γ in the presence of the SNPs having a direct effect on the outcome. We next go over several approaches for testing for an association between gene expression and outcome.

3.3 Methods for Testing and Estimation

3.3.1 Transcriptome Wide Association Studies (TWAS)

The Transcriptome Wide Association Study (TWAS) statistic uses summary statistics to test for an association between gene expression and a phenotype of interest (Gusev et al.,

2016a). The TWAS does not necessarily estimate the γ above but does provide a valid test for the association between the gene of interest and the outcome. The TWAS test statistic is:

$$Z_{TWAS} = \frac{\hat{\boldsymbol{\beta}}_E^{*T} \boldsymbol{\Sigma}^{-1} \boldsymbol{Z}_G^*}{\sqrt{\hat{\boldsymbol{\beta}}_E^{*T} \boldsymbol{\Sigma}^{-1} \hat{\boldsymbol{\beta}}_E^*}},$$

where $\boldsymbol{Z}_G^* = \hat{\boldsymbol{\beta}}_G^* \cdot \boldsymbol{\sigma}_G^{*-1}$ is a column vector of the marginal test statistics. The test statistic is then compared to a standard normal to assess significance. The TWAS can use either the marginal or conditional eQTL estimates as weights. If using the conditional, substitute $\hat{\boldsymbol{\beta}}_E$ for $\hat{\boldsymbol{\beta}}_E^*$ in the equation above.

3.3.2 "Toby Johnson" or MR Estimator

The "Toby Johnson" estimator of γ is the summary Mendelian Randomization estimate and the basis for all estimates considered for the rest of the paper (Johnson, 2012). For it to be an unbiased estimate of γ, it requires that there be no direct effects, i.e. $\boldsymbol{\theta} = \mathbf{0}$, or the direct effects be orthogonal to a transformation of $\hat{\boldsymbol{\beta}}_E^*$. It also estimates γ assuming that the SNPs are independent ($\boldsymbol{\Sigma} = \boldsymbol{I}_J$), therefore using the marginal estimates of the eQTL and GWAS effects. The Toby Johnson estimate (from here on the MR estimate):

$$\hat{\gamma}_{MR} = \frac{\hat{\boldsymbol{\beta}}_E^{*T} \boldsymbol{V}^{-1} \hat{\boldsymbol{\beta}}_G^*}{\hat{\boldsymbol{\beta}}_E^{*T} \boldsymbol{V}^{-1} \hat{\boldsymbol{\beta}}_E^*},$$

where $\boldsymbol{V}$ is a diagonal matrix with $v_{jj} = \sigma_{G,j}^{*2} = var(\hat{\beta}_{G,j}^*)$. The estimate can be rewritten as:

$$\hat{\gamma}_{MR} = \frac{\sum_{j=1}^{J} \hat{\beta}_{E,j}^* \hat{\beta}_{G,j}^* \sigma_{G,j}^{-2*}}{\sum_{j=1}^{J} \hat{\beta}_{E,j}^{2*} \sigma_{G,j}^{-2*}},$$

and we estimate the variance as:

$$\text{var}(\hat{\gamma}_{MR}) = \frac{\hat{\sigma}_{MR}^2}{\sum_{j=1}^{J} \left(\hat{\beta}_{E,j}^{2*} \sigma_{G,j}^{-2*} \right)},$$

$$\hat{\sigma}_{MR}^2 = \frac{1}{J-1} \sum_{j=1}^{J} \sigma_{G,j}^{-2*} (\hat{\beta}_{G,j}^* - \hat{\beta}_{E,j}^* \hat{\gamma}_{MR})^2.$$

The test statistic is then:

$$Z_{MR} = \frac{\hat{\gamma}_{MR}}{\sqrt{\text{var}(\hat{\gamma}_{MR})}} = \frac{\sum_{j=1}^{J} \hat{\beta}^*_{E,j}\hat{\beta}^*_{G,j}\sigma^{-2*}_{G,j}}{\sigma_{MR}\sqrt{\sum_{j=1}^{J} \hat{\beta}^{2*}_{E,j}\sigma^{-2*}_{G,j}}}.$$

For testing, we compare Z_{MR} to the quantiles of a t-distribution with $J-1$ degrees of freedom. If the SNPs are in LD or if there are direct effects, $\hat{\gamma}_{MR}$ can lead to incorrect inference. The MR estimate can be framed as a weighted linear regression without an intercept of the marginal GWAS estimates on the marginal eQTL estimates with weights equal to $\sigma^{-2*}_{G,j}$.

3.3.3 MR-Egger Estimate

If the MR estimate can be thought of as a weighted linear regression without an intercept, the MR-Egger extends the MR by including an intercept (λ) to the weighted linear regression. It assumes that $E(\hat{\boldsymbol{\beta}}^*_G|\hat{\boldsymbol{\beta}}^*_E) = \lambda \mathbf{1}_J + \hat{\boldsymbol{\beta}}^*_E\gamma$. The idea is that λ will account for some of the direct effect. The estimates are (same $\boldsymbol{V}$ above):

$$\begin{bmatrix} \hat{\lambda} \\ \hat{\gamma}_{MRE} \end{bmatrix} = \begin{bmatrix} \mathbf{1}_J^T\boldsymbol{V}^{-1}\mathbf{1}_J & \mathbf{1}_J^T\boldsymbol{V}^{-1}\hat{\boldsymbol{\beta}}^*_E \\ \hat{\boldsymbol{\beta}}^{*T}_E\boldsymbol{V}^{-1}\mathbf{1}_J & \hat{\boldsymbol{\beta}}^{*T}_E\boldsymbol{V}^{-1}\hat{\boldsymbol{\beta}}^*_E \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{1}_J^T\boldsymbol{V}^{-1} \\ \hat{\boldsymbol{\beta}}^{*T}_E\boldsymbol{V}^{-1} \end{bmatrix} \hat{\boldsymbol{\beta}}^*_G,$$

the estimate of γ is:

$$\hat{\gamma}_{MRE} = \frac{(\sum_{j=1}^{J}\sigma^{-2*}_{G,j})(\sum_{j=1}^{J}\hat{\beta}^*_{G,j}\hat{\beta}^*_{E,j}\sigma^{-2*}_{G,j}) - (\sum_{j=1}^{J}\hat{\beta}^*_{E,j}\sigma^{-2*}_{G,j})(\sum_{j=1}^{J}\hat{\beta}^*_{G,j}\sigma^{-2*}_{G,j})}{(\sum_{j=1}^{J}\sigma^{-2*}_{G,j})(\sum_{j=1}^{J}\sigma^{-2*}_{G,j}\hat{\beta}^{2*}_{E,j}) - (\sum_{j=1}^{J}\hat{\beta}^*_{E,j}\sigma^{-2*}_{G,j})^2}.$$

The variance of the estimate is:

$$var(\hat{\gamma}_{MRE}) = \sigma^2_{MRE}\frac{\sum_{j=1}^{J}\sigma^{-2*}_{G,j}}{(\sum_{j=1}^{J}\sigma^{-2*}_{G,j})(\sum_{j=1}^{J}\sigma^{-2*}_{G,j}\hat{\beta}^{2*}_{E,j}) - (\sum_{j=1}^{J}\hat{\beta}^*_{E,j}\sigma^{-2*}_{G,j})^2},$$

$$\hat{\sigma}^2_{MRE} = \frac{1}{J-2}\sum_{j=1}^{J}\sigma^{-2*}_{G,j}(\hat{\beta}^*_{G,j} - \hat{\lambda} - \hat{\beta}^*_{E,j}\hat{\gamma}_{MRE})^2.$$

The test statistic is:

$$Z_{MRE} = \frac{(\sum_{j=1}^{J}\sigma^{-2*}_{G,j})(\sum_{j=1}^{J}\hat{\beta}^*_{G,j}\hat{\beta}^*_{E,j}\sigma^{-2*}_{G,j}) - (\sum_{j=1}^{J}\hat{\beta}^*_{E,j}\sigma^{-2*}_{G,j})(\sum_{j=1}^{J}\hat{\beta}^*_{G,j}\sigma^{-2*}_{G,j})}{\hat{\sigma}_{MRE}\sqrt{\sum_{j=1}^{J}\sigma^{-2*}_{G,j}}\sqrt{(\sum_{j=1}^{J}\sigma^{-2*}_{G,j})(\sum_{j=1}^{J}\sigma^{-2*}_{G,j}\hat{\beta}^{2*}_{E}) - (\sum_{j=1}^{J}\hat{\beta}^*_{E,j}\sigma^{-2*}_{G,j})^2}}.$$

To test, we compare to the quantiles of a t-distribution with $J-2$ degrees of freedom. If the SNPs are in LD, the test for $\hat{\gamma}_{MRE}$ may lead to incorrect inference due to the

variance of $\hat{\gamma}_{MRE}$ being misspecified. If there are direct effects, and they do not vary from SNP to SNP, and are independent of the eQTL effects, this test statistic will properly take into account the direct effect and properly adjust for potential co-localization of disease SNPs with eQTL SNPs (Instrument Strength Independent of Direct Effect or InSIDE condition) (Bowden et al., 2015).

3.3.4 LD Aware MR (LDA MR)

The LDA MR estimate of γ is an extension of the MR by relaxing the assumption of the SNPs being independent. It still requires that there are no direct effects. Recall that $\boldsymbol{\Sigma}_G = \text{cov}(\hat{\boldsymbol{\beta}}_G)$, the LDA MR estimator is then:

$$\hat{\gamma}_{LDMR} = \frac{\hat{\boldsymbol{\beta}}_E^T \boldsymbol{\Sigma}_G^{-1} \hat{\boldsymbol{\beta}}_G}{\hat{\boldsymbol{\beta}}_E^T \boldsymbol{\Sigma}_G^{-1} \hat{\boldsymbol{\beta}}_E},$$

and we estimate the variance as:

$$\text{var}(\hat{\gamma}_{LDMR}) = \frac{\hat{\sigma}^2_{LDMR}}{\hat{\boldsymbol{\beta}}_E^T \boldsymbol{\Sigma}_G^{-1} \hat{\boldsymbol{\beta}}_E},$$
$$\hat{\sigma}^2_{LDMR} = \frac{1}{J-1}(\hat{\boldsymbol{\beta}}_G - \hat{\boldsymbol{\beta}}_E \hat{\gamma}_{LDMR})^T \boldsymbol{\Sigma}_G^{-1} (\hat{\boldsymbol{\beta}}_G - \hat{\boldsymbol{\beta}}_E \hat{\gamma}_{LDMR}).$$

The test statistic is then:

$$Z_{LDMR} = \frac{\hat{\gamma}_{LDA-MR}}{\sqrt{var(\hat{\gamma})_{LDMR}}} = \frac{\hat{\boldsymbol{\beta}}_E^T \boldsymbol{\Sigma}_G^{-1} \hat{\boldsymbol{\beta}}_G}{\hat{\sigma}_{LDMR} \sqrt{\hat{\boldsymbol{\beta}}_E^T \boldsymbol{\Sigma}_G^{-1} \hat{\boldsymbol{\beta}}_E}}.$$

Similar to the MR, we compare to a t-distribution with $J-1$ df. If there are direct effects, $\hat{\gamma}_{LDMR}$ can lead to incorrect inference. Just as the MR estimate is a weighted linear regression without an intercept, the LDA MR estimate is a weighted linear regression without an intercept using the weight matrix $\boldsymbol{\Sigma}_G^{-1}$. If $\sigma^*_{G,j} = v$ for all SNPs, the LDA MR and the TWAS test statistic on the marginal eQTL will be proportional to each other by $\hat{\sigma}_{LDMR}$. The proof is provided in Appendix 3.10.2.

3.3.5 LDA MR-Egger

The LDA MR-Egger extends the MR-Egger by incorporating the LD structure of the SNPs. The estimate is:

$$\begin{bmatrix} \hat{\lambda}_{LD} \\ \hat{\gamma}_{LDMRE} \end{bmatrix} = \begin{bmatrix} \mathbf{1}_J^T \boldsymbol{\Sigma}_G^{-1} \mathbf{1}_J & \mathbf{1}_J^T \boldsymbol{\Sigma}_G^{-1} \hat{\boldsymbol{\beta}}_E \\ \hat{\boldsymbol{\beta}}_E^T \boldsymbol{\Sigma}_G^{-1} \mathbf{1}_J & \hat{\boldsymbol{\beta}}_E^T \boldsymbol{\Sigma}_G^{-1} \hat{\boldsymbol{\beta}}_E \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{1}_J^T \boldsymbol{\Sigma}_G^{-1} \\ \hat{\boldsymbol{\beta}}_E^{*T} \boldsymbol{\Sigma}_G^{-1} \end{bmatrix} \hat{\boldsymbol{\beta}}_G,$$

and our estimate of γ:

$$\hat{\gamma}_{LDMRE} = \frac{(\mathbf{1}_J^T \boldsymbol{\Sigma}_G^{-1} \mathbf{1}_J)(\hat{\boldsymbol{\beta}}_E^T \boldsymbol{\Sigma}_G^{-1} \hat{\boldsymbol{\beta}}_G) - (\mathbf{1}_J^T \boldsymbol{\Sigma}_G^{-1} \hat{\boldsymbol{\beta}}_E)(\mathbf{1}_J^T \boldsymbol{\Sigma}_G^{-1} \hat{\boldsymbol{\beta}}_G)}{(\mathbf{1}_J^T \boldsymbol{\Sigma}_G^{-1} \mathbf{1}_J)(\hat{\boldsymbol{\beta}}_E^T \boldsymbol{\Sigma}_G^{-1} \hat{\boldsymbol{\beta}}_E) - (\mathbf{1}_J^T \boldsymbol{\Sigma}_G^{-1} \hat{\boldsymbol{\beta}}_E)^2}.$$

The variance is estimated as:

$$var(\hat{\gamma}_{LDMRE}) = \hat{\sigma}^2_{LDMRE} \frac{\mathbf{1}_J^T \boldsymbol{\Sigma}_G^{-1} \mathbf{1}_J}{(\mathbf{1}_J^T \boldsymbol{\Sigma}_G^{-1} \mathbf{1}_J)(\hat{\boldsymbol{\beta}}_E^T \boldsymbol{\Sigma}_G^{-1} \hat{\boldsymbol{\beta}}_E) - (\mathbf{1}_J^T \boldsymbol{\Sigma}_G^{-1} \hat{\boldsymbol{\beta}}_E)^2},$$

$$\sigma^2_{LDMRE} = \frac{1}{J-2} (\hat{\boldsymbol{\beta}}_G - \hat{\lambda}_{LD} - \hat{\boldsymbol{\beta}}_E \hat{\gamma}_{LDMRE})^T \boldsymbol{\Sigma}_G^{-1} (\hat{\boldsymbol{\beta}}_G - \hat{\lambda}_{LD} - \hat{\boldsymbol{\beta}}_E \hat{\gamma}_{LDMRE}).$$

The test statistic is then:

$$\begin{aligned} Z_{LDMRE} &= \frac{\hat{\gamma}_{LDMRE}}{\sqrt{var(\hat{\gamma}_{LDMRE})}} \\ &= \frac{(\mathbf{1}_J^T \boldsymbol{\Sigma}_G^{-1} \mathbf{1}_J)(\hat{\boldsymbol{\beta}}_E^T \boldsymbol{\Sigma}_G^{-1} \hat{\boldsymbol{\beta}}_G) - (\mathbf{1}_J^T \boldsymbol{\Sigma}_G^{-1} \hat{\boldsymbol{\beta}}_E)(\mathbf{1}_J^T \boldsymbol{\Sigma}_G^{-1} \hat{\boldsymbol{\beta}}_G)}{\hat{\sigma}_{LDMRE} \sqrt{\mathbf{1}_J^T \boldsymbol{\Sigma}_G^{-1} \mathbf{1}_J} \sqrt{(\mathbf{1}_J^T \boldsymbol{\Sigma}_G^{-1} \mathbf{1}_J)(\hat{\boldsymbol{\beta}}_E^T \boldsymbol{\Sigma}_G^{-1} \hat{\boldsymbol{\beta}}_E) - (\mathbf{1}_J^T \boldsymbol{\Sigma}_G^{-1} \hat{\boldsymbol{\beta}}_E)^2}}. \end{aligned}$$

For testing, we compare to a t-distribution with $J-2$ degrees of freedom. This estimate is a combination of the approaches from LDA MR and the MR-Egger. The LDA MR-Egger will provide valid inference in the same scenarios as the MR-Egger, but in addition when the SNPs are in LD. It will still not provide valid inference when the direct effects are variable, or if the direct effect is not independent of the SNP to gene expression effect. Details of all tests are provided in Table 3.1.

3.4 Bias of aforementioned methods

We first focus on the estimates that do not incorporate an intercept, the MR and the LDA MR. First the MR estimate:

$$\hat{\gamma}_{MR} = \frac{\hat{\boldsymbol{\beta}}_E^{*T} \boldsymbol{V}^{-1} \hat{\boldsymbol{\beta}}_G^*}{\hat{\boldsymbol{\beta}}_E^{*T} \boldsymbol{V}^{-1} \hat{\boldsymbol{\beta}}_E^*}.$$

Method	Adjusts for Direct Effect	Accounts for LD	Distribution
TWAS	No	Yes	$Z_{TWAS} \sim N(0,1)$
MR	No	No	$Z_{MR} \sim t(df = J-1)$
MR-Egger	Yes	No	$Z_{MRE} \sim t(df = J-2)$
LDA MR	No	Yes	$Z_{LDMR} \sim t(df = J-1)$
LDA MR-Egger	Yes	Yes	$Z_{LDMRE} \sim t(df = J-2)$

Table 3.1: Overview of the five tests considered in this paper.

Note that $\hat{\boldsymbol{\beta}}_G^* = \boldsymbol{\Sigma}\hat{\boldsymbol{\beta}}_G$ and $\hat{\boldsymbol{\beta}}_E^* = \boldsymbol{\Sigma}\hat{\boldsymbol{\beta}}_E$. Then $E(\hat{\boldsymbol{\beta}}_G^*) = E(\boldsymbol{\Sigma}\hat{\boldsymbol{\beta}}_G) = \boldsymbol{\Sigma}\boldsymbol{\beta}_E\gamma + \boldsymbol{\Sigma}\boldsymbol{\theta}$. We assume that $\hat{\boldsymbol{\beta}}_E$ and $\hat{\boldsymbol{\beta}}_G$ were estimated from different samples and thus are independent, $E(\hat{\boldsymbol{\beta}}_G|\hat{\boldsymbol{\beta}}_E) = E(\hat{\boldsymbol{\beta}}_G)$. Using this gives us:

$$E(\hat{\gamma}_{MR}|\hat{\boldsymbol{\beta}}_E^*) = \gamma\frac{\hat{\boldsymbol{\beta}}_E^{*T}\boldsymbol{V}^{-1}\boldsymbol{\Sigma}\boldsymbol{\beta}_E}{\hat{\boldsymbol{\beta}}_E^{*T}\boldsymbol{V}^{-1}\hat{\boldsymbol{\beta}}_E^*} + \frac{\hat{\boldsymbol{\beta}}_E^{*T}\boldsymbol{V}^{-1}\boldsymbol{\Sigma}\boldsymbol{\theta}}{\hat{\boldsymbol{\beta}}_E^{*T}\boldsymbol{V}^{-1}\hat{\boldsymbol{\beta}}_E^*}.$$

Then use that $\hat{\boldsymbol{\beta}}_E^* = \boldsymbol{\Sigma}\hat{\boldsymbol{\beta}}_E$:

$$E(\hat{\gamma}_{MR}|\hat{\boldsymbol{\beta}}_E^*) = \gamma\frac{\hat{\boldsymbol{\beta}}_E^{*T}\boldsymbol{V}^{-1}\boldsymbol{\Sigma}\boldsymbol{\beta}_E}{\hat{\boldsymbol{\beta}}_E^{*T}\boldsymbol{V}^{-1}\boldsymbol{\Sigma}\hat{\boldsymbol{\beta}}_E} + \frac{\hat{\boldsymbol{\beta}}_E^{*T}\boldsymbol{V}^{-1}\boldsymbol{\Sigma}\boldsymbol{\theta}}{\hat{\boldsymbol{\beta}}_E^{*T}\boldsymbol{V}^{-1}\boldsymbol{\Sigma}\hat{\boldsymbol{\beta}}_E}.$$

The term $\hat{\boldsymbol{\beta}}_E^*$ was estimated from a sample size of N_E. As $N_E \to \infty$, we have that $\hat{\boldsymbol{\beta}}_E^* \to \boldsymbol{\Sigma}\boldsymbol{\beta}_E$ and the first term goes to γ. Assuming N_E is sufficiently large:

$$E(\hat{\gamma}_{MR}|\hat{\boldsymbol{\beta}}_E^*) \approx \gamma + \frac{\hat{\boldsymbol{\beta}}_E^{*T}\boldsymbol{V}^{-1}\boldsymbol{\Sigma}\boldsymbol{\theta}}{\hat{\boldsymbol{\beta}}_E^{*T}\boldsymbol{V}^{-1}\boldsymbol{\Sigma}\hat{\boldsymbol{\beta}}_E}.$$

Now, unless $\boldsymbol{\theta} = \mathbf{0}$ or the transformation $\hat{\boldsymbol{\beta}}_E^{*T}\boldsymbol{V}^{-1}\boldsymbol{\Sigma}$ is orthogonal to $\boldsymbol{\theta}$, the MR will be biased. Next we assess the LDA MR:

$$\hat{\gamma}_{LDMR} = \frac{\hat{\boldsymbol{\beta}}_E^T\boldsymbol{\Sigma}_G^{-1}\hat{\boldsymbol{\beta}}_G}{\hat{\boldsymbol{\beta}}_E^T\boldsymbol{\Sigma}_G^{-1}\hat{\boldsymbol{\beta}}_E},$$

$$E(\hat{\gamma}_{LDMR}|\hat{\boldsymbol{\beta}}_E) = \gamma\frac{\hat{\boldsymbol{\beta}}_E^T\boldsymbol{\Sigma}_G^{-1}\boldsymbol{\beta}_E}{\hat{\boldsymbol{\beta}}_E^T\boldsymbol{\Sigma}_G^{-1}\hat{\boldsymbol{\beta}}_E} + \frac{\hat{\boldsymbol{\beta}}_E^T\boldsymbol{\Sigma}_G^{-1}\boldsymbol{\theta}}{\hat{\boldsymbol{\beta}}_E^T\boldsymbol{\Sigma}_G^{-1}\hat{\boldsymbol{\beta}}_E}.$$

As $N_E \to \infty$, $\hat{\boldsymbol{\beta}}_E \to \boldsymbol{\beta}_E$ and the first term will go to γ. Assuming that N_E is sufficiently large, we have:

$$E(\hat{\gamma}_{LDMR}|\hat{\boldsymbol{\beta}}_E^*) \approx \gamma + \frac{\hat{\boldsymbol{\beta}}_E^T\boldsymbol{\Sigma}_G^{-1}\boldsymbol{\theta}}{\hat{\boldsymbol{\beta}}_E^T\boldsymbol{\Sigma}_G^{-1}\hat{\boldsymbol{\beta}}_E}.$$

If $\boldsymbol{\theta} \neq \mathbf{0}$ or if the transformation $\hat{\boldsymbol{\beta}}_E^T \boldsymbol{\Sigma}_G^{-1}$ is not orthogonal to $\boldsymbol{\theta}$, then the estimate will be biased. As J increases, the N_E needed for $\hat{\boldsymbol{\beta}}_E \to \boldsymbol{\beta}_E$ will also increase. We next look at the MR-Egger and again use that $\hat{\boldsymbol{\beta}}_G^* = \boldsymbol{\Sigma}\hat{\boldsymbol{\beta}}_G$ and $\hat{\boldsymbol{\beta}}_E^* = \boldsymbol{\Sigma}\hat{\boldsymbol{\beta}}_E$.

$$\begin{aligned}
\hat{\gamma}_{MRE} &= \frac{(\mathbf{1}_J^T \boldsymbol{V}^{-1} \mathbf{1}_J)(\hat{\boldsymbol{\beta}}_E^{*T} \boldsymbol{V}^{-1} \hat{\boldsymbol{\beta}}_G^*) - (\mathbf{1}_J^T \boldsymbol{V}^{-1} \hat{\boldsymbol{\beta}}_E^*)(\mathbf{1}_J^T \boldsymbol{V}^{-1} \hat{\boldsymbol{\beta}}_G^*)}{(\mathbf{1}_J^T \boldsymbol{V}^{-1} \mathbf{1}_J)(\hat{\boldsymbol{\beta}}_E^{*T} \boldsymbol{V}^{-1} \hat{\boldsymbol{\beta}}_E^*) - (\mathbf{1}_J^T \boldsymbol{V}^{-1} \hat{\boldsymbol{\beta}}_E^*)^2} \\
&= \frac{(\mathbf{1}_J^T \boldsymbol{V}^{-1} \mathbf{1}_J)(\hat{\boldsymbol{\beta}}_E^{*T} \boldsymbol{V}^{-1} \boldsymbol{\Sigma}\hat{\boldsymbol{\beta}}_G) - (\mathbf{1}_J^T \boldsymbol{V}^{-1} \boldsymbol{\Sigma}\hat{\boldsymbol{\beta}}_E)(\mathbf{1}_J^T \boldsymbol{V}^{-1} \boldsymbol{\Sigma}\hat{\boldsymbol{\beta}}_G)}{(\mathbf{1}_J^T \boldsymbol{V}^{-1} \mathbf{1}_J)(\hat{\boldsymbol{\beta}}_E^{*T} \boldsymbol{V}^{-1} \boldsymbol{\Sigma}\hat{\boldsymbol{\beta}}_E) - (\mathbf{1}_J^T \boldsymbol{V}^{-1} \boldsymbol{\Sigma}\hat{\boldsymbol{\beta}}_E)^2}, \\
E(\hat{\gamma}_{MRE} | \hat{\boldsymbol{\beta}}_E^*) &= \gamma \frac{(\mathbf{1}_J^T \boldsymbol{V}^{-1} \mathbf{1}_J)(\hat{\boldsymbol{\beta}}_E^{*T} \boldsymbol{V}^{-1} \boldsymbol{\Sigma}\boldsymbol{\beta}_E) - (\mathbf{1}_J^T \boldsymbol{V}^{-1} \boldsymbol{\Sigma}\hat{\boldsymbol{\beta}}_E)(\mathbf{1}_J^T \boldsymbol{V}^{-1} \boldsymbol{\Sigma}\boldsymbol{\beta}_E)}{(\mathbf{1}_J^T \boldsymbol{V}^{-1} \mathbf{1}_J)(\hat{\boldsymbol{\beta}}_E^{*T} \boldsymbol{V}^{-1} \boldsymbol{\Sigma}\hat{\boldsymbol{\beta}}_E) - (\mathbf{1}_J^T \boldsymbol{V}^{-1} \boldsymbol{\Sigma}\hat{\boldsymbol{\beta}}_E)^2} \\
&\quad + \frac{(\mathbf{1}_J^T \boldsymbol{V}^{-1} \mathbf{1}_J)(\hat{\boldsymbol{\beta}}_E^{*T} \boldsymbol{V}^{-1} \boldsymbol{\Sigma}\boldsymbol{\theta}) - (\mathbf{1}_J^T \boldsymbol{V}^{-1} \boldsymbol{\Sigma}\hat{\boldsymbol{\beta}}_E)(\mathbf{1}_J^T \boldsymbol{V}^{-1} \boldsymbol{\Sigma}\boldsymbol{\theta})}{(\mathbf{1}_J^T \boldsymbol{V}^{-1} \mathbf{1}_J)(\hat{\boldsymbol{\beta}}_E^{*T} \boldsymbol{V}^{-1} \boldsymbol{\Sigma}\hat{\boldsymbol{\beta}}_E) - (\mathbf{1}_J^T \boldsymbol{V}^{-1} \boldsymbol{\Sigma}\hat{\boldsymbol{\beta}}_E)^2}.
\end{aligned}$$

Assume N_E is sufficiently large so that $\hat{\boldsymbol{\beta}}_E \approx \boldsymbol{\beta}_E$, making the first term γ.

$$E(\hat{\gamma}_{MRE} | \hat{\boldsymbol{\beta}}_E^*) \approx \gamma + \frac{(\mathbf{1}_J^T \boldsymbol{V}^{-1} \mathbf{1}_J)(\hat{\boldsymbol{\beta}}_E^{*T} \boldsymbol{V}^{-1} \boldsymbol{\Sigma}\boldsymbol{\theta}) - (\mathbf{1}_J^T \boldsymbol{V}^{-1} \boldsymbol{\Sigma}\hat{\boldsymbol{\beta}}_E)(\mathbf{1}_J^T \boldsymbol{V}^{-1} \boldsymbol{\Sigma}\boldsymbol{\theta})}{(\mathbf{1}_J^T \boldsymbol{V}^{-1} \mathbf{1}_J)(\hat{\boldsymbol{\beta}}_E^{*T} \boldsymbol{V}^{-1} \boldsymbol{\Sigma}\hat{\boldsymbol{\beta}}_E) - (\mathbf{1}_J^T \boldsymbol{V}^{-1} \boldsymbol{\Sigma}\hat{\boldsymbol{\beta}}_E)^2}.$$

The second term will be zero if there is no direct effect ($\boldsymbol{\theta} = \mathbf{0}$), or the mean of $\boldsymbol{\Sigma}\boldsymbol{\theta}$ is a constant and $\boldsymbol{\theta}$ is independent of $\hat{\boldsymbol{\beta}}_E$. The numerator of the second term is a function of the sample univariate covariance between $\boldsymbol{\Sigma}\boldsymbol{\theta}$ and $\hat{\boldsymbol{\beta}}_E^*$ weighted by $\boldsymbol{V}^{-1}$. This independence between $\hat{\boldsymbol{\beta}}_E$ and $\boldsymbol{\theta}$ is the Instrument Strength Independent of Direct Effect (InSIDE) condition (Bowden et al., 2015). This numerator will go to 0 as the effective number of SNPs increase. We finally look at the LDA MR-Egger estimate:

$$\hat{\gamma}_{LDMRE} = \frac{(\mathbf{1}_J^T \boldsymbol{\Sigma}_G^{-1} \mathbf{1}_J)(\hat{\boldsymbol{\beta}}_E^T \boldsymbol{\Sigma}_G^{-1} \hat{\boldsymbol{\beta}}_G) - (\mathbf{1}_J^T \boldsymbol{\Sigma}_G^{-1} \hat{\boldsymbol{\beta}}_E)(\mathbf{1}_J^T \boldsymbol{\Sigma}_G^{-1} \hat{\boldsymbol{\beta}}_G)}{(\mathbf{1}_J^T \boldsymbol{\Sigma}_G^{-1} \mathbf{1}_J)(\hat{\boldsymbol{\beta}}_E^T \boldsymbol{\Sigma}_G^{-1} \hat{\boldsymbol{\beta}}_E) - (\mathbf{1}_J^T \boldsymbol{\Sigma}_G^{-1} \hat{\boldsymbol{\beta}}_E)^2}.$$

Again, using that $\hat{\boldsymbol{\beta}}_E$ and $\hat{\boldsymbol{\beta}}_G$ are independent, we have that:

$$\begin{aligned}
E(\hat{\gamma}_{LDMRE} | \hat{\boldsymbol{\beta}}_E) &= \gamma \frac{(\mathbf{1}_J^T \boldsymbol{\Sigma}_G^{-1} \mathbf{1}_J)(\hat{\boldsymbol{\beta}}_E^T \boldsymbol{\Sigma}_G^{-1} \boldsymbol{\beta}_E) - (\mathbf{1}_J^T \boldsymbol{\Sigma}_G^{-1} \hat{\boldsymbol{\beta}}_E)(\mathbf{1}_J^T \boldsymbol{\Sigma}_G^{-1} \boldsymbol{\beta}_E)}{(\mathbf{1}_J^T \boldsymbol{\Sigma}_G^{-1} \mathbf{1}_J)(\hat{\boldsymbol{\beta}}_E^T \boldsymbol{\Sigma}_G^{-1} \hat{\boldsymbol{\beta}}_E) - (\mathbf{1}_J^T \boldsymbol{\Sigma}_G^{-1} \hat{\boldsymbol{\beta}}_E)^2} \\
&\quad + \frac{(\mathbf{1}_J^T \boldsymbol{\Sigma}_G^{-1} \mathbf{1}_J)(\hat{\boldsymbol{\beta}}_E^T \boldsymbol{\Sigma}_G^{-1} \boldsymbol{\theta}) - (\mathbf{1}_J^T \boldsymbol{\Sigma}_G^{-1} \hat{\boldsymbol{\beta}}_E)(\mathbf{1}_J^T \boldsymbol{\Sigma}_G^{-1} \boldsymbol{\theta})}{(\mathbf{1}_J^T \boldsymbol{\Sigma}_G^{-1} \mathbf{1}_J)(\hat{\boldsymbol{\beta}}_E^T \boldsymbol{\Sigma}_G^{-1} \hat{\boldsymbol{\beta}}_E) - (\mathbf{1}_J^T \boldsymbol{\Sigma}_G^{-1} \hat{\boldsymbol{\beta}}_E)^2}.
\end{aligned}$$

As with the other estimates we assume that N_E is sufficiently large:

$$E(\hat{\gamma}_{LDMRE} | \hat{\boldsymbol{\beta}}_E) \approx \gamma + \frac{(\mathbf{1}_J^T \boldsymbol{\Sigma}_G^{-1} \mathbf{1}_J)(\hat{\boldsymbol{\beta}}_E^T \boldsymbol{\Sigma}_G^{-1} \boldsymbol{\theta}) - (\mathbf{1}_J^T \boldsymbol{\Sigma}_G^{-1} \hat{\boldsymbol{\beta}}_E)(\mathbf{1}_J^T \boldsymbol{\Sigma}_G^{-1} \boldsymbol{\theta})}{(\mathbf{1}_J^T \boldsymbol{\Sigma}_G^{-1} \mathbf{1}_J)(\hat{\boldsymbol{\beta}}_E^T \boldsymbol{\Sigma}_G^{-1} \hat{\boldsymbol{\beta}}_E) - (\mathbf{1}_J^T \boldsymbol{\Sigma}_G^{-1} \hat{\boldsymbol{\beta}}_E)^2}.$$

The numerator of the second term is a function of the sample univariate covariance between $\hat{\boldsymbol{\beta}}_E$ and $\boldsymbol{\theta}$ weighted by $\boldsymbol{\Sigma}_G^{-1}$. If $\boldsymbol{\theta}$ is independent of $\hat{\boldsymbol{\beta}}_E$ and the mean of $\boldsymbol{\theta}$ is a constant-vector, then the term will go to 0 as the number of effective SNPs (J_{EFF}) in the loci increases (Appendix 3.10.3). The above results give us:

$$E(\hat{\gamma}_{LDMRE}) \to \gamma \text{ as } N_E \to \infty \text{ and } J_{EFF} \to \infty.$$

In practice at an individual loci, this assumption of the effective number of SNPs increasing is likely unreasonable. There are a finite number of SNPs in the genome and the loci. If $\boldsymbol{\theta}$ is a constant however, the estimate will converge to γ.

3.5 Simulation

Parameter	Definition	Values taken
h_E^2	Proportion of Variability in $\boldsymbol{M}$ explained by $\boldsymbol{G}$	0.01, 0.20, 0.50
$h_{E\to Y}^2$	Proportion of Variability in $\boldsymbol{Y}$ explained by $\boldsymbol{M}$	0, 0.005, 0.01
$h_{G\to Y}^2$	Proportion of Variability in $\boldsymbol{Y}$ explained by $\boldsymbol{G}$	0, 0.005, 0.01
N	Sample Size of GWAS	5000
N_E	Sample Size of EQTL	1000
ρ	Correlation between SNPs. AR structure	0.125, 0.9
J	Number of SNPs in the Loci	50,300
τ	Strength of Pleiotropic Effects	0,10
L	Cholesky Decomposition of $\boldsymbol{\Sigma}$	Function of ρ and J

Table 3.2: Parameters that were varied across all simulations.

We examine the empirical performance of four methods: the MR, MR-Egger, LDA MR, and the LDA MR-Egger. Under our simulation scenario, the TWAS and the LDA MR are approximately the same ($\sigma_{G,j}^*$ is a constant). For each simulation, we generated summary eQTL and GWAS statistics from a multivariate normal distribution as opposed to individual level data. We fixed the sample of the eQTL study to 1000 (N_E) and the sample size of the GWAS to 5000 (N). We varied the number of SNPs at the loci (J), the proportion of variation in $\boldsymbol{Y}$ explained by $\boldsymbol{G}$ ($h_{G\to Y}^2$) and $\boldsymbol{M}$ ($h_{E\to Y}^2$), proportion of variation in $\boldsymbol{M}$ explained by $\boldsymbol{G}$ (h_e^2), the variance of the direct effect (τ), and the LD matrix

($\boldsymbol{\Sigma}$, AR(J) structure, $\boldsymbol{\Sigma}_{i,j} = \rho^{|i-j|}$). More details along with which values were taken are given in Table 3.2. The process and order for the generation of the simulation data is provided in Table 3.3.

Step	Procedure	Mathematically
1	Generate the true eQTL	$\boldsymbol{\beta}_E \sim N_J(0, \boldsymbol{I}_J)$
2	Proportion of variability in $\boldsymbol{M}$ due to $\boldsymbol{G}$	$\sigma_E^2 = \frac{h_e^2}{\boldsymbol{\beta}_E^T \boldsymbol{\Sigma} \boldsymbol{\beta}_E}$
3	Rescale true eQTL	$\boldsymbol{\beta}_E = \sigma_E \boldsymbol{\beta}_E$
4	Set Expression to outcome Effect	$\gamma^2 = h^2_{E \to Y}$
5	Generate potential Direct Effect	$\boldsymbol{\theta} = e + \tau; e \sim N_J(0, \boldsymbol{I}_J)$
6	Proportion of variability in $\boldsymbol{Y}$ due directly to $\boldsymbol{G}$	$\sigma^2 = \frac{h^2_{G \to Y}}{\boldsymbol{\theta}^T \boldsymbol{\Sigma} \boldsymbol{\theta}}$
7	Rescale Direct Effects	$\boldsymbol{\theta} = \sigma \boldsymbol{\theta}$
8	Generate GWAS effects	$\boldsymbol{\beta}_G = \boldsymbol{\theta} + \gamma \boldsymbol{\beta}_E$
9	Generate estimated eQTL	$\hat{\boldsymbol{\beta}}_E^* \sim \boldsymbol{\sigma} \boldsymbol{\beta}_E + L^T \epsilon_E;$ $\epsilon_E \sim N_J(0, \frac{1-h_E^2}{N_E} \boldsymbol{I}_J)$
10	Generate estimated GWAS	$\hat{\boldsymbol{\beta}}_G * \sim \boldsymbol{\Sigma} \boldsymbol{\beta}_G + L^T \boldsymbol{\epsilon}_G$ $\boldsymbol{\epsilon}_G \sim N_J(0, \boldsymbol{I}_J \frac{1-h_E^2 h^2_{E \to Y} - h^2_{G \to Y}}{N})$

Table 3.3: Procedure to simulate data for given values in Table 3.2

We performed 5x10^4 simulations for each combination of parameters (486 different combinations). In each simulation, we generated a new true $\boldsymbol{\beta}_E$ and $\boldsymbol{\beta}_G$, which are a function of $h^2_{G \to Y}, h^2_e, \tau$ and J . Generating a new "true" parameter value, better represents different eQTL and GWAS patterns across the genome. The procedure detailed in Table 3.3 is thus repeated 50K times for all combinations of the parameters in Table 3.2. Type I error and power were evaluated at 0.05.

As the true $\boldsymbol{\beta}_E$ and $\boldsymbol{\beta}_G$ were drawn from a random normal with mean 0, this will make the γ estimates appear unbiased. Therefore if we take the mean across the simulations for a given parameter combinations, it will appear that the estimates are unbiased due to the "true" eQTL and GWAS effects being a random draw centered at zero. To account for this, we next performed a set of 10^4 simulations with a fixed true value to highlight potential bias. We generate one true $\boldsymbol{\beta}_E$ and $\boldsymbol{\theta}$ for each value of J that is then held constant for all simulations. We then vary the parameters in Table 3.2. Therefore,

Steps 1 and 5 of Table 3.3 are not performed in this set of simulations, or more accurately, only performed once for both J=50 and J=300.

3.6 Real Data Application

We also performed an analysis to detect genes that are associated with risk of Breast Cancer. The marginal GWAS summary statistics were from a recent GWAS on Breast Cancer within women of European descent (Michailidou et al., 2017). SNP data was meta-analyzed from 13 various platforms: a new OncoArray (Amos et al., 2017), iCOGS, and eleven other GWAS. More details on the QC procedure can be found in Michailidou et al (Michailidou et al., 2017). After QC, the study consisted of 11.8 million SNPs, with 105,974 controls and 122,977 cases. The GWAS estimates were calculated on the non-standardized minor allele counts, and therefore were transformed using the minor allele frequency as detailed in Appendix 3.10.1.

Expression weights were calculated from GTEx along with LD information in breast tissue in an overall sample of 183 individuals (Lonsdale et al., 2013). We restricted our analysis to the set of transcripts that were deemed heritable using GCTA (Yang et al., 2011) analyzing SNPs within 500 BP of the gene boundary. The expression weights were calculated on standardize minor allele counts of SNPs (mean zero, variance one) and were conditionally estimated using Bayesian Sparse Linear Mixed Models (Zhou et al., 2013). A gene was deemed heritable if the GCTA p-value was less than the Bonferroni threshold of 0.05 (adjusting for 27,945 tests) and was cis-heritable across all GTEx tissues. We were left with 683 potential gene transcripts to analyze.

We analyze the data using the TWAS, LDA MR, and LDA MR-Egger. We did not examine the MR and MR-Egger as we were testing for cis-signals as opposed to genome wide, and therefore the SNPs were in LD. We took the overlap of the SNPs from the GWAS with the available SNP correlations from GTEx. If the effects were estimated with respect to a different reference allele, we changed the sign for that eQTL effect estimate. In total, the 683 genes corresponded to 191,583 unique SNPs.

3.7 Simulation Results

We first examine the type I error results when J=50 (Figure 3.1). The results for J=300 are similar and provided in Supplement Figure 4.27. Starting in the top left and moving right (Figure 3.1 A1), we see that when there is little LD and no direct effect of the SNPs, all of the approaches have the correct type I error. When there is little LD and the direct effect is variable, the MR and MR-Egger approaches have modestly inflated type I error rates. Finally, when there is low LD and a constant direct effect, only the LDA MR-Egger has correct type I error. If the SNPs are in high LD (second row of Figure 3.1) and there is no direct effect, the MR and the MR-Egger have inflated type I error due to misspecification of the variance. When there is a variable direct effect, all four approaches have inflated type I error. Finally, when there is strong linkage and a constant direct effect, only the LDA MR-Egger has the correct type I error.

We next examined the power when there is little to no LD between the SNPs and no direct effect (Figure 3.2 A1). Under this situation, all four approaches had correct type I error (Figure 3.1 A1 , Supplementary Figure 4.27 A1). Regardless of the magnitude of the effect of $\boldsymbol{M}$ on $\boldsymbol{Y}$, we see similar power amongst all four methods. There is a slightly larger power for the LDA MR compared to the LDA MR-Egger when the SNPs only explain 20% of the variation in the gene expression, but once the SNPs explain 50% of the variation in $\boldsymbol{M}$, all approaches have approximately equal power. In each individual plot, we see that as the SNPs explain more of the variation (become better instruments), we have an increase in power.

The decrease in power from J=50 to J=300 is due to an increase in the signal to noise when predicting expression levels using SNPs. For J=50, a larger proportion of the variation explained by SNPs is shared by each individual SNP leading to more precise estimates. When J=300, a smaller proportion of that same amount of variation is explained by each SNP in a larger set of SNPs. Assuming the sample size in the reference panel used to estimate $\hat{\boldsymbol{\beta}}_E$ is the same, the sampling error in the SNP-specific estimates $\hat{\beta}_{E,j}$ is the same for J=50 and J=300. We have held the proportion of variance explained by the SNPs constant, decreasing each individuals SNPs effect (Table 3.3 Step 2) making it

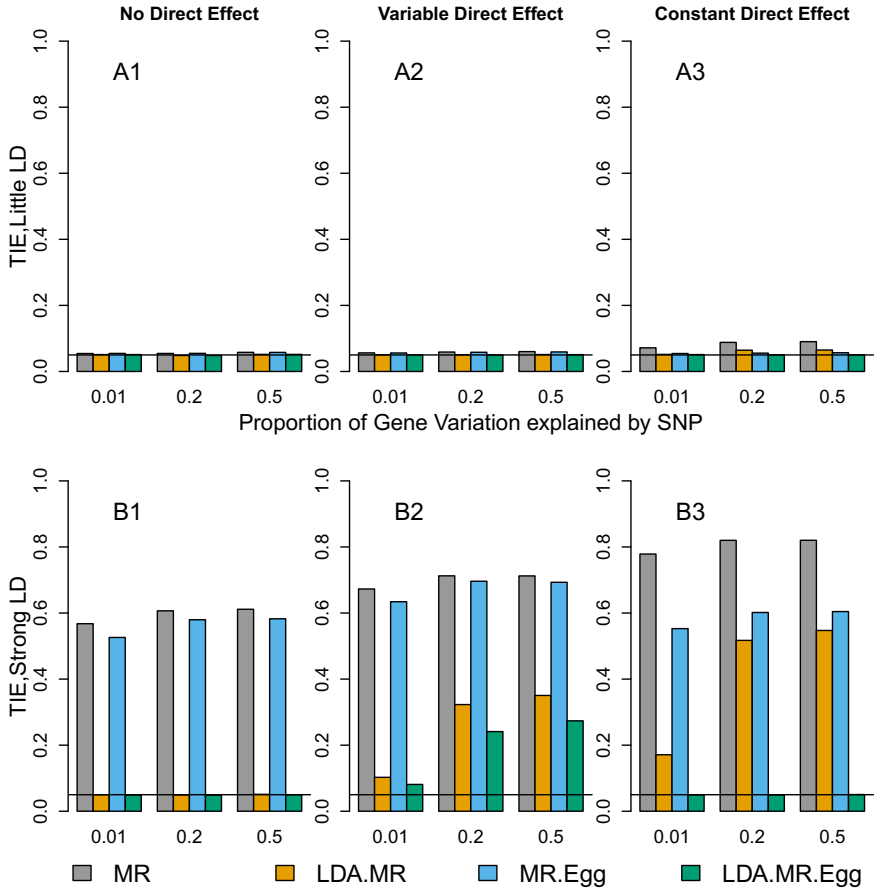

Figure 3.1: Type I error results when J=50. Each bar represents results over 5x10^4 simulations. Evaluated at $\alpha = 0.05$. First panel represent when low LD (plots with A). Second panel represents when strong LD (plots with B). From left to right correspond to: no direct effect, variable direct effect, and a constant direct effect.

harder to estimate. Tye noise has remained constant while the signal has decreased.

When there is low LD and a constant direct effect of the SNPs on the outcome (Table 3.4), the MR-Egger has less power than the LDA MR-Egger to detect an association when J=50. This may be due to the LDA MR-Egger gaining some information by accounting for the weak LD structure. If J=300, there is a decrease in power compared to when J=50 regardless of the presence of a direct effect in these two methods. At J=300, the LDA MR-Egger has slightly higher power than the MR-Egger. When the SNPs explain 50% of the variation in $\boldsymbol{M}$, both methods have power greater than 80% regardless of the effect size (Table 3.4). We did not report the power of the MR or LDA MR, as they did not have proper type I error when there is a constant direct effect (Figure 1 and Supplement

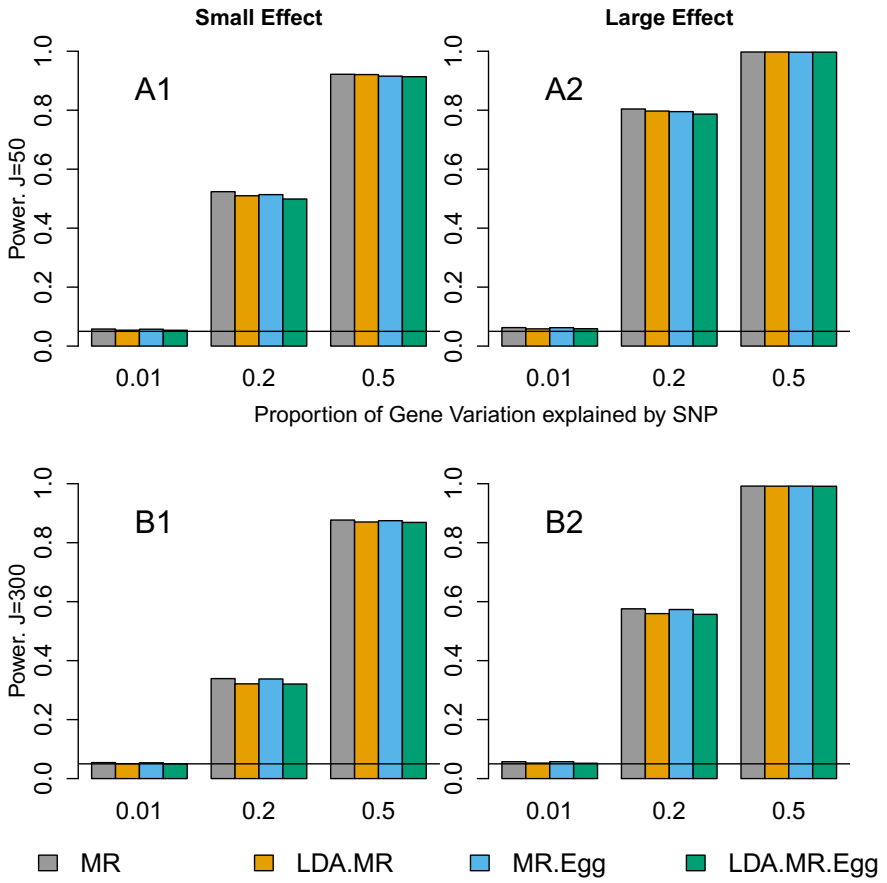

Figure 3.2: Power results when there is little to no LD and no direct effect. Each bar represents results over 5x10^4 simulations. Evaluated at $\alpha = 0.05$. First row represents J=50 and second row when J=300. From left to right: when $\gamma^2 = 0.005$ and $\gamma^2 = 0.01$.

Figure 1).

Finally, we examine the power when there is strong LD amongst the SNPs (Table 3.5). We do not assess the MR or MR-Egger for this scenario as they do not account for LD. When J=50, and there is no direct effect, the LDA MR has more power than the LDA MR-Egger, though the difference in power is less pronounced when $h_E^2 = 0.5$ and $\boldsymbol{M}$ has a strong effect on $\boldsymbol{Y}$. When J=300, the two methods have comparable power (Table 3.4), with the LDA MR-Egger having slightly less power than the LDA MR. When there is a direct effect, the LDA MR-Egger has approximately equal power to when there was no direct effect.

Comparing the power of the LDA MR-Egger when there is strong vs small LD

(Tables 3.4 and 3.5), and J=50, and there is a small mediated effect with $h^2_E = 0.5$, we have less power compared to when there is little LD (LDA MR-Egger 0.914 to 0.784). When J=300, there is a smaller drop-in power for small to strong LD, with the LDA MR-Egger power going from 0.869 to 0.840. The LDA MR has equal power comparing when the SNPs are strongly correlated vs weakly correlated.

We next examine the bias of our estimates (Figure 3.3). We here show the results when there is strong LD (Figure 3.3). The results for low LD are in Supplement Figure 4.28. Here we have fixed $\boldsymbol{\beta}_E$ and $\boldsymbol{\theta}$ effects for all simulations.

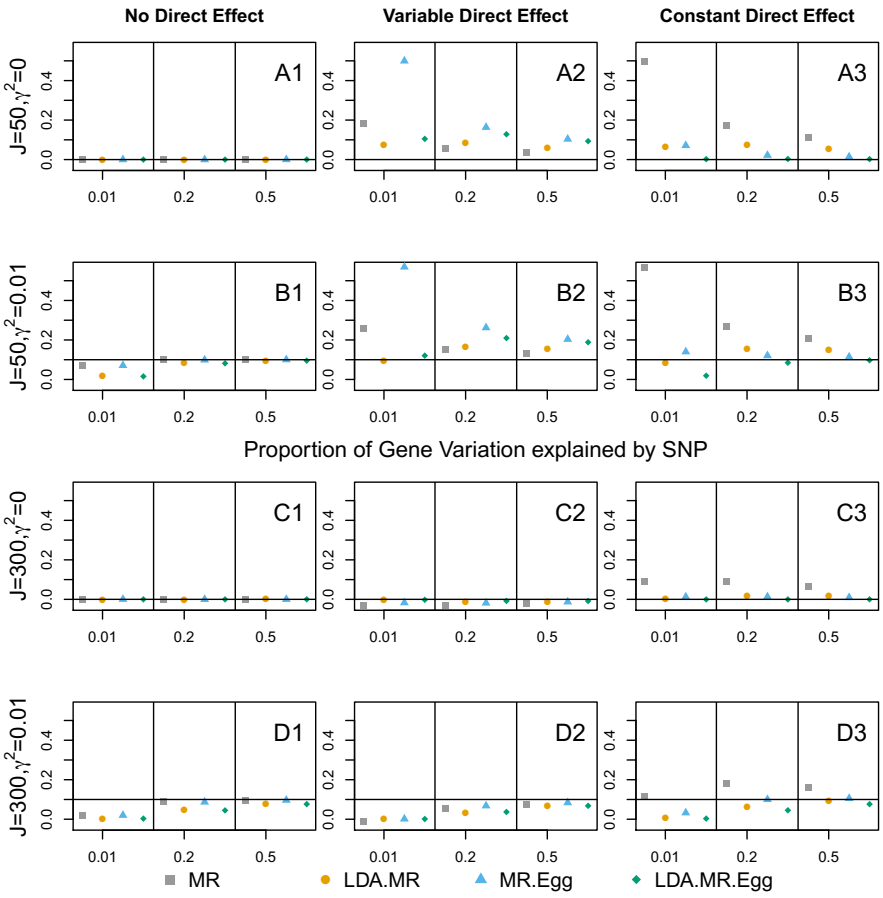

Figure 3.3: Bias plots for when there is strong LD in the SNP set. First row corresponds to J=50, $\gamma = 0$ (plots with A). Second panel (plots with B) when $J = 50$ and $\gamma^2 = 0.01$. Third panel (plots with C) when $J = 300$ and $\gamma = 0$. Final panel (plots with D) $J = 300$ and $\gamma^2 = 0.01$. From left to right: no direct effect, variable direct effect, constant direct effect.

As expected based on the section on bias of estimates, when $\gamma = 0$ and $\boldsymbol{\theta} = \mathbf{0}$, all of the estimates are unbiased (Figure 3.3 A1 and C1). When there is an effect on the

outcome and no direct effect, we see that the non LD aware approaches converge faster to the truth than the LD aware (Figure 3.3 B1 and D1). While the non LDA converge faster, recall that they have incorrect type I error. We also see attenuation bias when $\gamma \neq 0$, with estimates improving as the SNPs become better instruments. The attenuation bias is larger for J=300 relative to J=50, for the same reason that we saw a decrease in power from J=50 to J=300: an increase in signal to noise. When there is a variable direct effect, all of the approaches are biased (second column). Our empirical bias results depend on the particular values of $\boldsymbol{\beta}_E$ and $\boldsymbol{\theta}$. All of the estimates for the mediated effect of gene expression include a weighted covariance between $\boldsymbol{\beta}_E$ and $\boldsymbol{\theta}$. Even if the average across all genes for this covariance is zero (as would be the case under the InSIDE condition), for any particular gene it is non-zero, and can be large. Under the InSIDE condition, the average absolute magnitude of the bias term is a decreasing function of J. This is why we see less bias estimates when J=300, when there is a variable direct effect. Finally, when there is a constant direct effect, only the LDA MR-Egger is unbiased when $\gamma = 0$. When J=300 and $\gamma \neq 0$, we see the LDA MR -Egger having attenuation bias.

3.8 Real Data Results

Of the 683 genes tested, 74 were called significant by at least one approach (TWAS, LDA MR or LDA MR-Egger, Supplementary Tables 4.13 to 4.19). This large proportion of genes being significant is not surprising given that we pre-selected loci where we knew the SNPs were strong instruments (via our GCTA heritability analysis). Comparing the TWAS vs the LDA MR-Egger (Figure 3.4A), there were 19 genes that were significant by the TWAS but not by LDA MR-Egger, and 13 that were called by both. With the LDA MR and the LDA MR-Egger, there was much more agreement due to the same weight matrix being used, but still the LDA MR called twenty-three genes as significant that the LDA-MR Egger did not (Figure 3.4B). Twenty-seven gene probes were called significant by both the LDA-MR and the LDA-MR Egger (Supplement Table 4.16). There were eleven probes called significant by all three approaches. A detailed list of which probes were found significant by which method is provided in Supplementary Tables 4.13 to 4.19. We did not

see much agreement between the TWAS and LDA as the TWAS was calculated using the conditional, and $\sigma^*_{G,j}$ was not a constant in every loci (Figure 3.4 C).

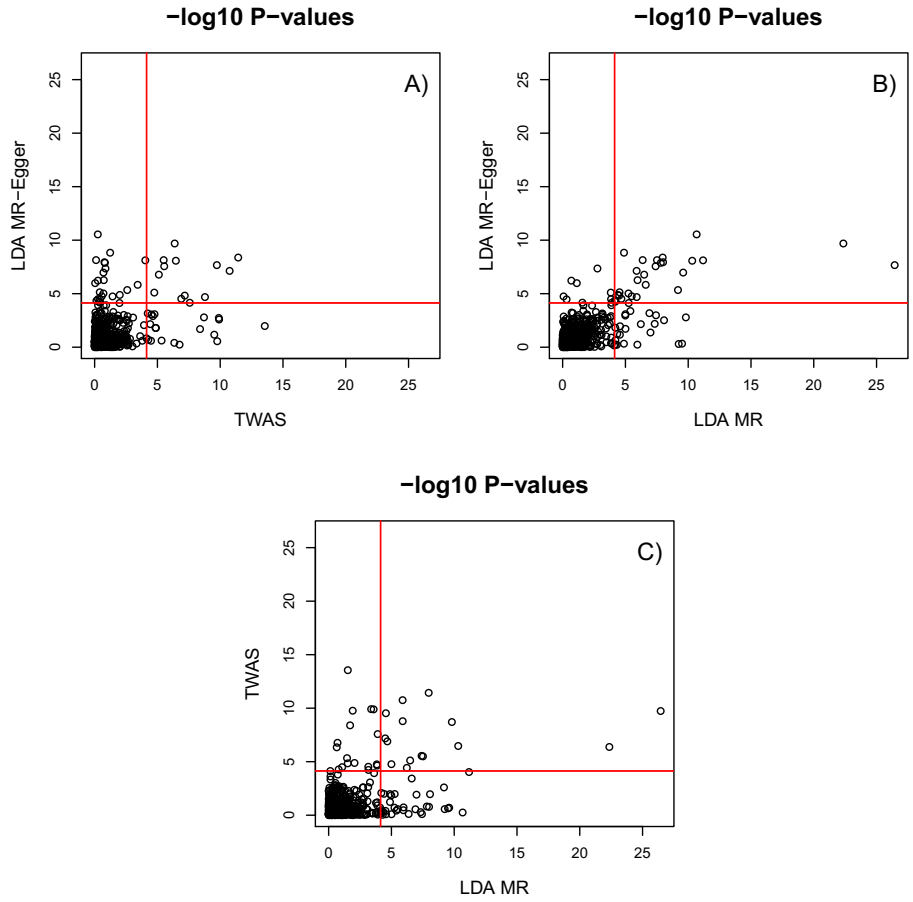

Figure 3.4: -log10 p-values for 683 genes. Line is the Bonferroni cutoff of -log10(.05/683). A) TWAS vs LDA MR-Egger. B) LDA MR vs TWAS and C) LDA MR vs TWAS.

One gene we will highlight that was called significant by TWAS was SETD9, with a p-value of 1.72e-10 at cytoband 5q11.2. The p-value for this probe from the LDA MR and the LDA MR-Egger was 0.012 and 0.278 respectively. A fine mapping analysis of this locus found four functional candidate SNPs in a sample of approximately 100K women of European descent (Glubb et al., 2015). These four candidate functional SNPs were associated with an increase in activity of MAP3K1, another gene in this loci. SETD9 was ruled as not the gene of interest as it had no association with these four candidate SNPs. MAP3K1 is located 94K BP from SETD9. MAP3K1 did not pass our cis-heritable tissue threshold and was not included in our analysis.

Another association of interest was at the 2q33.1 locus with Caspase 8 (CASP8) the most significant probe by the TWAS (p-value 2.8e-14), but not significant by the LDA MR-Egger (p-value 0.01). A conditional TWAS using nearby gene expression that was done by our group, found that upon conditioning for nearby gene ALS2CR12, CASP8 was no longer called significant. This was done on a separate data set with different gene expression weights however. This gene did not pass the cis-heritable threshold in GTEx and was not included in our analysis. The gene ALS2CR12 is located 1000 BP from CASP8.

3.9 Discussion

In this paper, we propose a new LD aware MR-Egger estimate of the effect of gene expression on an outcome using just summary statistics. This method properly accounts for both LD and when there is a constant direct effect. The LDA MR and the summary MR and MR-Egger are not proper estimates under these situations as they either do not account for the direct effect (LDA MR, MR) or the LD (MR and MR-Egger). For scenarios where the LDA MR is a valid test, the LDA MR-Egger has comparable or slightly lower power. However, the researchers can gain confidence that they are correctly adjusting for pleiotropy when the pleiotropic effect is constant. We also provided a 1-1 relationship between the TWAS and the traditional LDA MR when the standard errors of the marginal GWAS effects are constant. In our real data application, the three approaches (TWAS, LDA MR and LDA MR-Egger) all picked up on different results, with the LDA MR-Egger potentially correctly calling some results as null that the TWAS did not.

In our simulations, we saw that the type I error was correct when J=50, there was small LD, and a variable direct effect (Figure 3.1). This lack of inflation is due to our inclusion of a residual variance term in our estimates, and how we generated the simulation data (Appendix E). Briefly, the inclusion of the residual term helps to account for some of the variance in the GWAS effects that are due to the variation in the direct effects.

The validity of the InSIDE condition is not guaranteed. In practice, the SNPs may act through other intermediate genes that are in high correlation with the gene of inter-

est. However, it is difficult to assess the validity of these assumptions as it is rare to have expression, outcome, and SNP data for a large set of individuals. The additional assumption of the mean of the direct effect being a constant is likely true across the genome, but may not hold at individual loci. If this assumption is violated, the estimates may be bias.

Future methods could incorporate eQTL results from multiple nearby genes when there is SNP overlap. We would then have γ_k and $\hat{\boldsymbol{\beta}}_{E,k}$ for k=1...K for the K genes in the region, with information on what eQTL maps to what gene. This could lead to a conditional analyses, and could help parse out the direct effect signal. As mentioned in the text, LDA MR-Egger does not work when the direct effects are variable, and this conditional analysis incorporating nearby genes with an overlap of eQTL's could help correct this. A similar conditional TWAS approach was proposed by Gusev et al (Gusev et al., 2016b). The potential model is given in Appendix 3.10.5.

While we focused on the simple weighted linear regressions, there are extensions that could incorporate the variance of the eQTL data and estimate the parameters via likelihood maximization. These estimates do not have a closed form solution and require an iterative algorithm. A possible estimation algorithm is given in Appendix 3.10.5. It will provide more precise estimates of λ and γ, but at the cost of higher variance. This is to be expected as we now incorporates the variance of the eQTL estimates. This should be looked into more detail in future research to evaluate convergence and the proper covariance and standard errors of these estimates. Burgess et al, proposed a similar approach using Bayesian MCMC or a numerical optimization but for the LDA MR (Burgess et al., 2016).

While we focused on the case of gene expression data, the LDA MR-Egger can easily be extended to an arbitrary mediator when the instruments are correlated. The approaches we consider are all a function of three things: the correlation of the instruments, the strength of the instruments, and the association between the instruments and the outcome not taking into account the intermediate variable. It is thus easily extendable and executable.

In practice, we caution interpretation of the γ terms due to quality control and preprocessing of the data making it difficult to infer meaningful biological interpretation of

γ. We are also faced with the issue of attenuation bias, as we have estimates of the true eQTL effects. Despite this, the LDA MR-Egger is still a valid test for the effect of the gene on more often than the other approaches in most scenarios and can inform the direction of the effect.

In summary, we have extended the use of LDA MR and MR-Egger into LDA MR-Egger, a useful tool for when the genetic instruments are correlated and disease and eQTL SNPs have colocalized. This method of incorporating summary statistics from different sources can help in discovering novel loci as well as narrowing in on the susceptible gene in the region. We provided equations for their bias as well as evaluated their performance empirically. Further work can be done to account for the variation in the eQTL estimates and in de-noising the direct effects.

3.10 Appendix

3.10.1 Transformation of GWAS or eQTL statistics

Here we discuss howto transform the GWAS or eQTL estimates calculated on the minor allele count to what would have been observed on standardized genomes. Let $\hat{\beta}^*_{G,j,M}$, $\hat{\sigma}^*_{G,j,M}$, and $\hat{\beta}^*_{E,j,M}$ represent the marginal GWAS estimate, marginal GWAS standard error, and marginal eQTL estimate for SNP j when estimated on the minor allele count of that SNP. To transform to the value observed under the standardized SNP j (p_j represent the minor allele frequency for SNP j):

$$\hat{\beta}^*_{G,j} = \frac{\hat{\beta}^*_{G,j,M}}{\sqrt{2p_j(1-p_j)}},$$
$$\hat{\beta}^*_{E,j} = \frac{\hat{\beta}^*_{E,j,M}}{\sqrt{2p_j(1-p_j)}},$$
$$\hat{\sigma}^*_{G,j} = \frac{\hat{\sigma}^*_{G,j,M}}{\sqrt{2p_j(1-p_j)}}.$$

3.10.2 Relationship between TWAS and LDA MR

If $\sigma^*_{G,j} = v$ for all SNPs, the TWAS and LDA-MR will be proportional to each other. Recall that:

$$Z_{TWAS} = \frac{\hat{\boldsymbol{\beta}}_E^{*T} \boldsymbol{\Sigma}^{-1} \boldsymbol{Z}^*_G}{\sqrt{\hat{\boldsymbol{\beta}}_E^{*T} \boldsymbol{\Sigma}^{-1} \hat{\boldsymbol{\beta}}^*_E}}.$$

If $\sigma^*_{G,j} = v$, then $\boldsymbol{Z}^*_G = \hat{\boldsymbol{\beta}}^*_G / v$. Also, recall $\hat{\boldsymbol{\beta}}_E = \boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\beta}}^*_E$ and $\hat{\boldsymbol{\beta}}_G = \boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\beta}}^*_G$, we then have:

$$\frac{\hat{\boldsymbol{\beta}}_E^{*T} \boldsymbol{\Sigma}^{-1} \boldsymbol{Z}^*_G}{\sqrt{\hat{\boldsymbol{\beta}}_E^{*T} \boldsymbol{\Sigma}^{-1} \hat{\boldsymbol{\beta}}^*_E}} = \frac{\hat{\boldsymbol{\beta}}_E^{*T} \boldsymbol{\Sigma}^{-1} \hat{\boldsymbol{\beta}}^*_G v^{-1}}{\sqrt{\hat{\boldsymbol{\beta}}_E^{*T} \boldsymbol{\Sigma}^{-1} \hat{\boldsymbol{\beta}}^*_E}} \frac{v^{-1}}{v^{-1}} = \frac{\hat{\boldsymbol{\beta}}_E^{*T} \boldsymbol{\Sigma}^{-1} v^{-2} \hat{\boldsymbol{\beta}}^*_G}{\sqrt{\hat{\boldsymbol{\beta}}_E^{*T} \boldsymbol{\Sigma}^{-1} v^{-2} \hat{\boldsymbol{\beta}}^*_E}}.$$

Recall that $\boldsymbol{\Sigma}^{-1} v^{-2} = \boldsymbol{\Sigma}_G$ when all marginal variances arerqual and therefore $\boldsymbol{\Sigma}_G^{-1} = \boldsymbol{\Sigma} v^{-2}$.

$$Z_{TWAS} = \frac{\hat{\boldsymbol{\beta}}_E^{T} v^{-2} \hat{\boldsymbol{\beta}}^*_G}{\sqrt{\hat{\boldsymbol{\beta}}_E^{T} v^{-2} \hat{\boldsymbol{\beta}}^*_E}} = \frac{\hat{\boldsymbol{\beta}}_E^{T} v^{-2} \boldsymbol{\Sigma} \boldsymbol{\Sigma}^{-1} \hat{\boldsymbol{\beta}}^*_G}{\sqrt{\hat{\boldsymbol{\beta}}_E^{T} v^{-2} \boldsymbol{\Sigma} \boldsymbol{\Sigma}^{-1} \hat{\boldsymbol{\beta}}^*_E}} = \frac{\hat{\boldsymbol{\beta}}_E^{T} \boldsymbol{\Sigma}_G \hat{\boldsymbol{\beta}}_G}{\sqrt{\hat{\boldsymbol{\beta}}_E^{T} \boldsymbol{\Sigma}_G \hat{\boldsymbol{\beta}}_E}} = Z_{LDAMR} \hat{\sigma}_{LDAMR}.$$

The desired result.

3.10.3 Convergence of the LDA MR-Egger Parameters

The LDA MR-Egger estimates are:

$$\begin{bmatrix} \hat{\lambda}_{LD} \\ \hat{\gamma}_{LDMRE} \end{bmatrix} = \begin{bmatrix} \mathbf{1}_J^T \boldsymbol{\Sigma}_G^{-1} \mathbf{1}_J & \mathbf{1}_J^T \boldsymbol{\Sigma}_G^{-1} \hat{\boldsymbol{\beta}}_E \\ \hat{\boldsymbol{\beta}}_E^T \boldsymbol{\Sigma}_G^{-1} \mathbf{1}_J & \hat{\boldsymbol{\beta}}_E^T \boldsymbol{\Sigma}_G^{-1} \hat{\boldsymbol{\beta}}_E \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{1}_J^T \boldsymbol{\Sigma}_G^{-1} \\ \hat{\boldsymbol{\beta}}_E^{*T} \boldsymbol{\Sigma}_G^{-1} \end{bmatrix} \hat{\boldsymbol{\beta}}_G,$$

$$\begin{aligned} E(\hat{\gamma}_{LDMRE} | \hat{\boldsymbol{\beta}}_E) = \gamma & \frac{(\mathbf{1}_J^T \boldsymbol{\Sigma}_G^{-1} \mathbf{1}_J)(\hat{\boldsymbol{\beta}}_E^T \boldsymbol{\Sigma}_G^{-1} \boldsymbol{\beta}_E) - (\mathbf{1}_J^T \boldsymbol{\Sigma}_G^{-1} \hat{\boldsymbol{\beta}}_E)(\mathbf{1}_J^T \boldsymbol{\Sigma}_G^{-1} \boldsymbol{\beta}_E)}{(\mathbf{1}_J^T \boldsymbol{\Sigma}_G^{-1} \mathbf{1}_J)(\hat{\boldsymbol{\beta}}_E^T \boldsymbol{\Sigma}_G^{-1} \hat{\boldsymbol{\beta}}_E) - (\mathbf{1}_J^T \boldsymbol{\Sigma}_G^{-1} \hat{\boldsymbol{\beta}}_E)^2} \\ & + \frac{(\mathbf{1}_J^T \boldsymbol{\Sigma}_G^{-1} \mathbf{1}_J)(\hat{\boldsymbol{\beta}}_E^T \boldsymbol{\Sigma}_G^{-1} \boldsymbol{\theta}) - (\mathbf{1}_J^T \boldsymbol{\Sigma}_G^{-1} \hat{\boldsymbol{\beta}}_E)(\mathbf{1}_J^T \boldsymbol{\Sigma}_G^{-1} \boldsymbol{\theta})}{(\mathbf{1}_J^T \boldsymbol{\Sigma}_G^{-1} \mathbf{1}_J)(\hat{\boldsymbol{\beta}}_E^T \boldsymbol{\Sigma}_G^{-1} \hat{\boldsymbol{\beta}}_E) - (\mathbf{1}_J^T \boldsymbol{\Sigma}_G^{-1} \hat{\boldsymbol{\beta}}_E)^2}. \end{aligned}$$

Assuming that N_E is sufficiently large such that $\hat{\boldsymbol{\beta}}_E \approx \boldsymbol{\beta}_E$:

$$E(\hat{\gamma}_{LDMRE} | \hat{\boldsymbol{\beta}}_E) \approx \gamma + \frac{(\mathbf{1}_J^T \boldsymbol{\Sigma}_G^{-1} \mathbf{1}_J)(\hat{\boldsymbol{\beta}}_E^T \boldsymbol{\Sigma}_G^{-1} \boldsymbol{\theta}) - (\mathbf{1}_J^T \boldsymbol{\Sigma}_G^{-1} \hat{\boldsymbol{\beta}}_E)(\mathbf{1}_J^T \boldsymbol{\Sigma}_G^{-1} \boldsymbol{\theta})}{(\mathbf{1}_J^T \boldsymbol{\Sigma}_G^{-1} \mathbf{1}_J)(\hat{\boldsymbol{\beta}}_E^T \boldsymbol{\Sigma}_G^{-1} \hat{\boldsymbol{\beta}}_E) - (\mathbf{1}_J^T \boldsymbol{\Sigma}_G^{-1} \hat{\boldsymbol{\beta}}_E)^2}.$$

Focusing on the second term, it can be rewritten as:

$$\frac{(\mathbf{1}_J^T \boldsymbol{\Sigma}_G^{-1} \mathbf{1}_J)(\hat{\boldsymbol{\beta}}_E^T \boldsymbol{\Sigma}_G^{-1} \boldsymbol{\theta}) - (\hat{\boldsymbol{\beta}}_E^T \boldsymbol{\Sigma}_G^{-1} \mathbf{1}_J)(\mathbf{1}_J^T \boldsymbol{\Sigma}_G^{-1} \boldsymbol{\theta})}{(\mathbf{1}_J^T \boldsymbol{\Sigma}_G^{-1} \mathbf{1}_J)(\hat{\boldsymbol{\beta}}_E^T \boldsymbol{\Sigma}_G^{-1} \hat{\boldsymbol{\beta}}_E) - (\mathbf{1}_J^T \boldsymbol{\Sigma}_G^{-1} \hat{\boldsymbol{\beta}}_E)^2}.$$

The numerator is the sample univariate covariance between $\boldsymbol{\theta}$ and $\hat{\boldsymbol{\beta}}_E$ weighted by $\boldsymbol{\Sigma}_G^{-1}$, times $(\mathbf{1}_J^T\boldsymbol{\Sigma}_G^{-1}\mathbf{1}_J)^2$. The denominator is the sample univariate variance of $\hat{\boldsymbol{\beta}}_E$ weighted by $\boldsymbol{\Sigma}_G^{-1}$ times $(\mathbf{1}_J^T\boldsymbol{\Sigma}_G^{-1}\mathbf{1}_J)^2$, If $\boldsymbol{\theta}$ is independent of $\hat{\boldsymbol{\beta}}_E$ and the mean of $\boldsymbol{\theta}$ is a constant-the term will go to 0 as $J_{EFF} \to \infty$. J_{EFF} denotes the effective number of SNPs at the loci. We take the expectation of the numerator conditional on $\hat{\boldsymbol{\beta}}_E$:

$$E_\theta\left(E(\hat{\gamma}_{LDMRE}|\hat{\boldsymbol{\beta}}_E)|\hat{\boldsymbol{\beta}}_E\right)$$
$$= (\mathbf{1}_J^T\boldsymbol{\Sigma}_G^{-1}\mathbf{1}_J)\hat{\boldsymbol{\beta}}_E^T\boldsymbol{\Sigma}_G^{-1}E(\boldsymbol{\theta}|\hat{\boldsymbol{\beta}}_E) - (\hat{\boldsymbol{\beta}}_E^T\boldsymbol{\Sigma}_G^{-1}\mathbf{1}_J)\mathbf{1}_J^T\boldsymbol{\Sigma}_G^{-1}E(\boldsymbol{\theta}|\hat{\boldsymbol{\beta}}_E).$$

If $\boldsymbol{\theta}$ is independent of $\hat{\boldsymbol{\beta}}_E$, then $E(\boldsymbol{\theta}|\hat{\boldsymbol{\beta}}_E) = E(\boldsymbol{\theta}) = \mu_\theta$ and if $\mu_\theta = c\mathbf{1}_J$, the above reduces to:

$$(\mathbf{1}_J^T\boldsymbol{\Sigma}_G^{-1}\mathbf{1}_J)\hat{\boldsymbol{\beta}}_E^T\boldsymbol{\Sigma}_G^{-1}c\mathbf{1}_J - (\hat{\boldsymbol{\beta}}_E^T\boldsymbol{\Sigma}_G^{-1}\mathbf{1}_J)\mathbf{1}_J^T\boldsymbol{\Sigma}_G^{-1}c\mathbf{1}_J$$
$$= c(\mathbf{1}_J^T\boldsymbol{\Sigma}_G^{-1}\mathbf{1}_J)\left(\hat{\boldsymbol{\beta}}_E^T\boldsymbol{\Sigma}_G^{-1}\mathbf{1}_J - \hat{\boldsymbol{\beta}}_E^T\boldsymbol{\Sigma}_G^{-1}\mathbf{1}_J\right) = 0.$$

The above results taken together gives us:

$$E(\hat{\gamma}_{LDMRE}) \to \gamma \text{ as } N_E \to \infty \text{ and } J_{EFF} \to \infty.$$

We next look at the intercept

$$\hat{\lambda}_{LD} = \frac{(\hat{\boldsymbol{\beta}}_E^T\Sigma_G^{-1}\hat{\boldsymbol{\beta}}_E)(\mathbf{1}_J^T\boldsymbol{\Sigma}_G^{-1}\hat{\boldsymbol{\beta}}_G) - (\mathbf{1}_J^T\boldsymbol{\Sigma}_G^{-1}\hat{\boldsymbol{\beta}}_E)\hat{\boldsymbol{\beta}}_E^T\boldsymbol{\Sigma}_G^{-1}\hat{\boldsymbol{\beta}}_G}{(\mathbf{1}_J^T\boldsymbol{\Sigma}_G^{-1}\mathbf{1}_J)(\hat{\boldsymbol{\beta}}_E^T\boldsymbol{\Sigma}_G^{-1}\hat{\boldsymbol{\beta}}_E) - (\mathbf{1}_J^T\boldsymbol{\Sigma}_G^{-1}\hat{\boldsymbol{\beta}}_E)^2},$$

Take the expectation:

$$E(\hat{\lambda}_{LD}|\hat{\boldsymbol{\beta}}_E) = \gamma\frac{(\hat{\boldsymbol{\beta}}_E^T\Sigma_G^{-1}\hat{\boldsymbol{\beta}}_E)(\mathbf{1}_J^T\boldsymbol{\Sigma}_G^{-1}\boldsymbol{\beta}_E) - (\mathbf{1}_J^T\boldsymbol{\Sigma}_G^{-1}\hat{\boldsymbol{\beta}}_E)\hat{\boldsymbol{\beta}}_E^T\boldsymbol{\Sigma}_G^{-1}\boldsymbol{\beta}_E}{(\mathbf{1}_J^T\boldsymbol{\Sigma}_G^{-1}\mathbf{1}_J)(\hat{\boldsymbol{\beta}}_E^T\boldsymbol{\Sigma}_G^{-1}\hat{\boldsymbol{\beta}}_E) - (\mathbf{1}_J^T\boldsymbol{\Sigma}_G^{-1}\hat{\boldsymbol{\beta}}_E)^2}$$
$$+\frac{(\hat{\boldsymbol{\beta}}_E^T\Sigma_G^{-1}\hat{\boldsymbol{\beta}}_E)(\mathbf{1}_J^T\boldsymbol{\Sigma}_G^{-1}\boldsymbol{\theta}) - (\mathbf{1}_J^T\boldsymbol{\Sigma}_G^{-1}\hat{\boldsymbol{\beta}}_E)\hat{\boldsymbol{\beta}}_E^T\boldsymbol{\Sigma}_G^{-1}\boldsymbol{\theta}}{(\mathbf{1}_J^T\boldsymbol{\Sigma}_G^{-1}\mathbf{1}_J)(\hat{\boldsymbol{\beta}}_E^T\boldsymbol{\Sigma}_G^{-1}\hat{\boldsymbol{\beta}}_E) - (\mathbf{1}_J^T\boldsymbol{\Sigma}_G^{-1}\hat{\boldsymbol{\beta}}_E)^2}.$$

As N_E goes to infinity, the first term will go to 0 as $\hat{\boldsymbol{\beta}}_E \to \boldsymbol{\beta}_E$.

$$E(\hat{\lambda}_{LD}|\hat{\boldsymbol{\beta}}_E) \approx \frac{(\hat{\boldsymbol{\beta}}_E^T\Sigma_G^{-1}\hat{\boldsymbol{\beta}}_E)(\mathbf{1}_J^T\boldsymbol{\Sigma}_G^{-1}\boldsymbol{\theta}) - (\mathbf{1}_J^T\boldsymbol{\Sigma}_G^{-1}\hat{\boldsymbol{\beta}}_E)\hat{\boldsymbol{\beta}}_E^T\boldsymbol{\Sigma}_G^{-1}\boldsymbol{\theta}}{(\mathbf{1}_J^T\boldsymbol{\Sigma}_G^{-1}\mathbf{1}_J)(\hat{\boldsymbol{\beta}}_E^T\boldsymbol{\Sigma}_G^{-1}\hat{\boldsymbol{\beta}}_E) - (\mathbf{1}_J^T\boldsymbol{\Sigma}_G^{-1}\hat{\boldsymbol{\beta}}_E)^2}.$$

Rewriting this, if $\boldsymbol{\theta}$ is independent of $\hat{\boldsymbol{\beta}}_E$ and $\mu_\theta = \mathbf{1}_J c$

$$E_\theta(E(\hat{\lambda}_{LD}|\hat{\boldsymbol{\beta}}_E)|\hat{\boldsymbol{\beta}}_E) \approx c\frac{(\hat{\boldsymbol{\beta}}_E^T\Sigma_G^{-1}\hat{\boldsymbol{\beta}}_E)(\mathbf{1}_J^T\boldsymbol{\Sigma}_G^{-1}\mathbf{1}_J) - (\mathbf{1}_J^T\boldsymbol{\Sigma}_G^{-1}\hat{\boldsymbol{\beta}}_E)\hat{\boldsymbol{\beta}}_E^T\boldsymbol{\Sigma}_G^{-1}\mathbf{1}_J}{(\mathbf{1}_J^T\boldsymbol{\Sigma}_G^{-1}\mathbf{1}_J)(\hat{\boldsymbol{\beta}}_E^T\boldsymbol{\Sigma}_G^{-1}\hat{\boldsymbol{\beta}}_E) - (\mathbf{1}_J^T\boldsymbol{\Sigma}_G^{-1}\hat{\boldsymbol{\beta}}_E)^2} = c.$$

Therefore if the InSIDE condition holds, and the mean is a constant:

$$E(\hat{\lambda}_{LD}|\hat{\boldsymbol{\beta}}_E) \rightarrow c \text{ as } N_E, J_{EFF} \rightarrow \infty.$$

3.10.4 Incorporation of the residual variance

Based on out simulations, the conditional GWAS estimates are (Table 3.3):

$$\hat{\boldsymbol{\beta}}_G = \boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\beta}}^*_G = \boldsymbol{\beta}_G + \boldsymbol{\Sigma}^{-1}L^T\boldsymbol{\epsilon}_G = \boldsymbol{\theta} + \gamma\boldsymbol{\beta}_E + \boldsymbol{\Sigma}^{-1}L^T\boldsymbol{\epsilon}_G.$$

We will focus on the null case ($\gamma = 0$). The true $\boldsymbol{\beta}_G$ is then:

$$\boldsymbol{\beta}_G = \boldsymbol{\theta} = \sigma * \boldsymbol{\theta}_1, \sigma = \sqrt{\frac{h^2_{G\rightarrow Y}}{\boldsymbol{\theta}_1^T\boldsymbol{\Sigma}\boldsymbol{\theta}_1}}.$$

Here, $\boldsymbol{\theta}_1$ denotes $\boldsymbol{\theta}$ before the rescalling in Step 7 (Table 3.3). The $\boldsymbol{\epsilon}_G$ is:

$$\boldsymbol{\epsilon}_G \sim N\left(0, \frac{1 - h^2_E * h^2_{E\rightarrow Y} - h^2_{G\rightarrow Y}}{N}\boldsymbol{I}_J\right).$$

Under the null, the conditional estimate is:

$$\hat{\boldsymbol{\beta}}_G = \boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\beta}}^*_G = \boldsymbol{\beta}_G + \boldsymbol{\Sigma}^{-1}L^T\boldsymbol{\epsilon}_G = \sigma\boldsymbol{\theta}_1 + \boldsymbol{\Sigma}^{-1}L^T\boldsymbol{\epsilon}_G,$$

and it's variance:

$$\boldsymbol{\Sigma}_G = \text{var}(\hat{\boldsymbol{\beta}}_G) = \text{var}(\sigma\boldsymbol{\theta}_1) + \frac{1 - h^2_E * h^2_{E\rightarrow Y} - h^2_{G\rightarrow Y}}{N}\boldsymbol{\Sigma}^{-1}L^TL\boldsymbol{\Sigma}^{-1}.$$

Recall that each simulation is a new draw of $\sigma\boldsymbol{\theta}_1$, it is variable.

$$v(\boldsymbol{\theta}) = \text{var}(\boldsymbol{\theta}) = \text{var}\left(\sqrt{\frac{h^2_{G\rightarrow Y}}{\boldsymbol{\theta}_1^T\boldsymbol{\Sigma}\boldsymbol{\theta}_1}}\boldsymbol{\theta}_1\right).$$

Under the case (Case 1) when there is no direct effect ($h^2_{G\rightarrow Y} = 0$)

$$\boldsymbol{\Sigma}_G = \frac{1 - h^2_E * h^2_{E\rightarrow Y} - h^2_{G\rightarrow Y}}{N}\boldsymbol{\Sigma}^{-1}L^TL\boldsymbol{\Sigma}^{-1} = \frac{1 - h^2_E * h^2_{E\rightarrow Y} - h^2_{G\rightarrow Y}}{N}\boldsymbol{\Sigma}^{-1}.$$

Case 2, when there is a direct effect but it is constant, $v(\boldsymbol{\theta}) \approx 0$:

$$\boldsymbol{\Sigma}_G \approx \frac{1 - h^2_E * h^2_{E\rightarrow Y} - h^2_{G\rightarrow Y}}{N}\boldsymbol{\Sigma}^{-1},$$

and finally (Case 3) when there is a direct effect and it's variable:

$$\boldsymbol{\Sigma}_G = \text{var}(\sigma\boldsymbol{\theta}_1) + \frac{1 - h_E^2 * h_{E\to Y}^2 - h_{G\to Y}^2}{N}\boldsymbol{\Sigma}^{-1}.$$

For the first two scenarios, the residual does not help us at all and will be near 1. For the last it does as it accounts for some of $v(\boldsymbol{\theta})$. We now examine the expected value of the residual variance estimate. In our simulations we do not take into account that $v(\boldsymbol{\theta})$ as it is unobserved. First define the following:

$$\begin{aligned}
W &= \boldsymbol{\Sigma}\frac{N}{1 - h_E^2 * h_{E\to Y}^2 - h_{G\to Y}^2}, \\
\boldsymbol{X} &= \begin{bmatrix} 1 & \hat{\boldsymbol{\beta}}_E \end{bmatrix}, \\
\boldsymbol{\Gamma} &= \begin{bmatrix} \lambda \\ \gamma \end{bmatrix}, \\
\hat{\boldsymbol{\Gamma}} &= (\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}\hat{\boldsymbol{\beta}}_G.
\end{aligned}$$

Under the first two cases $W = \boldsymbol{\Sigma}_G^{-1}$. Based on the above and what the true variance of $\hat{\boldsymbol{\beta}}_G$ is, we have:

$$\text{var}(\hat{\boldsymbol{\Gamma}}|\boldsymbol{X}) = (\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1} + (\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}v(\boldsymbol{\theta})\boldsymbol{W}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}.$$

We estimate the variance as:

$$\begin{aligned}
\hat{\text{var}}(\hat{\boldsymbol{\Gamma}}|X) &= \hat{\sigma}^2(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}, \\
\hat{\sigma^2} &= \frac{1}{J-2}(\hat{\boldsymbol{\beta}}_G - \boldsymbol{X}\hat{\boldsymbol{\Gamma}})^T\boldsymbol{W}(\hat{\boldsymbol{\beta}}_G - \boldsymbol{X}\hat{\boldsymbol{\Gamma}}).
\end{aligned}$$

Define $\boldsymbol{P}_E = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}$. Now if $E(\sigma\boldsymbol{\theta}_1) = 0$:

$$\begin{aligned}
E(\hat{\sigma}^2|\boldsymbol{X}) &= \frac{1}{J-2}\text{trace}(\boldsymbol{\Sigma}_G(\boldsymbol{I}_J - \boldsymbol{P}_E)^T\boldsymbol{W}(\boldsymbol{I}_J - \boldsymbol{P}_E)), \\
E(\hat{\sigma}^2|\boldsymbol{X}) &= 1 + \frac{\text{trace}(v(\boldsymbol{\theta})\boldsymbol{W}(\boldsymbol{I}_J - \boldsymbol{P}_E))}{J-2},
\end{aligned}$$

therefore our estimated variance is:

$$E(\hat{\sigma}^2(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}|\boldsymbol{X}) = (\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1} + \frac{\text{trace}(v(\boldsymbol{\theta})\boldsymbol{W}(\boldsymbol{I}_J - \boldsymbol{P}_E))}{J-2}(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}.$$

The true variance was:

$$\text{var}(\hat{\boldsymbol{\Gamma}}|X) = (\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1} + \boldsymbol{P}_E v(\boldsymbol{\theta})\boldsymbol{P}_E^T.$$

For small LD with variable direct effect, the extra term in our estimated variance helps with the second term in the true variance above.

3.10.5 Incorporation of Multiple Genes

We now incorporate K nearby genes into the analysis jointly. We propose the following:

$$\hat{\beta}_{G,j} = \lambda + \sum_{k=1}^{K} \gamma_k \hat{\beta}_{E,j,k} I(\text{SNP } j \text{ within 500 BP of Gene k}) + \epsilon_{G,j}.$$

We can then estimate each γ_k using a $K+1$ design matrix where the j^{th} row is $(1, \hat{\beta}_{E,j,1} I(\text{SNP j within 500 BP of gene 1}), \cdots, \hat{\beta}_{E,j,K} I(\text{SNP j within 500 BP of gene K}))$:

$$\hat{\boldsymbol{\beta}}_G = \mathbf{1}_J \lambda + \left(\hat{\boldsymbol{\beta}}_{E,K} \cdot P_{E,K}\right) \boldsymbol{\gamma} + \boldsymbol{\epsilon}_G,$$

where $\hat{\boldsymbol{\beta}}_{E,K}$ is a J by K matrix of eQTL effects, and $P_{E,K}$ is J by K indicator matrix of whether SNP j maps to gene k, "$\cdot$" is element wise multiplication between the matrices. $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \cdots, \gamma_K)^T$.

3.10.6 Incorporation of eQTL variance

We have the following distribution of our eQTL and GWAS estimates from two independent samples:

$$\begin{pmatrix} \hat{\boldsymbol{\beta}}_E \\ \hat{\boldsymbol{\beta}}_G \end{pmatrix} \sim N_{2J} \left(\begin{bmatrix} \boldsymbol{\beta}_E \\ \boldsymbol{\beta}_E \gamma + \lambda \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_E & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_G \end{bmatrix} \right),$$

where $\boldsymbol{\Sigma}_E = \text{var}(\hat{\boldsymbol{\beta}}_E)$ and $\boldsymbol{\Sigma}_G$ is defined as in the document. We are maximizing the likelihood:

$$\begin{aligned} \ell(\hat{\boldsymbol{\beta}}_E, \hat{\boldsymbol{\beta}}_G | \boldsymbol{\beta}_E, \gamma, \lambda) \propto & \frac{-1}{2} (\hat{\boldsymbol{\beta}}_E - \boldsymbol{\beta}_E)^T \boldsymbol{\Sigma}_E^{-1} (\hat{\boldsymbol{\beta}}_E - \boldsymbol{\beta}_E) \\ & - \frac{1}{2} (\hat{\boldsymbol{\beta}}_G - \gamma \boldsymbol{\beta}_E - \lambda)^T \boldsymbol{\Sigma}_G^{-1} (\hat{\boldsymbol{\beta}}_G - \gamma \boldsymbol{\beta}_E - \lambda) \end{aligned}$$

We use an iterative process to solve for $\boldsymbol{\beta}_E$, λ, and γ. The score for these parameters is:

$$\begin{aligned} \boldsymbol{U}_{\beta_E} &= \boldsymbol{\Sigma}_E^{-1} (\hat{\boldsymbol{\beta}}_E - \boldsymbol{\beta}_E) + \gamma \boldsymbol{\Sigma}_G^{-1} (\hat{\boldsymbol{\beta}}_G - \gamma \boldsymbol{\beta}_E - \lambda), \\ \boldsymbol{U}_{\gamma,\lambda} &= \begin{bmatrix} \mathbf{1}_J^T \\ \boldsymbol{\beta}_E^T \end{bmatrix} \boldsymbol{\Sigma}_G^{-1} (\hat{\boldsymbol{\beta}}_G - \gamma \boldsymbol{\beta}_E - \lambda). \end{aligned}$$

Given initial estimates of $\hat{\gamma}^{(0)}$ and $\hat{\lambda}^{(0)}$: we set the first value of $\boldsymbol{\beta}_E$ (denote $\tilde{\boldsymbol{\beta}}_E$ to avoid confusion with $\hat{\boldsymbol{\beta}}_E$) as:

$$\tilde{\boldsymbol{\beta}}_E^{(1)} = (\boldsymbol{\Sigma}_E^{-1} + \hat{\gamma}^{(0)} \boldsymbol{\Sigma}_G^{-1})^{-1} (\boldsymbol{\Sigma}_E^{-1} \hat{\boldsymbol{\beta}}_E + \hat{\gamma}^{(0)} \boldsymbol{\Sigma}_G^{-1} (\hat{\boldsymbol{\beta}}_G - \hat{\lambda}^{(0)}))$$

We then iterate back and forth estimating each until convergence at some preset tolerance:

$$\tilde{\boldsymbol{\beta}}_E^{(k)} = (\boldsymbol{\Sigma}_E^{-1} + \hat{\gamma}^{(k-1)}\boldsymbol{\Sigma}_G^{-1})^{-1}(\boldsymbol{\Sigma}_E^{-1}\hat{\boldsymbol{\beta}}_E + \hat{\gamma}^{(k-1)}\boldsymbol{\Sigma}_G^{-1}(\hat{\boldsymbol{\beta}}_G - \hat{\lambda}^{(k-1)})),$$

$$\begin{bmatrix} \hat{\lambda}^{(k)} \\ \hat{\gamma}^{(k)} \end{bmatrix} = \left(\begin{bmatrix} \mathbf{1}_J^T\boldsymbol{\Sigma}_G^{-1}\mathbf{1}_J & \mathbf{1}_J^T\boldsymbol{\Sigma}_G^{-1}\tilde{\boldsymbol{\beta}}_E^{(k)} \\ \tilde{\boldsymbol{\beta}}_E^{T(k)}\boldsymbol{\Sigma}_G^{-1}\mathbf{1}_J & \tilde{\boldsymbol{\beta}}_E^{T(k)}\boldsymbol{\Sigma}_G^{-1}\tilde{\boldsymbol{\beta}}_E^{(k)} \end{bmatrix} \right)^{-1} \begin{bmatrix} \mathbf{1}_J^T \\ \tilde{\boldsymbol{\beta}}_E^{T(k)} \end{bmatrix} \boldsymbol{\Sigma}_G^{-1}\hat{\boldsymbol{\beta}}_G.$$

This will provide potentially more precise estimates but with higher variances as they account for the variance in the eQTL estimates. Work still needs to be done to derive the variances analytically.

LD	J	h^2_E	Method	No Direct effect		Constant Direct Effect	
				$\gamma^2 = 0.005$	$\gamma^2 = 0.01$	$\gamma^2 = 0.005$	$\gamma^2 = 0.02$
			MR	0.058	0.063	—-	—-
			LDA MR	0.054	0.058	—-	—-
		0.01	MRE	0.057	0.062	0.055	0.057
			LDA MRE	0.054	0.059	0.055	0.059
			MR	0.524	0.804	—-	—-
			LDA MR	0.510	0.797	—-	—-
	J=50	0.2	MRE	0.514	0.795	0.302	0.508
			LDA MRE	0.499	0.787	0.498	0.788
			MR	0.922	0.997	—-	—-
			LDA MR	0.921	0.997	—-	—-
		0.5	MRE	0.915	0.997	0.658	0.911
$\rho = 0.125$			LDA MRE	0.914	0.997	0.915	0.997
			MR	0.054	0.057	—-	—-
			LDA MR	0.051	0.053	—-	—-
		0.01	MRE	0.054	0.057	0.052	0.051
			LDA MR-E	0.050	0.052	0.052	0.051
			MR	0.339	0.576	—-	—-
	J=300		LDA MR	0.321	0.560	—-	—-
		0.2	MRE	0.338	0.573	0.288	0.501
			LDA MRE	0.321	0.557	0.325	0.562
			MR	0.877	0.992	—-	—-
			LDA MR	0.870	0.992	—-	—-
		0.5	MRE	0.875	0.992	0.812	0.980
			LDA MRE	0.869	0.991	0.869	0.992

Table 3.4: Power results from simulations evaluated at $\alpha = 0.05$ little LD. MRE stands for MR-Egger

LD	J	h_E^2	Method	No Direct effect		Constant Direct Effect	
				$\gamma^2 = 0.005$	$\gamma^2 = 0.01$	$\gamma^2 = 0.005$	$\gamma^2 = 0.02$
$\rho = 0.9$	J=50	0.01	LDA MR	0.054	0.059	—-	—-
			LDA MRE	0.052	0.055	0.053	0.055
		0.2	LDA MR	0.513	0.798	—-	—-
			LDA MRE	0.389	0.629	0.389	0.637
		0.5	LDA MR	0.919	0.998	—-	—-
			LDA MRE	0.784	0.943	0.788	0.941
	J=300	0.01	LDA MR	0.049	0.049	—-	—-
			LDA MRE	0.050	0.049	0.051	0.051
		0.2	LDA MR	0.318	0.560	—-	—-
			LDA MRE	0.294	0.523	0.300	0.524
		0.5	LDA MR	0.869	0.992	—-	—-
			LDA MRE	0.840	0.985	0.842	0.986

Table 3.5: Power results from simulations evaluated at $\alpha = 0.05$ large LD. MRE stands for MR-Egger

References

AGHA, G., HAJJ, H., RIFAS-SHIMAN, S. L., JUST, A. C., HIVERT, M. F., BURRIS, H. H., LIN, X., LITONJUA, A. A., OKEN, E., DEMEO, D. L., GILLMAN, M. W. and BACCARELLI, A. A. (2016). Birth weight-for-gestational age is associated with DNA methylation at birth and in childhood. *Clin Epigenetics* **8** 118.

AIKIO, M., ELAMAA, H., VICENTE, D., IZZI, V., KAUR, I., SEPPINEN, L., SPEEDY, H. E., KAMINSKA, D., KUUSISTO, S., SORMUNEN, R., HELJASVAARA, R., JONES, E. L., MUILU, M., JAUHIAINEN, M., PIHLAJAMAKI, J., SAVOLAINEN, M. J., SHOULDERS, C. C. and PIHLAJANIEMI, T. (2014). Specific collagen XVIII isoforms promote adipose tissue accrual via mechanisms determining adipocyte number and affect fat deposition. *Proc. Natl. Acad. Sci. U.S.A.* **111** E3043–3052.

AMOS, C. I., DENNIS, J., WANG, Z., BYUN, J., SCHUMACHER, F. R., GAYTHER, S. A., CASEY, G., HUNTER, D. J. ET AL. (2017). The oncoarray consortium: A network for understanding the genetic architecture of common cancers. *Cancer Epidemiol Biomarkers Prev* **26** 126–135.

BARBEIRA, A., SHAH, K. P., TORRES, J. M., WHEELER, H. E., TORSTENSON, E. S., EDWARDS, T., GARCIA, T., BELL, G. I., NICOLAE, D., COX, N. J. and IM, H. K. (2016). Metaxcan: Summary statistics based gene-level association method infers accurate predixcan results. *bioRxiv* .

BARFIELD, R. T., ALMLI, L. M., KILARU, V., SMITH, A. K., MERCER, K. B., DUNCAN, R., KLENGEL, T., MEHTA, D., BINDER, E. B., EPSTEIN, M. P., RESSLER, K. J. and CONNEELY, K. N. (2014). Accounting for population stratification in DNA methylation studies. *Genet. Epidemiol.* **38** 231–241.

BARON, R. M. and KENNY, D. A. (1986). The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *J Pers Soc Psychol* **51** 1173–1182.

BELL, B., ROSE, C. L. and DAMON, A. (1972). The Normative Aging Study: an interdisciplinary and longitudinal study of health and aging. *The International Journal of Aging and Human Development* **3** 5–17.

BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)* **57** 289–300.

BOWDEN, J., DAVEY SMITH, G. and BURGESS, S. (2015). Mendelian randomization with invalid instruments: effect estimation and bias detection through egger regression. *Int J Epidemiol* **44** 512–25.

BREEZE, C. E., PAUL, D. S., VAN DONGEN, J., BUTCHER, L. M., AMBROSE, J. C., BARRETT, J. E., LOWE, R., RAKYAN, V. K., IOTCHKOVA, V., FRONTINI, M., DOWNES, K., OUWEHAND, W. H., LAPERLE, J., JACQUES, P. E., BOURQUE, G., BERGMANN, A. K., SIEBERT, R., VELLENGA, E., SAEED, S., MATARESE, F., MARTENS, J. H., STUNNENBERG, H. G., TESCHENDORFF, A. E., HERRERO, J., BIRNEY, E., DUNHAM, I. and BECK, S. (2016). eFORGE: a tool for identifying cell type-specific signal in epigenomic data. *Cell Rep* **17** 2137–2150.

BURGESS, S., DUDBRIDGE, F. and THOMPSON, S. G. (2016). Combining information on multiple instrumental variables in mendelian randomization: comparison of allele score and summarized data methods. *Stat Med* **35** 1880–906.

CHEN, Y. A., LEMIRE, M., CHOUFANI, S., BUTCHER, D. T., GRAFODATSKAYA, D., ZANKE, B. W., GALLINGER, S., HUDSON, T. J. and WEKSBERG, R. (2013). Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics* **8** 203–209.

CUMMINGS, M., MERONE, L., KEEBLE, C., BURLAND, L., GRZELINSKI, M., SUTTON, K., BEGUM, N., THACOOR, A., GREEN, B., SARVESWARAN, J., HUTSON, R. and ORSI, N. M. (2015). Preoperative neutrophil:lymphocyte and platelet:lymphocyte ratios predict endometrial cancer survival. *Br. J. Cancer* **113** 311–320.

DE JONG, R. A., LEFFERS, N., BOEZEN, H. M., TEN HOOR, K. A., VAN DER ZEE, A. G., HOLLEMA, H. and NIJMAN, H. W. (2009). Presence of tumor-infiltrating lymphocytes is an independent prognostic factor in type I and II endometrial cancer. *Gynecol. Oncol.* **114** 105–110.

DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)* **39** 1–38.

DU, P., ZHANG, X., HUANG, C.-C., JAFARI, N., KIBBE, W. A., HOU, L. and LIN, S. M. (2010). Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC bioinformatics* **11** 587.

ENDERS, C. K., FAIRCHILD, A. J. and MACKINNON, D. P. (2013). A Bayesian Approach for Estimating Mediation Effects with Missing Data. *Multivariate Behav Res* **48** 340–369.

GAMAZON, E. R., WHEELER, H. E., SHAH, K. P., MOZAFFARI, S. V., AQUINO-MICHAELS, K., CARROLL, R. J., EYLER, A. E., DENNY, J. C., CONSORTIUM, G. T., NICOLAE, D. L., COX, N. J. and IM, H. K. (2015). A gene-based association method for mapping traits using reference transcriptome data. *Nat Genet* **47** 1091–8.

GLUBB, D. M., MARANIAN, M. J., MICHAILIDOU, K., POOLEY, K. A., MEYER, K. B., KAR, S., CARLEBUR, S., O'REILLY, M., BETTS, J. A., HILLMAN, K. M., KAUFMANN, S., BEESLEY, J., CANISIUS, S., HOPPER, J. L., SOUTHEY, M. C., TSIMIKLIS, H., ET AL. (2015). Fine-scale mapping of the 5q11.2 breast cancer locus reveals at least three independent risk variants regulating map3k1. *Am J Hum Genet* **96** 5–20.

GUSEV, A., KO, A., SHI, H., BHATIA, G., CHUNG, W., PENNINX, B. W., JANSEN, R., DE GEUS, E. J., BOOMSMA, D. I., WRIGHT, F. A., SULLIVAN, P. F., NIKKOLA, E.,

ALVAREZ, M., CIVELEK, M., LUSIS, A. J., LEHTIMAKI, T., RAITOHARJU, E., KAHONEN, M., SEPPALA, I., RAITAKARI, O. T., KUUSISTO, J., LAAKSO, M., PRICE, A. L., PAJUKANTA, P. and PASANIUC, B. (2016a). Integrative approaches for large-scale transcriptome-wide association studies. *Nat Genet* **48** 245–52.

GUSEV, A., MANCUSO, N., FINUCANE, H. K., RESHEF, Y., SONG, L., SAFI, A., OH, E., MCCAROLL, S., NEALE, B., OPHOFF, R., O'DONOVAN, M. C., KATSANIS, N., CRAWFORD, G. E., SULLIVAN, P. F., PASANIUC, B. and PRICE, A. L. (2016b). Transcriptome-wide association study of schizophrenia and chromatin activity yields mechanistic disease insights. *bioRxiv* .

HORVATH, S. (2013). DNA methylation age of human tissues and cell types. *Genome biology* **14** 1–20.

HOUSEMAN, E. A., ACCOMANDO, W. P., KOESTLER, D. C., CHRISTENSEN, B. C., MARSIT, C. J., NELSON, H. H., WIENCKE, J. K. and KELSEY, K. T. (2012). DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC bioinformatics* **13** 1.

HOUSEMAN, E. A., MOLITOR, J. and MARSIT, C. J. (2014). Reference-free cell mixture adjustments in analysis of DNA methylation data. *Bioinformatics* **30** 1431–1439.

JAFFE, A. E. and IRIZARRY, R. A. (2014). Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome Biol* **15** R31.

JAFFE, A. E., MURAKAMI, P., LEE, H., LEEK, J. T., FALLIN, M. D., FEINBERG, A. P. and IRIZARRY, R. A. (2012). Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. *International journal of epidemiology* **41** 200–209.

JOHNSON, T. (2012). Efficient calculation for multi-snp genetic risk scores.

JOHNSON, W. E., LI, C. and RABINOVIC, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8** 118–127.

KUCZMARSKI, R. J., OGDEN, C. L., GRUMMER-STRAWN, L. M., FLEGAL, K. M., GUO,

S. S., WEI, R., MEI, Z., CURTIN, L. R., ROCHE, A. F. and JOHNSON, C. L. (2000). CDC growth charts: United States. *Adv Data* 1–27.

KUPERS, L. K., XU, X., JANKIPERSADSING, S. A., VAEZ, A., LA BASTIDE-VAN GEMERT, S., SCHOLTENS, S., NOLTE, I. M., RICHMOND, R. C., RELTON, C. L., FELIX, J. F., DUIJTS, L., VAN MEURS, J. B., TIEMEIER, H., JADDOE, V. W., WANG, X., CORPELEIJN, E. and SNIEDER, H. (2015). DNA methylation mediates the effect of maternal smoking during pregnancy on birthweight of the offspring. *Int J Epidemiol* **44** 1224–1237.

LEEK, J. T. and STOREY, J. D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet* **3** e161.

LIANG, K.-Y. and ZEGER, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73** 13–22.

LIANG, K.-Y., ZEGER, S. L. and QAQISH, B. (1992). Multivariate regression analyses for categorical data. *Journal of the Royal Statistical Society. Series B (Methodological)* **54** 3–40.

LIU, Y., ARYEE, M. J., PADYUKOV, L., FALLIN, M. D., HESSELBERG, E., RUNARSSON, A., REINIUS, L., ACEVEDO, N., TAUB, M., RONNINGER, M., SHCHETYNSKY, K., SCHEYNIUS, A., KERE, J., ALFREDSSON, L., KLARESKOG, L., EKSTROM, T. J. and FEINBERG, A. P. (2013). Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nat. Biotechnol.* **31** 142–147.

LIU, Z. and LIN, X. (2017). Testing of mediation effect in genome-wide epigenetic studies. *Manuscript* .

LOCKE, A. E., KAHALI, B., BERNDT, S. I., JUSTICE, A. E. ET AL. (2015). Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518** 197–206.

LOKK, K., MODHUKUR, V., RAJASHEKAR, B., MARTENS, K., MAGI, R., KOLDE, R., KOLTSINA, M., NILSSON, T. K., VILO, J., SALUMETS, A. ET AL. (2014). DNA methylome profiling of human tissues identifies global and tissue-specific methylation patterns. *Genome Biol* **15** r54.

LONG, J. S. and ERVIN, L. H. (2000). Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician* **54** 217–224.

LONSDALE, J., THOMAS, J., SALVATORE, M., PHILLIPS, R., LO, E., SHAD, S., HASZ, R., WALTERS, G., GARCIA, F. ET AL. (2013). The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45** 580–585.

LOUIS, T. A. (1982). Finding the observed information matrix when using the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)* **44** 226–233.

MA, B., WILKER, E. H., WILLIS-OWEN, S. A., BYUN, H.-M., WONG, K. C., MOTTA, V., BACCARELLI, A. A., SCHWARTZ, J., COOKSON, W. O., KHABBAZ, K. ET AL. (2014). Predicting DNA methylation level across human tissues. *Nucleic acids research* **42** 3515–3528.

MACKINNON, D. P., LOCKWOOD, C. M., HOFFMAN, J. M., WEST, S. G. and SHEETS, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychol Methods* **7** 83–104.

MACKINNON, J. G. and WHITE, H. (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics* **29** 305 – 325.

MANCL, L. A. and DEROUEN, T. A. (2001). A covariance estimator for GEE with improved small-sample properties. *Biometrics* **57** 126–134.

MCGREGOR, K., BERNATSKY, S., COLMEGNA, I., HUDSON, M., PASTINEN, T., LABBE, A. and GREENWOOD, C. M. (2016). An evaluation of methods correcting for cell-type heterogeneity in DNA methylation studies. *Genome Biol.* **17** 84.

MEISSNER, A., MIKKELSEN, T. S., GU, H., WERNIG, M., HANNA, J., SIVACHENKO, A., ZHANG, X., BERNSTEIN, B. E., NUSBAUM, C., JAFFE, D. B., GNIRKE, A., JAENISCH, R. and LANDER, E. S. (2008). Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* **454** 766–770.

MENDELSON, M. M., MARIONI, R. E., JOEHANES, R. and OTHERS. (2017). Association of body mass index with dna methylation and gene expression in blood cells and relations to cardiometabolic disease: A mendelian randomization approach. *PLOS Medicine* **14** 1–30.

MICHAILIDOU, K., LINDSTRM, S., DENNIS, J., BEESLEY, J., HUI, S., KAR, S., LEMAON, A., ET AL. (2017). Large-scale genetic association analysis identifies 65 new breast cancer susceptibility loci and predicts target genes. *In Prep* .

MONTAÑO, C. M., IRIZARRY, R. A., KAUFMANN, W. E., TALBOT, K., GUR, R. E., FEINBERG, A. P. and TAUB, M. A. (2013). Measuring cell-type specific differential methylation in human brain tissue. *Genome biology* **14** 1.

OKEN, E., BACCARELLI, A. A., GOLD, D. R., KLEINMAN, K. P., LITONJUA, A. A., DE MEO, D., RICH-EDWARDS, J. W., RIFAS-SHIMAN, S. L., SAGIV, S., TAVERAS, E. M., WEISS, S. T., BELFORT, M. B., BURRIS, H. H., CAMARGO, C. A., HUH, S. Y., MANTZOROS, C., PARKER, M. G. and GILLMAN, M. W. (2015). Cohort profile: project viva. *Int J Epidemiol* **44** 37–48.

PAN, W. and WALL, M. M. (2002). Small-sample adjustments in using the sandwich variance estimator in generalized estimating equations. *Statistics in medicine* **21** 1429–1441.

PANNI, T., MEHTA, A. J., SCHWARTZ, J. D., BACCARELLI, A. A., JUST, A. C., WOLF, K., WAHL, S., CYRYS, J., KUNZE, S., STRAUCH, K. ET AL. (2016). A genome-wide analysis of DNA methylation and fine particulate matter air pollution in three study populations: KORA F3, KORA F4, and the Normative Aging Study. *Environmental health perspectives* **124** 983?990.

PASANIUC, B., ZAITLEN, N., SHI, H., BHATIA, G., GUSEV, A., PICKRELL, J., HIRSCHHORN, J., STRACHAN, D. P., PATTERSON, N. and PRICE, A. L. (2014). Fast and accurate imputation of summary statistics enhances evidence of functional enrichment. *Bioinformatics* **30** 2906–14.

PELOSO, G. M., AUER, P. L., BIS, J. C., VOORMAN, A., MORRISON, A. C., STITZIEL, N. O. ET AL. (2014). Association of low-frequency and rare coding-sequence variants with blood lipids and coronary heart disease in 56,000 whites and blacks. *Am. J. Hum. Genet.* **94** 223–232.

PETERS, T. J., BUCKLEY, M. J., STATHAM, A. L., PIDSLEY, R., SAMARAS, K., LORD, R. V., CLARK, S. J. and MOLLOY, P. L. (2015). De novo identification of differentially methylated regions in the human genome. *Epigenetics & chromatin* **8** 1.

PHIPSON, B., LEE, S., MAJEWSKI, I. J., ALEXANDER, W. S. and SMYTH, G. K. (2016). Robust hyperparameter estimation protects against hypervariable genes and improves power to detect differential expression. *ArXiv e-prints* .

RAHMANI, E., ZAITLEN, N., BARAN, Y., ENG, C., HU, D., GALANTER, J., OH, S., BURCHARD, E. G., ESKIN, E., ZOU, J. and HALPERIN, E. (2016). Sparse PCA corrects for cell type heterogeneity in epigenome-wide association studies. *Nat. Methods* **13** 443–445.

SCHULTZ, M. D., HE, Y., WHITAKER, J. W., HARIHARAN, M., MUKAMEL, E. A., LEUNG, D., RAJAGOPAL, N., NERY, J. R., URICH, M. A., CHEN, H., LIN, S., LIN, Y., JUNG, I., SCHMITT, A. D., SELVARAJ, S., REN, B., SEJNOWSKI, T. J., WANG, W. and ECKER, J. R. (2015). Human body epigenome maps reveal noncanonical DNA methylation variation. *Nature* **523** 212–216.

SHARP, G. C., LAWLOR, D. A., RICHMOND, R. C., FRASER, A., SIMPKIN, A., SUDERMAN, M., SHIHAB, H. A., LYTTLETON, O., MCARDLE, W., RING, S. M., GAUNT, T. R., DAVEY SMITH, G. and RELTON, C. L. (2015). Maternal pre-pregnancy BMI and gestational weight gain, offspring DNA methylation and later offspring adiposity: findings from the Avon Longitudinal Study of Parents and Children. *Int J Epidemiol* **44** 1288–1304.

SIEGMUND, K. D. (2011). Statistical approaches for the analysis of DNA methylation microarray data. *Human genetics* **129** 585–595.

SMITH, Z. D. and MEISSNER, A. (2013). DNA methylation: roles in mammalian development. *Nat. Rev. Genet.* **14** 204–220.

SMYTH, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* **3** Article3.

STOREY, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **64** 479–498.

SU, D., WANG, X., CAMPBELL, M. R., PORTER, D. K., PITTMAN, G. S., BENNETT, B. D., WAN, M., ENGLERT, N. A., CROWL, C. L., GIMPLE, R. N., ADAMSKI, K. N., HUANG, Z., MURPHY, S. K. and BELL, D. A. (2016). Distinct Epigenetic Effects of Tobacco Smoking in Whole Blood and among Leukocyte Subtypes. *PLoS ONE* **11** e0166486.

SUN, R., CARROLL, R. J., CHRISTIANI, D. C. and LIN, X. (2017). Testing for gene-environment interaction under environment misspecification. *Manuscript* .

TSAPROUNI, L. G., YANG, T.-P., BELL, J., DICK, K. J., KANONI, S., NISBET, J., VIÑUELA, A., GRUNDBERG, E., NELSON, C. P., MEDURI, E. ET AL. (2014). Cigarette smoking reduces DNA methylation levels at multiple genomic loci but the effect is partially reversible upon cessation. *Epigenetics* **9** 1382–1396.

VALERI, L. and VANDERWEELE, T. J. (2013). Mediation analysis allowing for exposure-mediator interactions and causal interpretation: theoretical assumptions and implementation with sas and spss macros. *Psychol Methods* **18** 137–50.

VANDERWEELE, T. and VANSTEELANDT, S. (2009). Conceptual issues concerning mediation, interventions and composition. *Statistics and Its Interface* **2** 457–468.

WHO (2010). Software for assessing growth and development of the world's children. http://www.who.int/childgrowth/software/en/.

WU, W. and JIA, F. (2013). A New Procedure to Test Mediation With Missing Data Through Nonparametric Bootstrapping and Multiple Imputation. *Multivariate Behav Res* **48** 663–691.

YANG, J., LEE, S. H., GODDARD, M. E. and VISSCHER, P. M. (2011). Gcta: a tool for genome-wide complex trait analysis. *Am J Hum Genet* **88** 76–82.

ZEILINGER, S., KÜHNEL, B., KLOPP, N., BAURECHT, H., KLEINSCHMIDT, A., GIEGER, C., WEIDINGER, S., LATTKA, E., ADAMSKI, J., PETERS, A. ET AL. (2013). Tobacco smoking leads to extensive genome-wide changes in DNA methylation. *PloS one* **8** e63812.

ZHANG, H., ZHENG, Y., ZHANG, Z., GAO, T., JOYCE, B., YOON, G., ZHANG, W., SCHWARTZ, J., JUST, A., COLICINO, E., VOKONAS, P., ZHAO, L., LV, J., BACCARELLI, A., HOU, L. and LIU, L. (2016). Estimating and testing high-dimensional mediation effects in epigenetic studies. *Bioinformatics* **32** 3150–3154.

ZHANG, Y., WILSON, R., HEISS, J., BREITLING, L. P., SAUM, K. U., SCHOTTKER, B., HOLLECZEK, B., WALDENBERGER, M., PETERS, A. and BRENNER, H. (2017). DNA methylation signatures in peripheral blood strongly predict all-cause mortality. *Nat Commun* **8** 14617.

ZHANG, Z. and WANG, L. (2013). Methods for mediation analysis with missing data. *Psychometrika* **78** 154–184.

ZHOU, X., CARBONETTO, P. and STEPHENS, M. (2013). Polygenic modeling with bayesian sparse linear mixed models. *PLoS Genet* **9** e1003264.

ZHU, Z., ZHANG, F., HU, H., BAKSHI, A., ROBINSON, M. R., POWELL, J. E., MONTGOMERY, G. W., GODDARD, M. E., WRAY, N. R., VISSCHER, P. M. and YANG, J. (2016). Integration of summary data from gwas and eqtl studies predicts complex trait gene targets. *Nat Genet* **48** 481–7.

ZOU, J., LIPPERT, C., HECKERMAN, D., ARYEE, M. and LISTGARTEN, J. (2014). Epigenome-wide association studies without the need for cell-type composition. *Nat Methods* **11** 309–11.

Supplementary Tables and Figures

4.1 Supplementary Material Paper 1

4.1.1 Supplementary Figures

Efficiency when MCAR

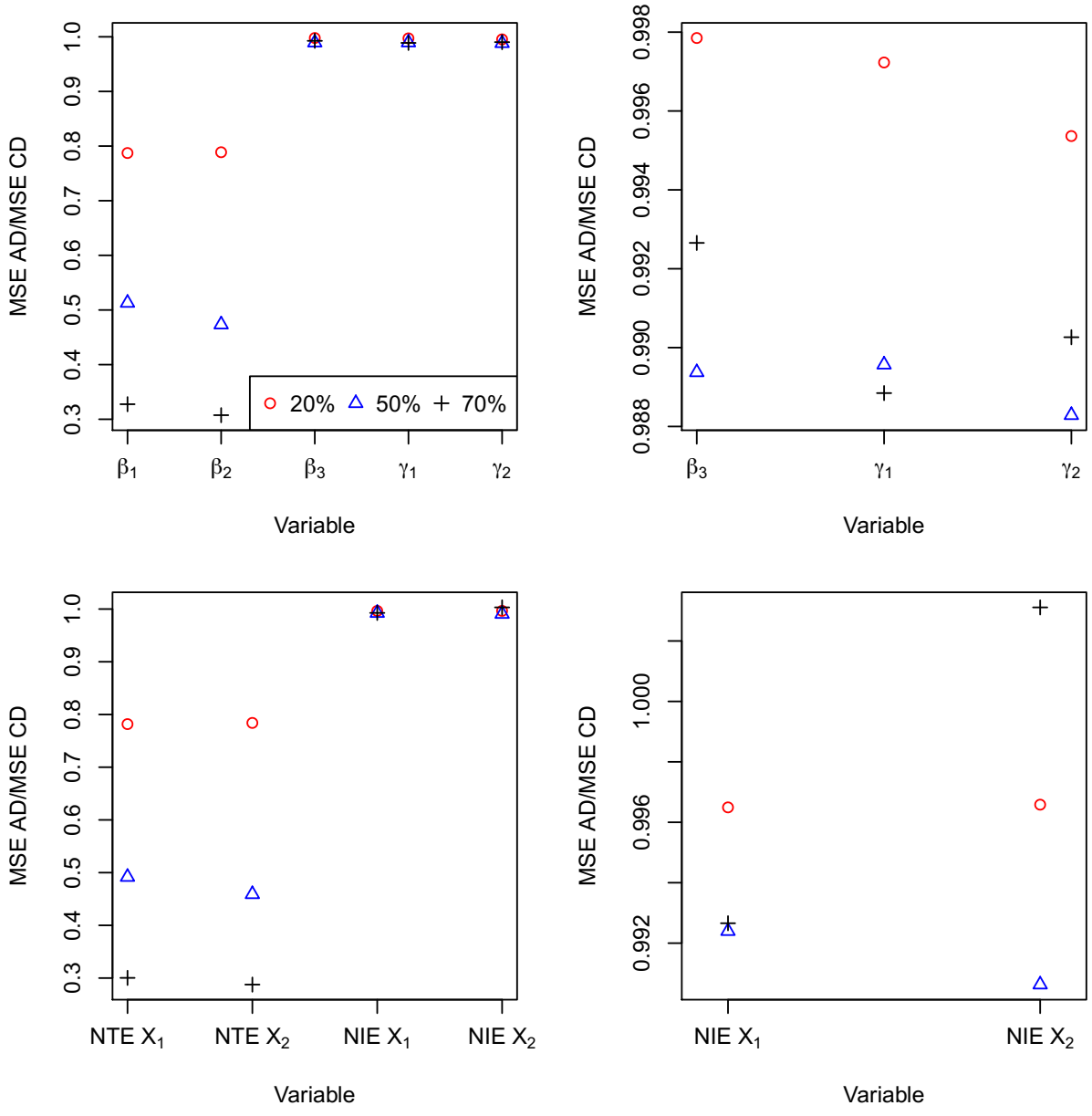

Figure 4.1: Efficiency when MCAR over 10^4 simulations comparing when use all available data vs just complete data.Top panel shows the efficiency for the parameters $\beta_1, \beta_2, \beta_3, \gamma_1$ and γ_2. Bottom panel for NTE and NIE. Right column of plots shows efficiency for β_3,γ_1,γ_2, and NIE's. X_1 is continuous and X_2 is categorical.

Global type I error simulations, MAR other level of missingness

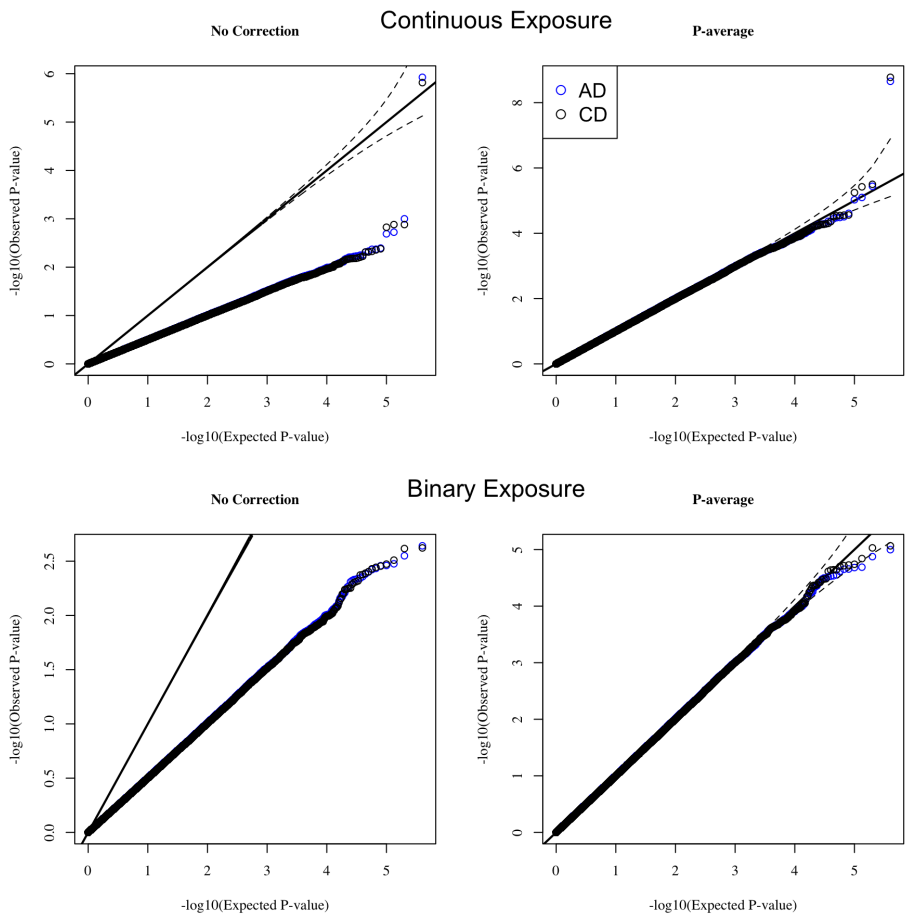

Figure 4.2: Results from genome wide type I error simulations when 20% missing and MAR. Done over 4x10^5 replications where .035% are in Case 1, .035% in Case 2 and the remaining in Case 3. Plots to the left shows p_{NIE}, plots to the right shows $p_{average}$. AD: All available data. CD: Just complete data.

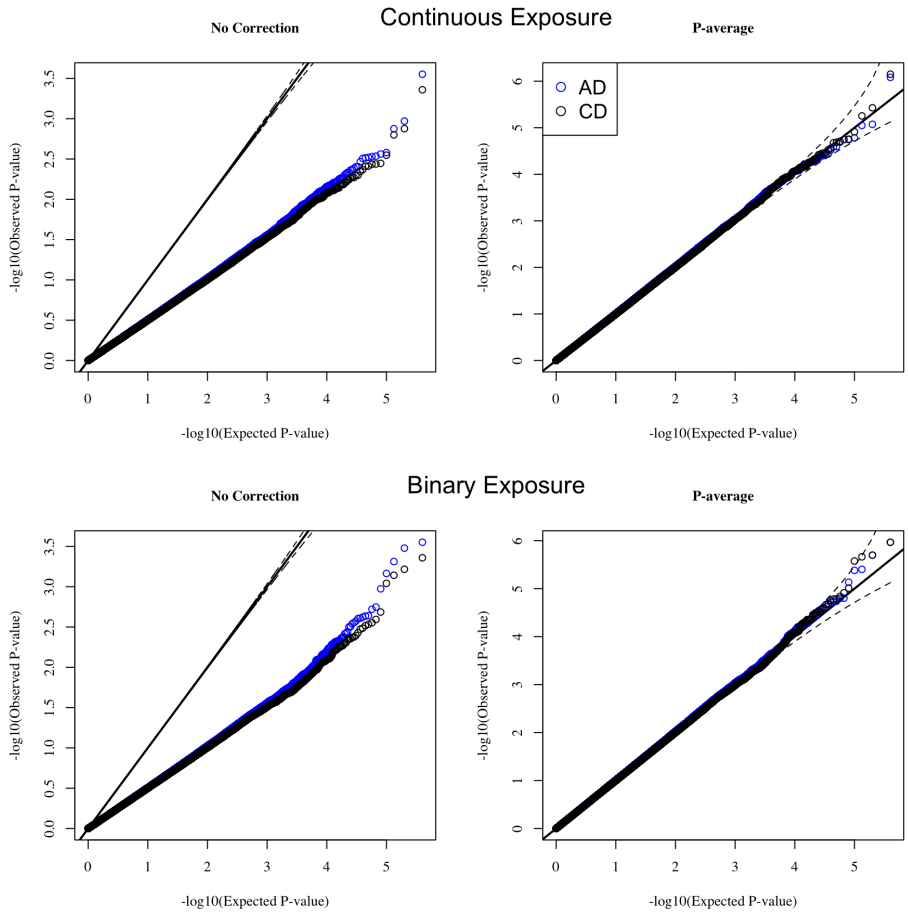

Figure 4.3: Results from genome wide type I error simulations when 70% missing and MAR. Done over 4x10^5 replications where .035% are in Case 1, .035% in Case 2 and the remaining in Case 3. Plots to the left shows p_{NIE}, plots to the right shows $p_{average}$. AD: All available data. CD: Just complete data.

Global type I error simulations, MCAR

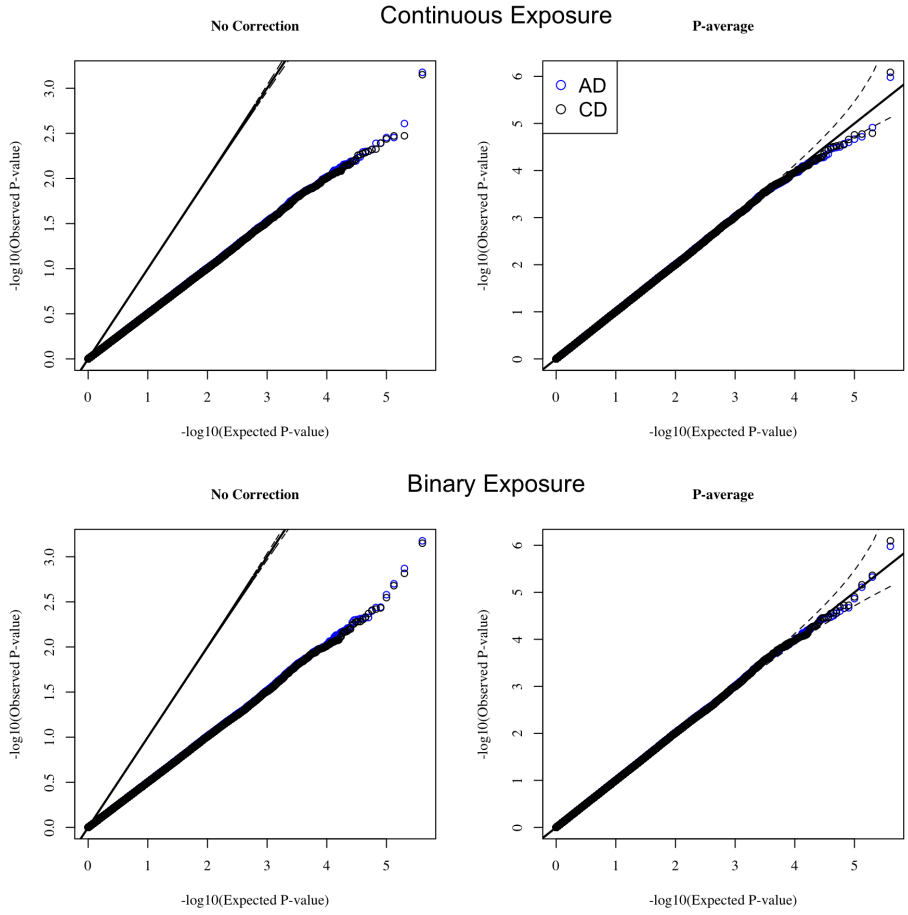

Figure 4.4: Results from genome wide type I error simulations when 20% missing and MCAR. Done over 4x10^5 replications where .035% are in Case 1, .035% in Case 2 and the remaining in Case 3. Plots to the left shows p_{NIE}, plots to the right shows $p_{average}$. AD: All available data. CD: Just complete data.

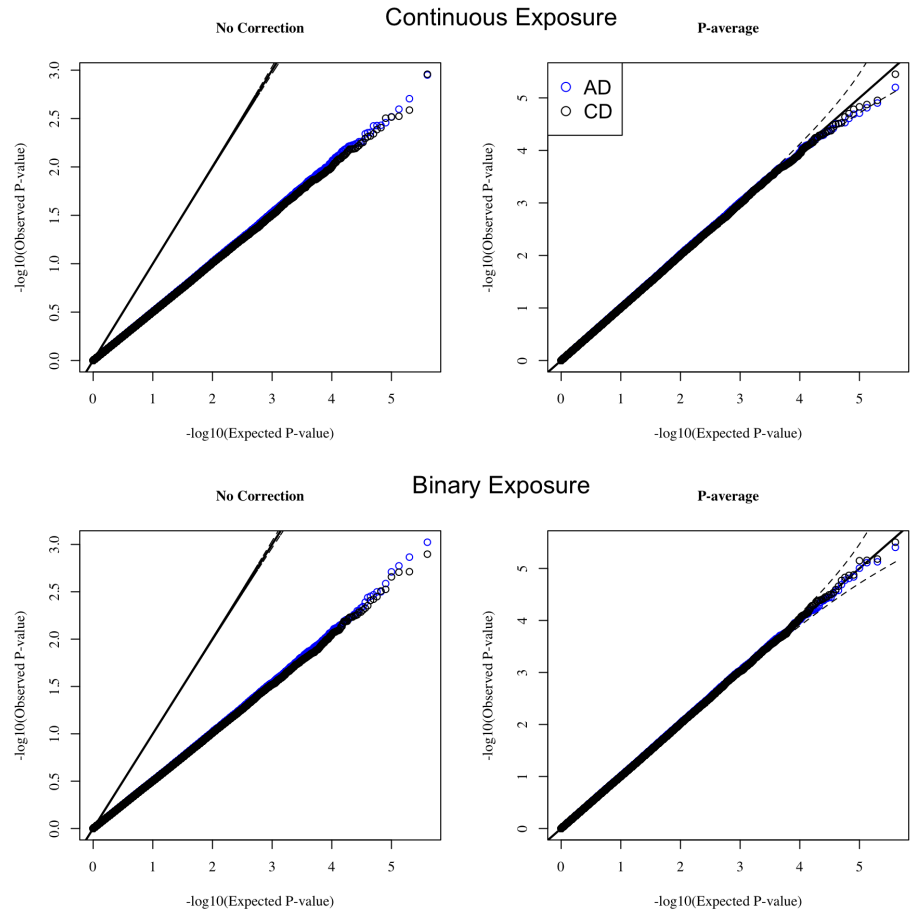

Figure 4.5: Results from genome wide type I error simulations when 50% missing and MCAR. Done over 4x10^5 replications where .035% are in Case 1, .035% in Case 2 and the remaining in Case 3. Plots to the left shows p_{NIE}, plots to the right shows $p_{average}$. AD: All available data. CD: Just complete data.

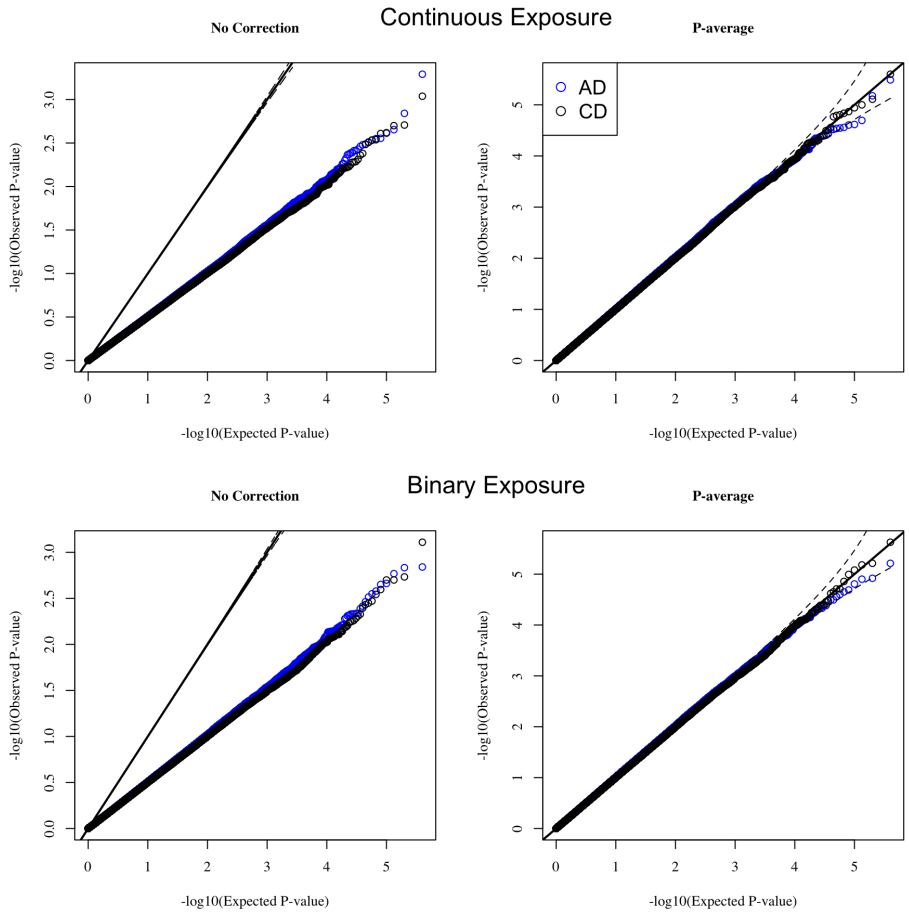

Figure 4.6: Results from genome wide type I error simulations when 70% missing and MCAR. Done over 4x10^5 replications where .035% are in Case 1, .035% in Case 2 and the remaining in Case 3. Plots to the left shows p_{NIE}, plots to the right shows $p_{average}$. AD: All available data. CD: Just complete data.

Power simulations, MCAR

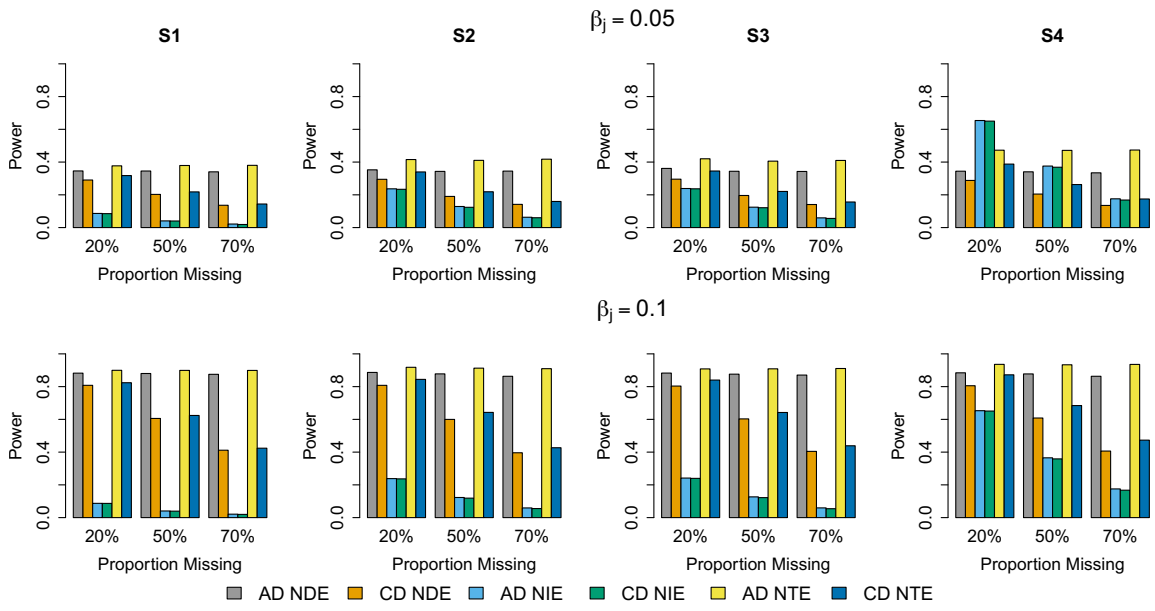

Figure 4.7: Power for continuous variable when MCAR, evaluated at $\alpha = 0.05$. Top panel shows when $\beta_1 = 0.05$, bottom panel when $\beta_1 = 0.1$. Moving from left to right: S1) $\gamma_1 = \beta_3 = 0.05$, S2) $\gamma_1 = 0.1, \beta_3 = 0.05$, S3) $\gamma_1 = 0.05, \beta_3 = 0.1$ and S4) $\gamma_1 = \beta_3 = 0.1$. AD: All available data. CD: Just complete data.

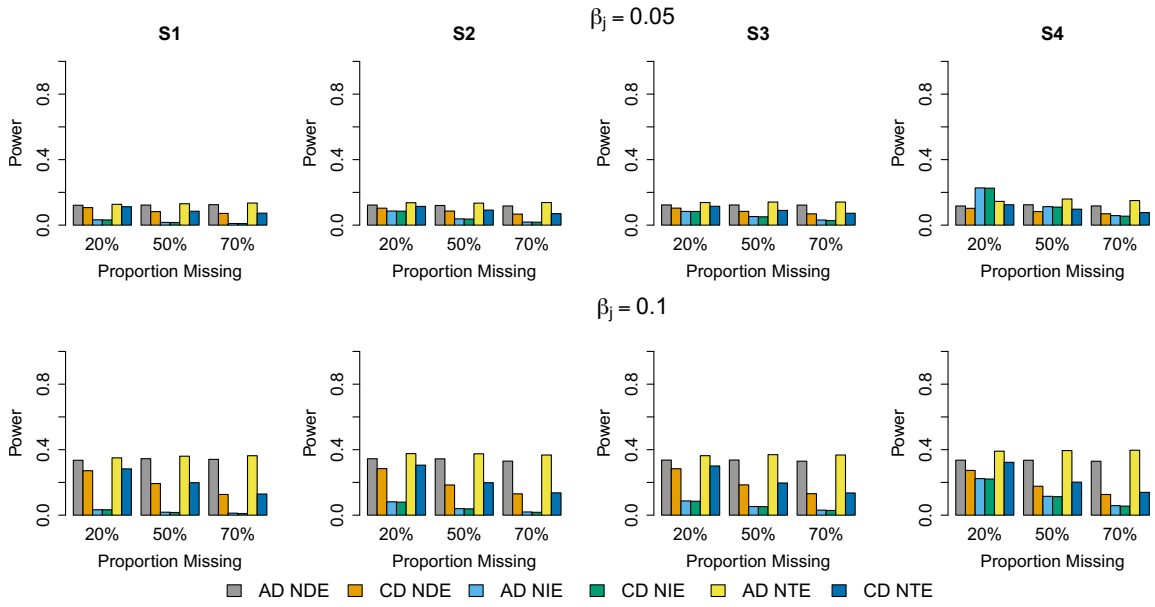

Figure 4.8: Power for categorical variable when MCAR, evaluated at $\alpha = 0.05$. Top panel shows when $\beta_2 = 0.05$, bottom panel when $\beta_2 = 0.1$. Moving from left to right: S1) $\gamma_1 = \beta_3 = 0.05$, S2) $\gamma_1 = 0.1, \beta_3 = 0.05$, S3) $\gamma_1 = 0.05, \beta_3 = 0.1$ and S4) $\gamma_1 = \beta_3 = 0.1$. AD: All available data. CD: Just complete data.

Association specific qq-plots in Project Viva

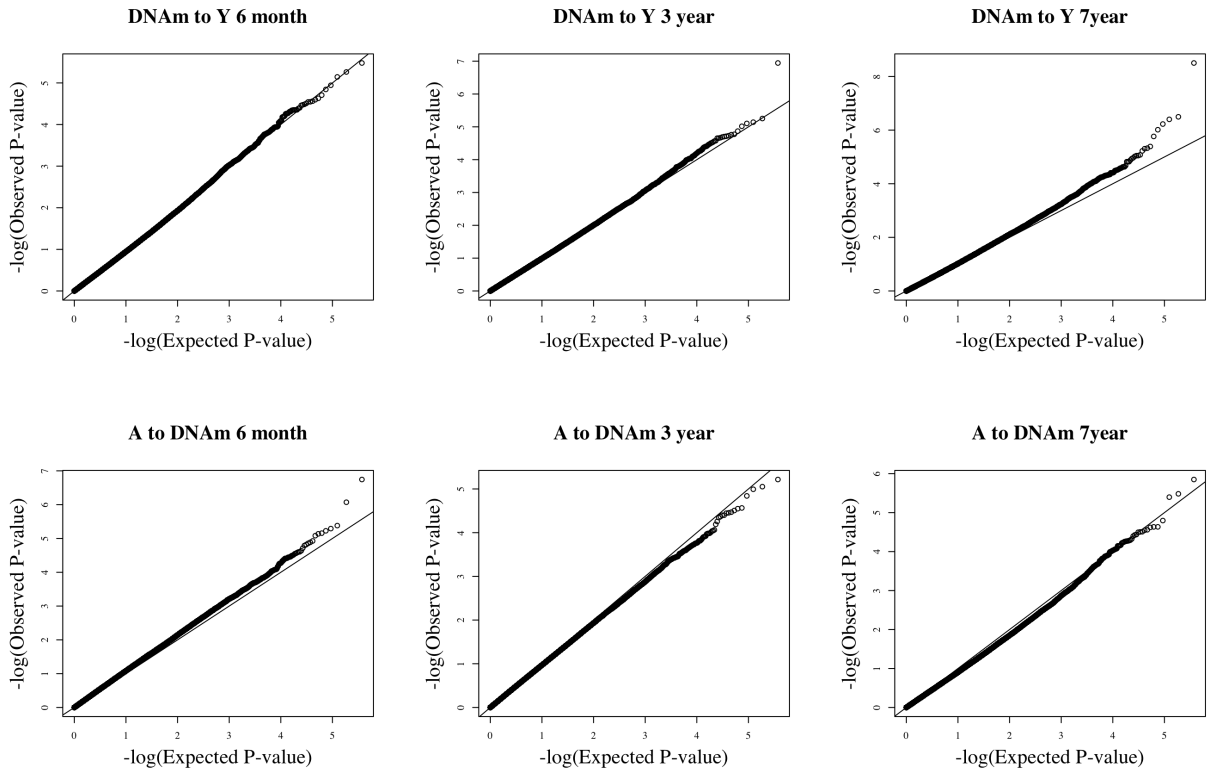

Figure 4.9: Results from analysis in Project Viva restricted to white mother-child pairs. First panel shows test for $\beta_M = 0$, second panel for $\gamma_A = 0$. From left to right: at 6 months, 3 years, and 7 years.

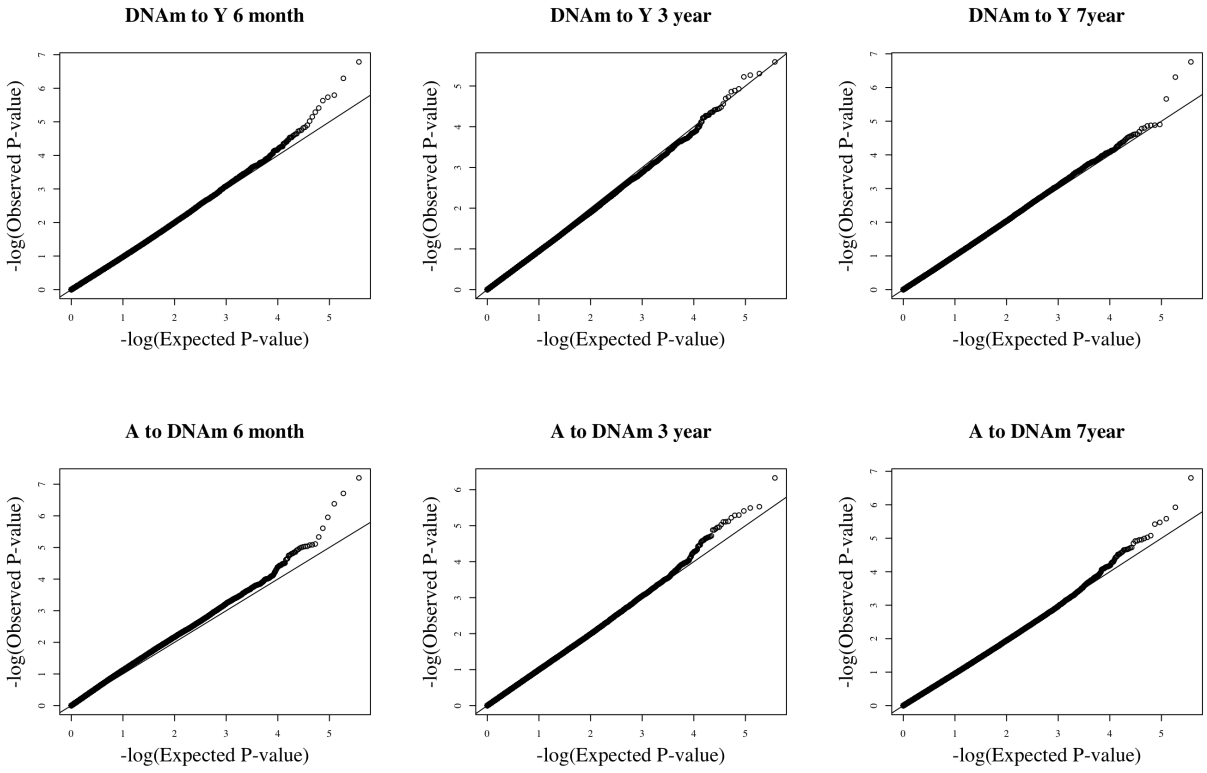

Figure 4.10: Results from analysis in Project Viva using every mother-child pair. First panel shows test for $\beta_M = 0$, second panel for $\gamma_A = 0$. From left to right: at 6 months, 3 years, and 7 years.

4.1.2 Supplementary Tables

Bias results when MCAR

Table 4.1: Mean estimate, empirical variance, and 95% coverage when MCAR for 10^4 simulations.

		All Available Data			Just Complete data		
Parameter	% Missing	Mean	Var	95% CI	Mean	Var	95% CI
$\underline{\beta_0}$	20%	0.200	0.002	0.952	0.200	0.002	0.950
	50%	0.200	0.002	0.949	0.199	0.003	0.948
	70%	0.200	0.002	0.953	0.202	0.006	0.948
$\underline{\beta_1}$	20%	0.140	0.001	0.950	0.140	0.001	0.950
	50%	0.140	0.001	0.952	0.140	0.002	0.950
	70%	0.139	0.001	0.948	0.140	0.003	0.953
$\underline{\beta_2}$	20%	0.140	0.004	0.951	0.140	0.005	0.949
	50%	0.140	0.004	0.949	0.140	0.009	0.950
	70%	0.139	0.004	0.953	0.137	0.014	0.948
$\underline{\beta_3}$	20%	0.140	0.001	0.951	0.140	0.001	0.951
	50%	0.141	0.002	0.947	0.141	0.002	0.949
	70%	0.140	0.003	0.942	0.139	0.003	0.947
$\underline{\gamma_0}$	20%	0.200	0.002	0.951	0.200	0.002	0.952
	50%	0.200	0.003	0.950	0.200	0.003	0.952
	70%	0.199	0.006	0.946	0.199	0.006	0.950
$\underline{\gamma_1}$	20%	0.140	0.001	0.946	0.140	0.001	0.947
	50%	0.141	0.002	0.947	0.141	0.002	0.949
	70%	0.141	0.003	0.943	0.141	0.003	0.946
$\underline{\gamma_2}$	20%	0.140	0.005	0.948	0.140	0.005	0.950
	50%	0.140	0.008	0.951	0.140	0.008	0.951
	70%	0.142	0.014	0.947	0.142	0.014	0.947
$\underline{\sigma^2_Y}$	20%	0.996	0.002	—	1.000	0.002	—
	50%	0.995	0.002	—	1.001	0.004	—
	70%	0.993	0.002	—	1.000	0.007	—
$\underline{\sigma^2_M}$	20%	0.996	0.002	—	0.999	0.002	—
	50%	0.995	0.004	—	1.001	0.004	—
	70%	0.990	0.007	—	1.000	0.007	—

Table 4.2: Mean estimate, empirical variance, and 95% coverage when MCAR for 10^4 simulations examining causal effects.

		All Available Data			Just Complete Data		
Parameter	% Missing	Mean	Var	95 CI	Mean	Var	95 CI
NIE Continuous	20%	0.0195	5.11e-05	0.931	0.0195	5.12e-05	0.931
	50%	0.0198	8.35e-05	0.927	0.0198	8.41e-05	0.929
	70%	0.0197	1.42e-04	0.916	0.0195	1.43e-04	0.915
NIE Categorical	20%	0.0195	1.34e-04	0.937	0.0195	1.35e-04	0.938
	50%	0.0197	2.25e-04	0.935	0.0197	2.27e-04	0.936
	70%	0.0198	4.23e-04	0.921	0.0197	4.22e-04	0.922
NTE Continuous	20%	0.1595	1.01e-03	0.953	0.1595	1.29e-03	0.949
	50%	0.1597	1.00e-03	0.952	0.1597	2.04e-03	0.950
	70%	0.1598	1.05e-03	0.947	0.1594	3.49e-03	0.950
NTE Categorical	20%	0.1595	4.28e-03	0.950	0.1592	5.46e-03	0.950
	50%	0.1588	4.13e-03	0.952	0.1584	9.01e-03	0.949
	70%	0.1590	4.34e-03	0.948	0.1595	1.51e-02	0.955

Type I error results when MCAR.

Table 4.3: Type I error evaluated at $\alpha = 0.05$ when MCAR over 5x10^4 simulations with n=1000. Examined under different TIE scenarios. Case 1: $\gamma_1 = \gamma_2 = \beta_1 = \beta_2 = 0$, $\beta_3 = 0.39$, Case 2: $\gamma_1 = \gamma_2 = 0.39$, Case 3: $\gamma_1 = \gamma_2 = \beta_1 = \beta_2 = \beta_3 = 0$.

Exposure	TIE Case	% Missing	All Available Data			Just Complete data		
			NDE	NIE	NTE	NDE	NIE	NTE
Continuous	Case 1	20%	0.050	0.052	0.050	0.050	0.052	0.050
		50%	0.050	0.051	0.051	0.051	0.052	0.050
		70%	0.049	0.052	0.050	0.051	0.053	0.050
	Case 2	20%	0.051	0.050	0.049	0.050	0.051	0.025
		50%	0.050	0.050	0.051	0.051	0.052	0.025
		70%	0.052	0.052	0.052	0.050	0.056	0.025
	Case 3	20%	0.051	0.002	0.050	0.051	0.002	0.050
		50%	0.050	0.003	0.050	0.050	0.003	0.049
		70%	0.050	0.003	0.052	0.050	0.003	0.051
Categorical	Case 1	20%	0.050	0.051	0.050	0.051	0.052	0.050
		50%	0.051	0.051	0.051	0.051	0.052	0.051
		70%	0.050	0.051	0.050	0.050	0.051	0.048
	Case 2	20%	0.051	0.050	0.051	0.052	0.051	0.043
		50%	0.050	0.049	0.051	0.049	0.051	0.044
		70%	0.050	0.046	0.051	0.051	0.050	0.044
	Case 3	20%	0.051	0.003	0.050	0.051	0.003	0.050
		50%	0.052	0.003	0.051	0.051	0.003	0.050
		70%	0.050	0.003	0.051	0.050	0.003	0.050

Table 4.4: Type I error evaluated at $\alpha = 0.01$ when MCAR over 5x10^4 simulations with n=1000. Examined under different TIE scenarios. Case 1: $\gamma_1 = \gamma_2 = \beta_1 = \beta_2 = 0$, $\beta_3 = 0.39$, Case 2: $\gamma_1 = \gamma_2 = 0.39$, Case 3: $\gamma_1 = \gamma_2 = \beta_1 = \beta_2 = \beta_3 = 0$.

Exposure	TIE Case	% Missing	EM			LWD		
			NDE	NIE	NTE	NDE	NIE	NTE
Continuous	Case 1	20%	0.010	0.011	0.010	0.010	0.011	0.010
		50%	0.010	0.010	0.011	0.011	0.010	0.010
		70%	0.010	0.012	0.011	0.010	0.011	0.011
	Case 2	20%	0.010	0.010	0.010	0.010	0.010	0.003
		50%	0.010	0.011	0.010	0.010	0.012	0.003
		70%	0.010	0.011	0.011	0.010	0.013	0.004
	Case 3	20%	0.010	0.000	0.010	0.010	0.000	0.010
		50%	0.010	0.000	0.010	0.010	0.000	0.010
		70%	0.009	0.000	0.010	0.010	0.000	0.010
Categorical	Case 1	20%	0.011	0.011	0.010	0.011	0.011	0.011
		50%	0.010	0.011	0.010	0.011	0.011	0.010
		70%	0.010	0.010	0.011	0.010	0.011	0.010
	Case 2	20%	0.011	0.010	0.011	0.011	0.010	0.008
		50%	0.010	0.010	0.010	0.011	0.011	0.008
		70%	0.010	0.008	0.011	0.011	0.009	0.009
	Case 3	20%	0.011	0.000	0.011	0.011	0.000	0.011
		50%	0.010	0.000	0.011	0.010	0.000	0.011
		70%	0.010	0.000	0.010	0.010	0.000	0.010

Association with having observed DNAm data in Project Viva

Table 4.5: Testing for covariates, outcome, and exposure being associated with observed DNAm data in whites

	6 months	3 years	7 years
N	746	798	655
Have DNAm	247	289	274
Pre Pregnancy	0.803	0.456	0.349
BMI Z-score	0.821	0.010	0.007
Dad BMI	0.912	0.631	0.283
Maternal Age	0.282	0.256	0.643
College Grad	0.144	0.050	0.115
New Parent	0.126	0.407	0.819
Married or Cohabited	0.305	0.035	0.156
Gender of Baby	0.620	0.764	0.336
Smoking Status	0.700	0.656	0.381

Table 4.6: Association with having observed DNAm in analysis including all individuals

	6 months	3 years	7 years
N	1090	1174	1027
Have DNAm	356	430	402
Pre Pregnancy	0.810	0.918	0.202
BMI Z-score	0.662	0.142	0.072
Dad BMI	0.762	0.617	0.866
Maternal Age	0.271	0.079	0.659
College Grad	0.496	0.009	0.603
New Parent	0.408	0.446	0.800
Married or Cohabited	0.288	0.046	0.701
Gender of Baby	0.259	0.490	0.295
Smoking Status	0.597	0.954	0.422
Race	0.492	0.506	0.042

4.1.3 More detailed proof for independence of $\hat{\gamma}_A$ and $\hat{\beta}_m$

Individuals with Complete Data

$$\boldsymbol{\Theta} = \begin{bmatrix} \beta_0 & \beta_a & \boldsymbol{\beta}_x^T & \beta_m & \gamma_0 & \gamma_a & \boldsymbol{\gamma}_x^T & \sigma_Y^2 & \sigma_M^2 \end{bmatrix}^T$$

$$
\begin{aligned}
\ell_i &= \log(\mathrm{L}(Y_i|M_i, \boldsymbol{\Theta}, \boldsymbol{X}_i, A_i) * \mathrm{L}(M_i|\boldsymbol{\Theta}, \boldsymbol{X}_i, A_i)) \\
&\propto -\frac{1}{2}\log(\sigma_Y^2) - \frac{(Y_i - \beta_0 - A_i\beta_a - \boldsymbol{X}_i\boldsymbol{\beta}_C - M_i\beta_m)^2}{2\sigma_Y^2} \\
&- \frac{1}{2}\log(\sigma_M^2) - \frac{(M_i - \gamma_0 - A_i\gamma_a - \boldsymbol{X}_i\boldsymbol{\gamma}_C)^2}{2\sigma_M^2}
\end{aligned}
$$

The expected information is:

$$
F_C = \begin{array}{c} \\ \beta_0 \\ \beta_a \\ \beta_m \\ \boldsymbol{\beta}_C \\ \gamma_0 \\ \gamma_a \\ \boldsymbol{\gamma}_C \\ \sigma_Y^2 \\ \sigma_M^2 \end{array}
\begin{array}{c}
\begin{array}{ccccccccc} \beta_0 & \beta_a & \beta_m & \boldsymbol{\beta}_C & \gamma_0 & \gamma_a & \boldsymbol{\gamma}_C & \sigma_Y^2 & \sigma_M^2 \end{array} \\
\begin{pmatrix}
C_{11} & C_{12} & C_{13} & C_{14} & 0 & 0 & 0 & 0 & 0 \\
C_{12} & C_{22} & C_{23} & C_{24} & 0 & 0 & 0 & 0 & 0 \\
C_{13} & C_{23} & C_{33} & C_{34} & 0 & 0 & 0 & 0 & 0 \\
C_{14} & C_{24} & C_{34} & C_{44} & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & C_{55} & C_{56} & C_{57} & 0 & 0 \\
0 & 0 & 0 & 0 & C_{56} & C_{66} & C_{67} & 0 & 0 \\
0 & 0 & 0 & 0 & C_{57} & C_{67} & C_{77} & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & C_{88} & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & C_{99}
\end{pmatrix}
\end{array}
$$

Individuals with Missing Data

$$
\begin{aligned}
\mathrm{L}(Y_i|\boldsymbol{\Theta}, \boldsymbol{X}_i, A_i) &\propto \frac{1}{\sqrt{(\sigma_Y^2 + \beta_m^2\sigma_M^2)}} \\
&\exp\left(\frac{-(Y_i - \beta_0 - A_i\beta_a - \boldsymbol{X}_i\boldsymbol{\beta}_C - (\gamma_0 + A_i\gamma_a + \boldsymbol{X}_i\boldsymbol{\gamma}_C)\beta_m)^2}{2\left(\sigma_Y^2 + \beta_m^2\sigma_M^2\right)}\right) \\
\ell(Y_i|\boldsymbol{\Theta}, \boldsymbol{X}_i, A_i) &\propto -\frac{1}{2}\log(\sigma_Y^2 + \beta_m^2\sigma_M^2) \\
&- \frac{(Y_i - \beta_0 - A_i\beta_a - \boldsymbol{X}_i\boldsymbol{\beta}_C - (\gamma_0 + A_i\gamma_a + \boldsymbol{X}_i\boldsymbol{\gamma}_C)\beta_m)^2}{2\left(\sigma_Y^2 + \beta_m^2\sigma_M^2\right)}
\end{aligned}
$$

The expected fisher information is

$$
F_M = \begin{array}{c} \\ \beta_0 \\ \beta_a \\ \beta_m \\ \boldsymbol{\beta}_C \\ \gamma_0 \\ \gamma_a \\ \boldsymbol{\gamma}_C \\ \sigma_Y^2 \\ \sigma_M^2 \end{array}
\begin{array}{c}
\begin{array}{ccccccccc} \beta_0 & \beta_a & \beta_m & \boldsymbol{\beta}_C & \gamma_0 & \gamma_a & \boldsymbol{\gamma}_C & \sigma_Y^2 & \sigma_M^2 \end{array} \\
\begin{pmatrix}
B_{11} & B_{12} & B_{13} & B_{14} & B_{15} & B_{16} & B_{17} & 0 & 0 \\
B_{12} & B_{22} & B_{23} & B_{24} & B_{25} & B_{26} & B_{27} & 0 & 0 \\
B_{13} & B_{23} & B_{33} & B_{34} & B_{35} & B_{36} & B_{37} & B_{38} & B_{39} \\
B_{14} & B_{24} & B_{34} & B_{44} & B_{45} & B_{46} & B_{47} & 0 & 0 \\
B_{15} & B_{25} & B_{35} & B_{45} & B_{55} & B_{56} & B_{57} & 0 & 0 \\
B_{16} & B_{26} & B_{36} & B_{46} & B_{56} & B_{66} & B_{67} & 0 & 0 \\
B_{17} & B_{27} & B_{37} & B_{47} & B_{57} & B_{67} & B_{77} & 0 & 0 \\
0 & 0 & B_{38} & 0 & 0 & 0 & 0 & B_{88} & B_{89} \\
0 & 0 & B_{39} & 0 & 0 & 0 & 0 & B_{89} & B_{99}
\end{pmatrix}
\end{array}
$$

The overall Fisher information is then $F_M + F_C$. We want to the third row, sixth column entry of the inverse of $F_M + F_C$. For individuals with missing data:

$$B_{11} = \sum_{i=1}^{n_m} \frac{-1}{\beta_m^2 \sigma_M^2 + \sigma_Y^2}$$

$$B_{12} = \sum_{i=1}^{n_m} \frac{-A_i}{\beta_m^2 \sigma_M^2 + \sigma_Y^2}$$

$$B_{14} = \sum_{i=1}^{n_m} \frac{-\boldsymbol{X}_i}{\beta_m^2 \sigma_M^2 + \sigma_Y^2}$$

$$B_{13} = \sum_{i=1}^{n_m} -\frac{\gamma_0 + A_i\gamma_a + \boldsymbol{X}_i\boldsymbol{\gamma}_C}{\beta_m^2 \sigma_M^2 + \sigma_Y^2} = B_{11}\gamma_0 + B_{12}\gamma_a + B_{14}\boldsymbol{\gamma}_C$$

$$B_{15} = \sum_{i=1}^{n_m} \frac{-\beta_m}{\beta_m^2 \sigma_M^2 + \sigma_Y^2} = B_{11}\beta_m$$

$$B_{16} = \sum_{i=1}^{n_m} \frac{-A_i\beta_m}{\beta_m^2 \sigma_M^2 + \sigma_Y^2} = B_{12}\beta_m$$

$$B_{17} = \sum_{i=1}^{n_m} \frac{-\boldsymbol{X}_i\beta_m}{\beta_m^2 \sigma_M^2 + \sigma_Y^2} = B_{14}\beta_m$$

$$B_{22} = \sum_{i=1}^{n_m} \frac{-A_i^2}{\beta_m^2 \sigma_M^2 + \sigma_Y^2}$$

$$B_{24} = \sum_{i=1}^{n_m} \frac{-A_i\boldsymbol{X}_i}{\beta_m^2 \sigma_M^2 + \sigma_Y^2}$$

$$B_{23} = \sum_{i=1}^{n_m} -\frac{A_i(\gamma_0 + A_i\gamma_a + \boldsymbol{X}_i\boldsymbol{\gamma}_C)}{\beta_m^2 \sigma_M^2 + \sigma_Y^2} = B_{12}\gamma_0 + B_{22}\gamma_a + B_{24}\boldsymbol{\gamma}_C$$

$$B_{25} = \sum_{i=1}^{n_m} \frac{-A_i\beta_m}{\beta_m^2 \sigma_M^2 + \sigma_Y^2} = B_{12}\beta_m$$

$$B_{26} = \sum_{i=1}^{n_m} \frac{-A_i^2\beta_m}{\beta_m^2 \sigma_M^2 + \sigma_Y^2} = B_{22}\beta_m$$

$$B_{27} = \sum_{i=1}^{n_m} \frac{-A_i\boldsymbol{X}_i\beta_m}{\beta_m^2 \sigma_M^2 + \sigma_Y^2} = B_{24}\beta_m$$

$$B_{44} = \sum_{i=1}^{n_m} \frac{-\boldsymbol{X}_i^T\boldsymbol{X}_i}{\beta_m^2 \sigma_M^2 + \sigma_Y^2}$$

$$B_{45} = \sum_{i=1}^{n_m} \frac{\boldsymbol{X}_i^T\beta_m}{\beta_m^2 \sigma_M^2 + \sigma_Y^2} = B_{14}^T\beta_m$$

$$B_{34} = \sum_{i=1}^{n_m} -\frac{(\gamma_0 + A_i\gamma_a + \boldsymbol{X}_i\boldsymbol{\gamma}_C)\boldsymbol{X}_i}{\beta_m^2\sigma_M^2 + \sigma_Y^2} = B_{14}\gamma_0 + B_{24}\gamma_a + \boldsymbol{\gamma}_C^T B_{44}$$

$$B_{35} = \sum_{i=1}^{n_m} -\frac{\beta_m(\gamma_0 + A_i\gamma_a + \boldsymbol{X}_i\boldsymbol{\gamma}_C)}{\beta_m^2\sigma_M^2 + \sigma_Y^2} = B_{15}\gamma_0 + B_{25}\gamma_a + B_{45}^T\boldsymbol{\gamma}_C$$

$$B_{46} = \sum_{i=1}^{n_m} \frac{\boldsymbol{X}_i^T A_i\beta_m}{\beta_m^2\sigma_M^2 + \sigma_Y^2} = B_{24}^T\beta_m = B_{27}^T$$

$$B_{36} = \sum_{i=1}^{n_m} -\frac{\beta_m A_i(\gamma_0 + A_i\gamma_a + \boldsymbol{X}_i\boldsymbol{\gamma}_C)}{\beta_m^2\sigma_M^2 + \sigma_Y^2} = B_{16}\gamma_0 + B_{26}\gamma_a + B_{46}^T\boldsymbol{\gamma}_C = B_{16}\gamma_0 + B_{26}\gamma_a + \boldsymbol{\gamma}_C^T B_{46}$$

$$B_{47} = \sum_{i=1}^{n_m} \frac{\boldsymbol{X}_i^T\boldsymbol{X}_i\beta_m}{\beta_m^2\sigma_M^2 + \sigma_Y^2} = B_{44}\beta_m$$

$$B_{37} = \sum_{i=1}^{n_m} -\frac{\beta_m\boldsymbol{X}_i(\gamma_0 + A_i\gamma_a + \boldsymbol{X}_i\boldsymbol{\gamma}_C)}{\beta_m^2\sigma_M^2 + \sigma_Y^2} = B_{34}\beta_m = B_{45}^T\gamma_0 + B_{27}^T\gamma_a + \boldsymbol{\gamma}_C^T B_{47}^T$$

For the observations with complete data:

$$C_{11} = \sum_{i=1}^{n_c} \frac{-1}{\sigma_Y^2}$$

$$C_{12} = \sum_{i=1}^{n_c} \frac{-A_i}{\sigma_Y^2}$$

$$C_{14} = \sum_{i=1}^{n_c} \frac{-\boldsymbol{X}_i}{\sigma_Y^2}$$

$$C_{13} = \sum_{i=1}^{n_c} -\frac{\gamma_0 + A_i\gamma_a + \boldsymbol{X}_i\boldsymbol{\gamma}_C}{\sigma_Y^2} = C_{11}\gamma_0 + C_{12}\gamma_a + C_{14}\boldsymbol{\gamma}_C$$

$$C_{22} = \sum_{i=1}^{n_c} \frac{-A_i^2}{\sigma_Y^2}$$

$$C_{24} = \sum_{i=1}^{n_c} \frac{-A_i\boldsymbol{X}_i}{\sigma_Y^2}$$

$$C_{23} = \sum_{i=1}^{n_c} -\frac{A_i(\gamma_0 + A_i\gamma_a + \boldsymbol{X}_i\boldsymbol{\gamma}_C)}{\sigma_Y^2} = C_{12}\gamma_0 + C_{22}\gamma_a + C_{24}\boldsymbol{\gamma}_C$$

$$C_{44} = \sum_{i=1}^{n_c} \frac{-\boldsymbol{X}_i^T\boldsymbol{X}_i}{\sigma_Y^2}$$

$$C_{34} = \sum_{i=1}^{n_c} -\frac{\boldsymbol{X}_i(\gamma_0 + A_i\gamma_a + \boldsymbol{X}_i\boldsymbol{\gamma}_C)}{\sigma_Y^2} = C_{14}\gamma_0 + C_{24}\gamma_a + \boldsymbol{\gamma}_C^T C_{44}$$

These specifications will help in simplifying the inverse. We do not want the inverse of

the whole matrix. We will rewrite $F_M + F_C$ as:

$$F_M + F_C = \begin{pmatrix} A & G \\ G^T & D \end{pmatrix}$$

Where we have reorganized the matrix so that F_M corresponds to β_M and γ_A. G corresponds to β_M and γ_A vs β_0, β_A, $\boldsymbol{\beta}_C$, γ_0, $\boldsymbol{\gamma}_C$, σ_Y^2, and σ_M^2 respectively. D is the Fisher information for those variables β_0, β_A, $\boldsymbol{\beta}_C$, γ_0, $\boldsymbol{\gamma}_C$, σ_Y^2, and σ_M^2. The sub matrices are

$$A = \begin{array}{c} \\ \beta_m \\ \gamma_a \end{array} \begin{array}{c} \begin{array}{cc} \beta_m & \gamma_a \end{array} \\ \begin{pmatrix} B_{33} + C_{33} & B_{36} \\ B_{36} & B_{66} + C_{66} \end{pmatrix} \end{array}$$

$$G = \begin{pmatrix} G_1 & G_2 \\ G_3 & 0 \end{pmatrix}$$

$$G_1 = \begin{array}{c} \\ \beta_m \end{array} \begin{array}{c} \begin{array}{ccccc} \beta_0 & \beta_a & \boldsymbol{\beta}_C & \gamma_0 & \boldsymbol{\gamma}_C \end{array} \\ \begin{pmatrix} B_{13} + C_{13} & B_{23} + C_{23} & B_{34} + C_{34} & B_{35} & B_{37} \end{pmatrix} \end{array}$$

$$G_2 = \begin{array}{c} \\ \beta_m \end{array} \begin{array}{c} \begin{array}{cc} \sigma_Y^2 & \sigma_M^2 \end{array} \\ \begin{pmatrix} B_{38} & B_{39} \end{pmatrix} \end{array}$$

$$G_3 = \begin{array}{c} \\ \gamma_a \end{array} \begin{array}{c} \begin{array}{ccccc} \beta_0 & \beta_a & \boldsymbol{\beta}_C & \gamma_0 & \boldsymbol{\gamma}_C \end{array} \\ \begin{pmatrix} B_{16} & B_{26} & B_{46}^T & B_{56} + C_{56} & B_{67} + C_{67} \end{pmatrix} \end{array}$$

$$D = \begin{pmatrix} D_1 & 0 \\ 0 & D_2 \end{pmatrix}$$

$$D_1 = \begin{array}{c} \\ \beta_0 \\ \beta_a \\ \boldsymbol{\beta}_C \\ \gamma_0 \\ \boldsymbol{\gamma}_C \end{array} \begin{array}{c} \begin{array}{ccccc} \beta_0 & \beta_a & \boldsymbol{\beta}_C & \gamma_0 & \boldsymbol{\gamma}_C \end{array} \\ \begin{pmatrix} B_{11} + C_{11} & B_{12} + C_{12} & B_{14} + C_{14} & B_{15} & B_{17} \\ B_{12} + C_{12} & B_{22} + C_{22} & B_{24} + C_{24} & B_{25} & B_{27} \\ B_{14}^T + C_{14}^T & B_{24}^T + C_{24}^T & B_{44} + C_{44} & B_{45} & B_{47} \\ B_{15} & B_{25} & B_{45}^T & B_{55} + C_{55} & B_{57} + C_{57} \\ B_{17}^T & B_{27}^T & B_{47}^T & B_{57}^T + C_{57}^T & B_{77} + C_{77} \end{pmatrix} \end{array}$$

$$D_2 = \begin{array}{c} \\ \sigma_Y^2 \\ \sigma_M^2 \end{array} \begin{array}{c} \begin{array}{cc} \sigma_Y^2 & \sigma_M^2 \end{array} \\ \begin{pmatrix} B_{88} + C_{88} & B_{89} \\ B_{89} & B_{99} + C_{99} \end{pmatrix} \end{array}$$

The covariance of $\hat{\beta}_m$ and $\hat{\gamma}_a$ will then be (using block matrix inversion)

$$-\left(A - GD^{-1}G^T\right)^{-1}$$

Focus on $GD^{-1}G^T$:

$$\begin{aligned} GD^{-1}G^T &= \begin{bmatrix} G_1 & G_2 \\ G_3 & 0 \end{bmatrix} \begin{bmatrix} D_1^{-1} & 0 \\ 0 & D_2^{-1} \end{bmatrix} \begin{bmatrix} G_1^T & G_3^T \\ G_2^T & 0^T \end{bmatrix} \\ &= \begin{bmatrix} G_1 D_1^{-1} G_1^T + G_2 D_2^{-1} G_2^T & G_1 D_1^{-1} G_3^T \\ G_3 D_1^{-1} G_1^T & G_3 D_1^{-} 1 G_3^T \end{bmatrix} \end{aligned}$$

The covariance of $\hat{\beta}_m$ and $\hat{\gamma}_a$ will be

$$-\left(\begin{bmatrix} B_{33} + C_{33} & B_{36} \\ B_{36} & B_{66} + C_{66} \end{bmatrix} - \begin{bmatrix} G_1 D_1^{-1} G_1^T + G_2 D_2^{-1} G_2^T & G_1 D_1^{-1} G_3^T \\ G_3 D_1^{-1} G_1^T & G_3 D_1^{-1} G_3^T \end{bmatrix} \right)^{-1}$$

Therefore for $\hat{\beta}_m$ and γ_a to have 0 covariance we need:

$$B_{36} - G_1 D_1^{-1} G_3^T = 0$$

We next simplify G_1:

$$\begin{aligned} G_1^T = \begin{bmatrix} B_{13} + C_{13} \\ B_{23} + C_{23} \\ B_{34}^T + C_{34}^T \\ B_{35} \\ B_{37}^T \end{bmatrix} &= \begin{bmatrix} (B_{11} + C_{11})\gamma_0 + (B_{12} + C_{12})\gamma_a + (B_{14} + C_{14})\boldsymbol{\gamma}_C \\ (B_{12} + C_{12})\gamma_0 + (B_{22} + C_{22})\gamma_a + (B_{24} + C_{24})\boldsymbol{\gamma}_C \\ (B_{14}^T + C_{14}^T)\gamma_0 + (B_{24}^T + C_{24}^T)\gamma_a + (B_{44} + C_{44})\boldsymbol{\gamma}_C \\ B_{15}\gamma_0 + B_{25}\gamma_a + B_{45}^T \boldsymbol{\gamma}_C \\ B_{17}^T \gamma_0 + B_{27}^T \gamma_a + B_{47}^T \boldsymbol{\gamma}_C \end{bmatrix} \\ &= D_1 \begin{bmatrix} \gamma_0 \\ \gamma_a \\ \boldsymbol{\gamma}_C \\ 0_{p+1} \end{bmatrix} \end{aligned}$$

Therefore:

$$\begin{aligned} & B_{36} - G_1 D_1^{-1} G_3^T = 0 \\ & B_{36} - \begin{bmatrix} \gamma_0 & \gamma_a & \boldsymbol{\gamma}_C^T & 0_{p+1}^T \end{bmatrix} D_1 D_1^{-1} G_3^T \\ & B_{36} - \begin{bmatrix} \gamma_0 & \gamma_a & \boldsymbol{\gamma}_C^T & 0_{p+1}^T \end{bmatrix} \begin{bmatrix} B_{16} \\ B_{26} \\ B_{46} \\ B_{56} + C_{56} \\ B_{67}^T + C_{67}^T \end{bmatrix} \\ & B_{36} - B_{16}\gamma_0 - B_{26}\gamma_a - \boldsymbol{\gamma}_C^T B_{46} = 0 \end{aligned}$$

End of proof as $B_{36} = B_{16}\gamma_0 + B_{26}\gamma_a + \boldsymbol{\gamma}_C^T B_{46}$.

4.2 Supplementary Material Paper 2

4.2.1 Supplementary Figures

Bias results for other unobserved covariances

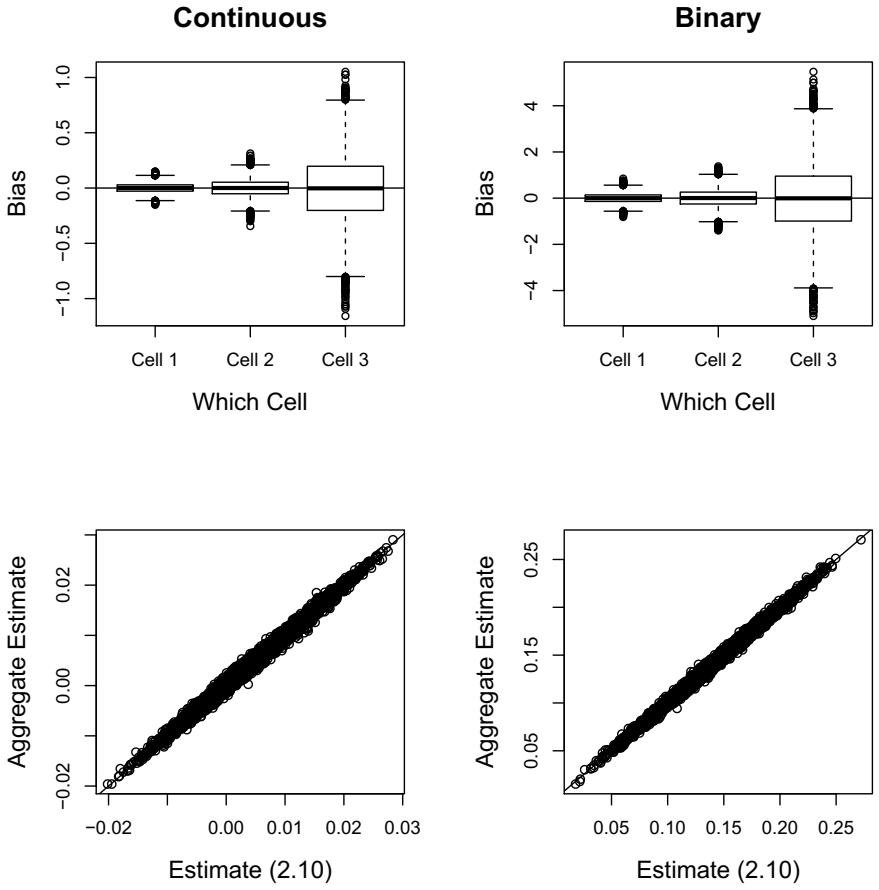

Figure 4.11: Bias and relationship between γ_X and (2.6.1). 10^4 replications, true variance is independent, assuming $\boldsymbol{D} = \sigma^2 \boldsymbol{I}_m$.

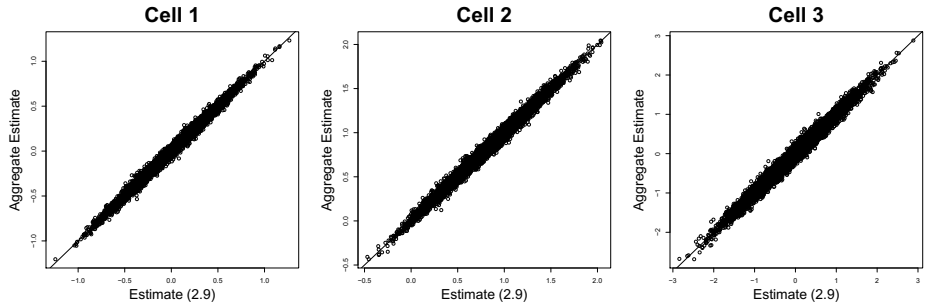

Figure 4.12: Comparing aggregate estimates of $\boldsymbol{\gamma}_\pi$ and (2.6.1) from 10^4. True variance is independent, assuming $\boldsymbol{D} = \sigma^2 \boldsymbol{I}_m$.

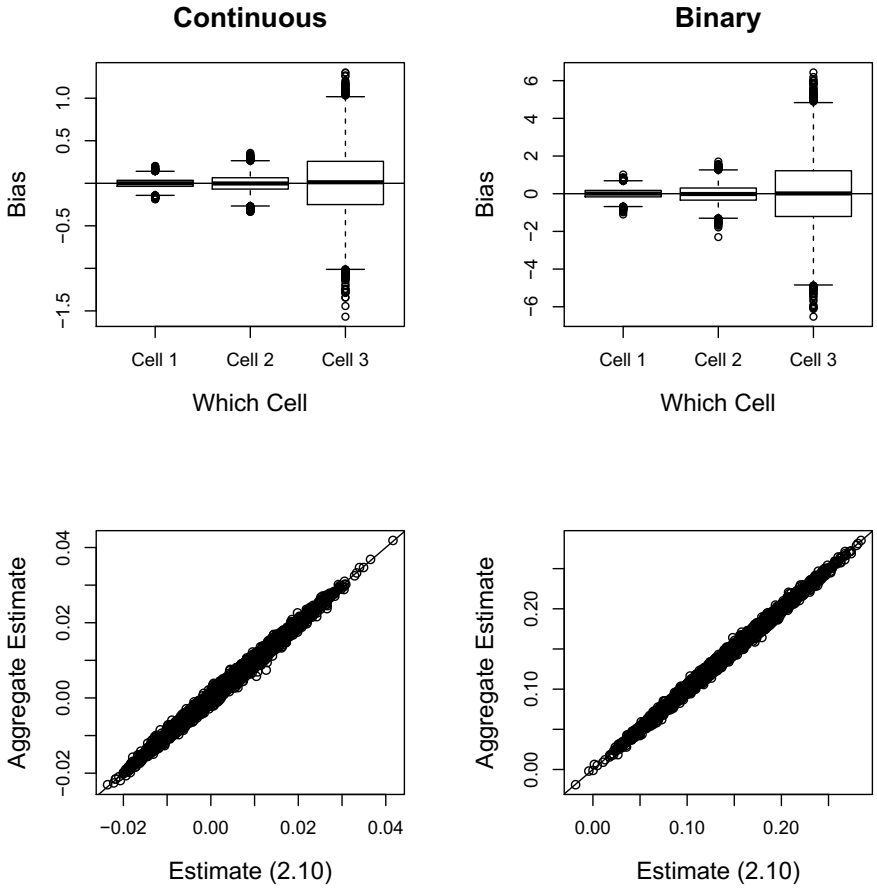

Figure 4.13: Bias and relationship between $\boldsymbol{\gamma}_X$ and (2.6.1). 10^4 replications, true variance is exchangeable, assuming $\boldsymbol{D} = \sigma^2 \boldsymbol{I}_m$.

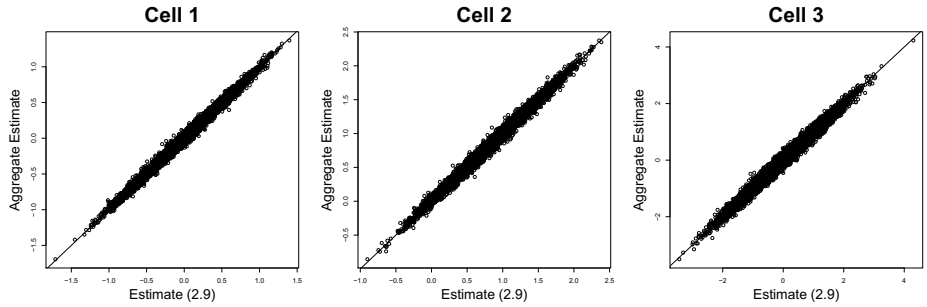

Figure 4.14: Comparing aggregate estimates of γ_π and (2.6.1) from 10^4. True variance is exchangeable, assuming $\boldsymbol{D} = \sigma^2 \boldsymbol{I}_m$.

Type I error results comparing to empirical distribution

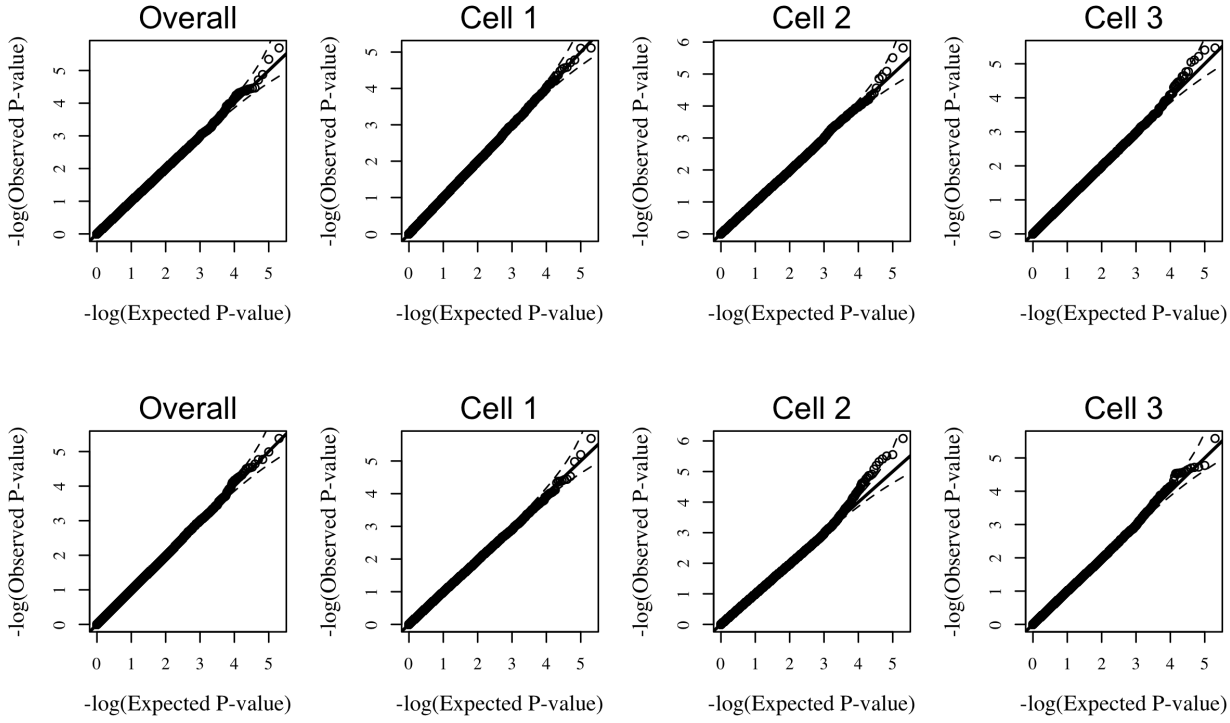

Figure 4.15: Distribution of p-values over 2x10^5 simulations compared to expected p-values. True covariance is independent. Top panel corresponds to continuous while bottom corresponds to binary variable. From left to right, testing for an effect in all cell types, testing cell type 1, cell 2, and cell type 3.

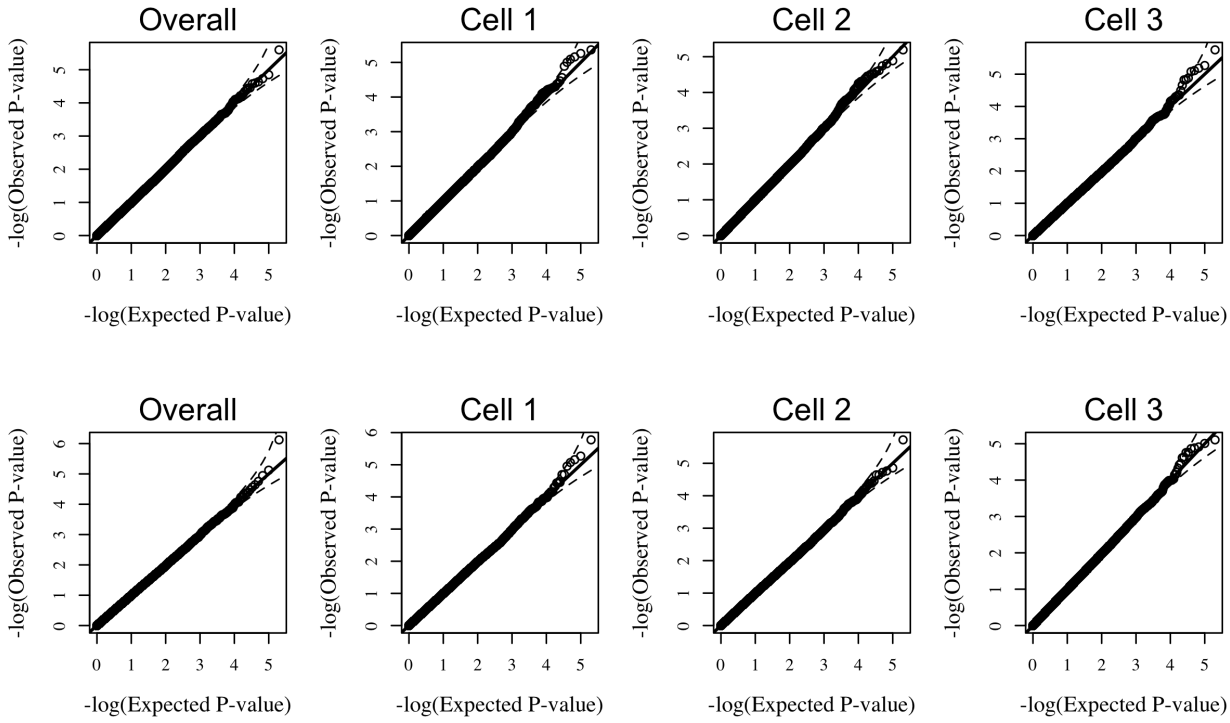

Figure 4.16: Distribution of p-values over $2x10^5$ simulations compared to expected p-values. True covariance is exchangeable. Top panel corresponds to continuous while bottom corresponds to binary variable. From left to right, testing for an effect in all cell types, testing cell type 1, cell 2, and cell type 3.

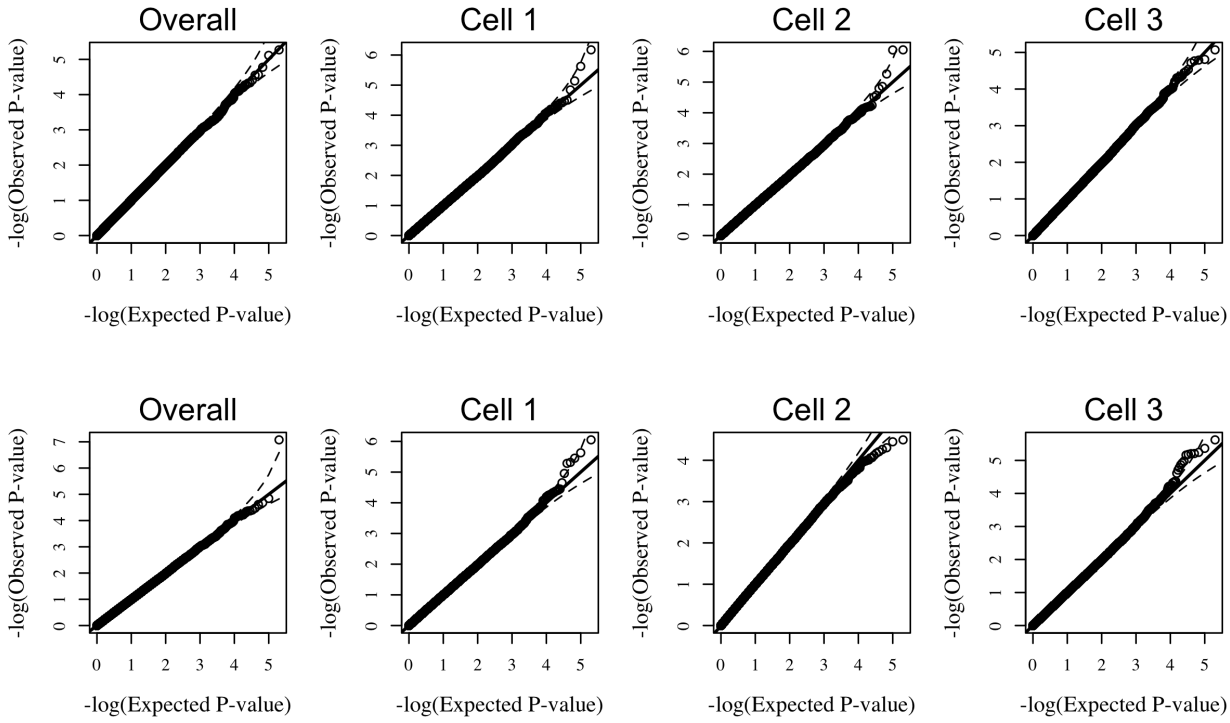

Figure 4.17: Distribution of p-values over $2x10^5$ simulations compared to expected p-values. True covariance is unstructured. Top panel corresponds to continuous while bottom corresponds to binary variable. From left to right, testing for an effect in all cell types, testing cell type 1, cell 2, and cell type 3.

Power results for larger effect sizes

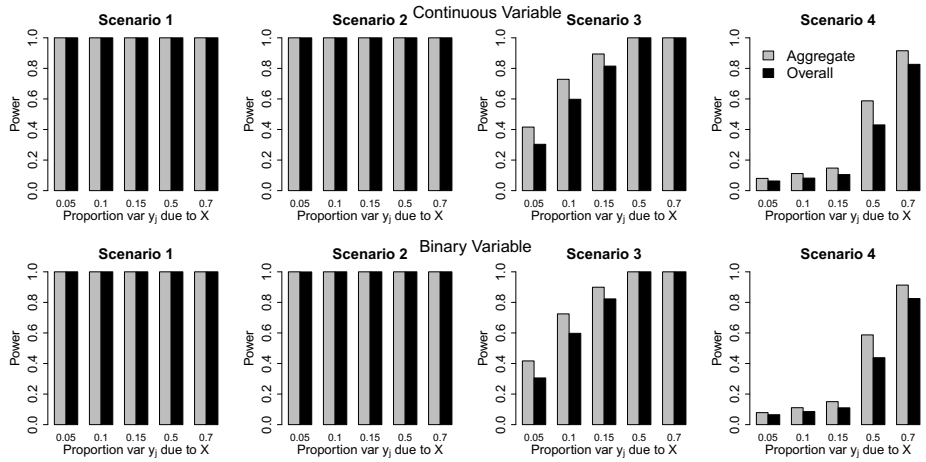

Figure 4.18: Power results of overall and aggregate tests when the true covariance is unstructured at $\alpha = 0.05$. Top panel corresponds to continuous while bottom corresponds to binary variable. From left to right, association in all cell types (Scenario 1), just cell 1(Scenario 2), just cell 2 (Scenario 3), just cell 3 (Scenario 4). Variables were set to explain more power than in main text to see when power to detect association in Scenarios 2 through 4.

Power results for other unobserved covariances

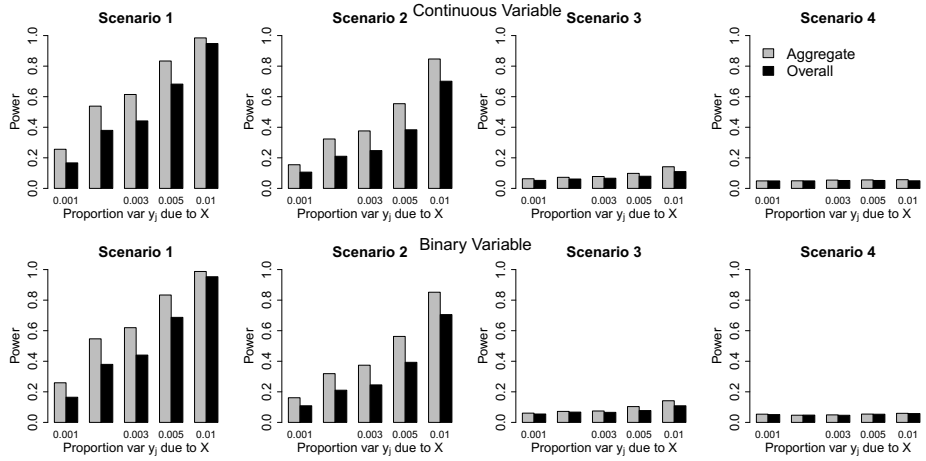

Figure 4.19: Power results of overall and aggregate tests when the true covariance is independent at $\alpha = 0.05$. Top panel corresponds to continuous while bottom corresponds to binary variable. From left to right, association in all cell types (Scenario 1), just cell 1(Scenario 2), just cell 2 (Scenario 3), just cell 3 (Scenario 4). Variables were set to explain .1%, .25%,.3%,.5%,1% of the variation in the $y_{i,j}$'s.

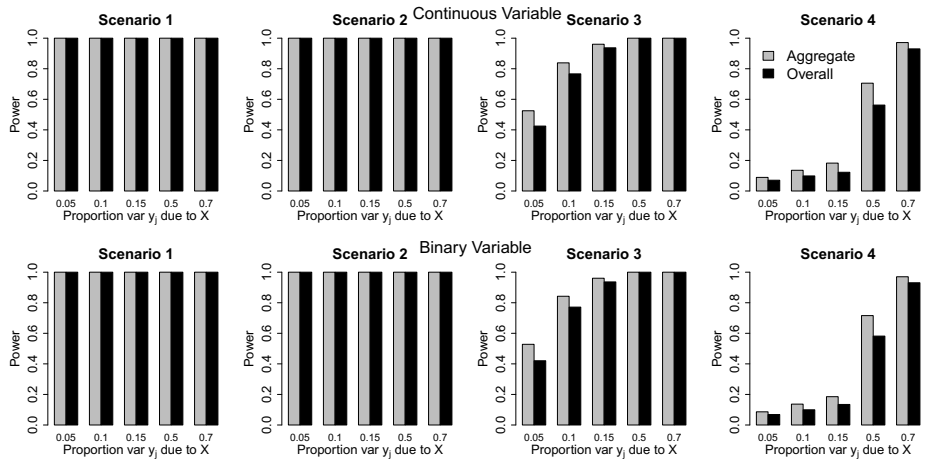

Figure 4.20: Power results of overall and aggregate tests when the true covariance is independent at $\alpha = 0.05$. Top panel corresponds to continuous while bottom corresponds to binary variable. From left to right, association in all cell types (Scenario 1), just cell 1(Scenario 2), just cell 2 (Scenario 3), just cell 3 (Scenario 4). Variables were set to explain more power than in main text to see when power to detect association in Scenarios 2 through 4.

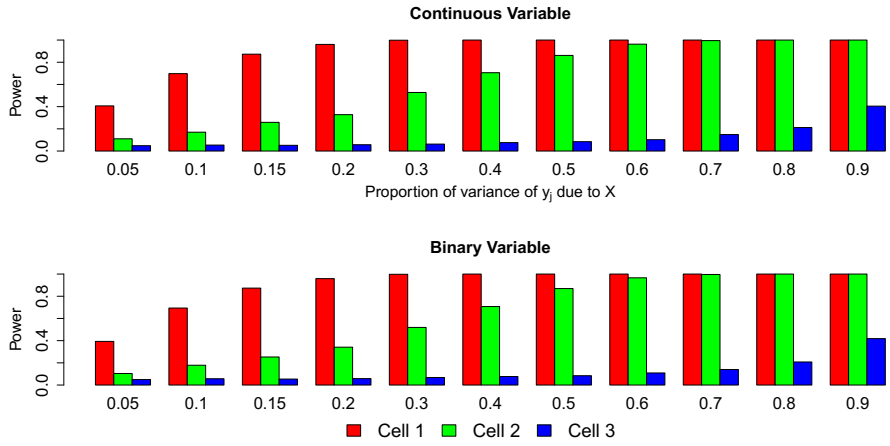

Figure 4.21: Power to detect individual cell specific effects. True covariance is independent, at $\alpha = 0.05$. Top panel is continuous variable, bottom panel is binary.

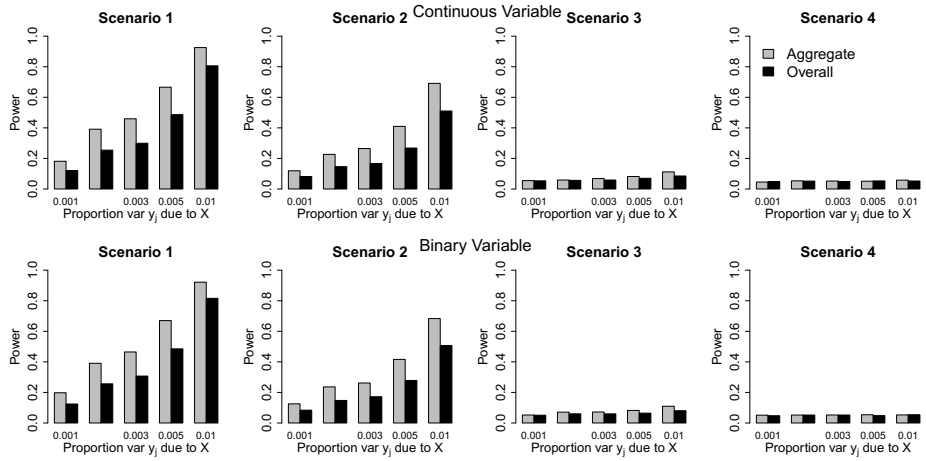

Figure 4.22: Power results of overall and aggregate tests when the true covariance is exchangeable at $\alpha = 0.05$. Top panel corresponds to continuous while bottom corresponds to binary variable. From left to right, association in all cell types (Scenario 1), just cell 1(Scenario 2), just cell 2 (Scenario 3), just cell 3 (Scenario 4). Variables were set to explain .1%, .25%,.3%,.5%,1% of the variation in the $y_{i,j}$'s.

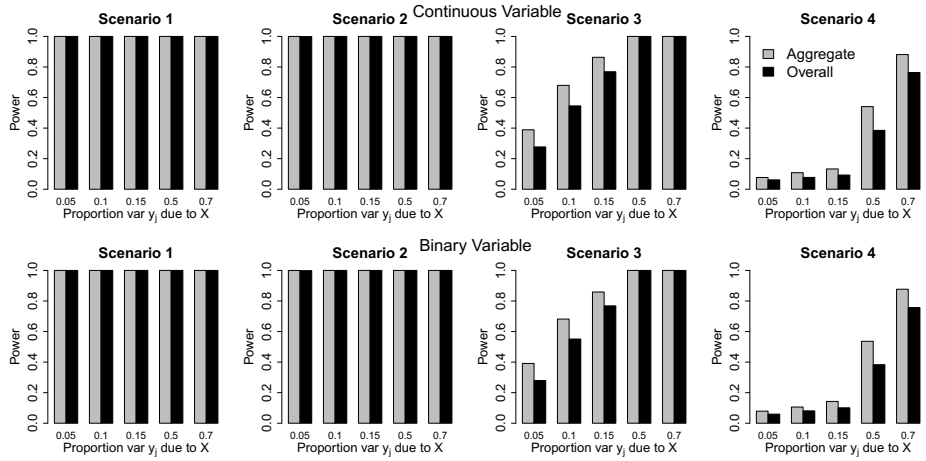

Figure 4.23: Power results of overall and aggregate tests when the true covariance is exchangeable at $\alpha = 0.05$. Top panel corresponds to continuous while bottom corresponds to binary variable. From left to right, association in all cell types (Scenario 1), just cell 1(Scenario 2), just cell 2 (Scenario 3), just cell 3 (Scenario 4). Variables were set to explain more power than in main text to see when power to detect association in Scenarios 2 through 4.

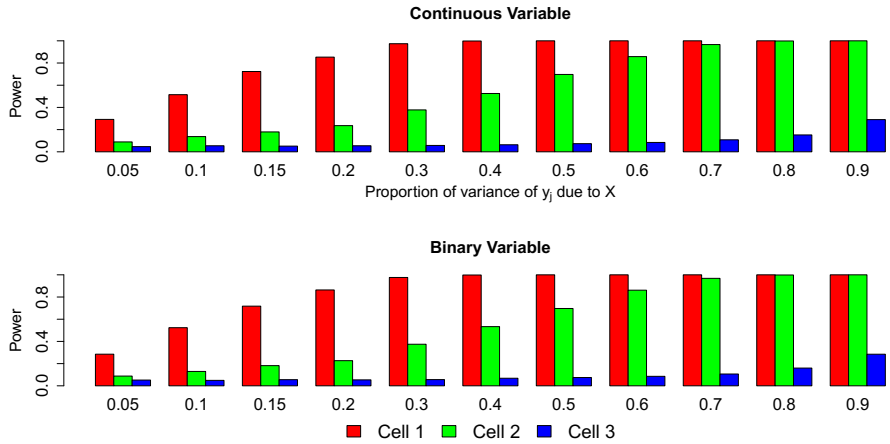

Figure 4.24: Power to detect individual cell specific effects. True covariance is exchangeable, at $\alpha = 0.05$. Top panel is continuous variable, bottom panel is binary.

Bias When X confounds π

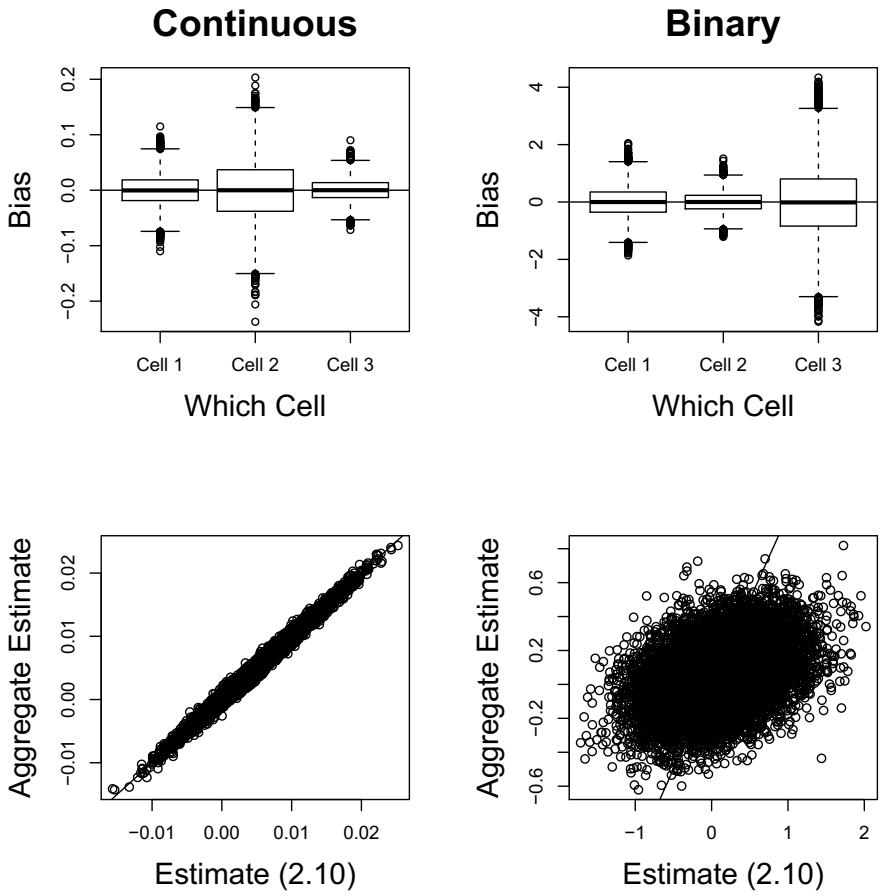

Figure 4.25: Results when X_2 confounds $\boldsymbol{\pi}$ distribution. Bias and relationship between aggregate estimates (2.6.1) 10^4 replications, true variance is independent, assuming $\boldsymbol{D} = \sigma^2 \boldsymbol{I}_m$.

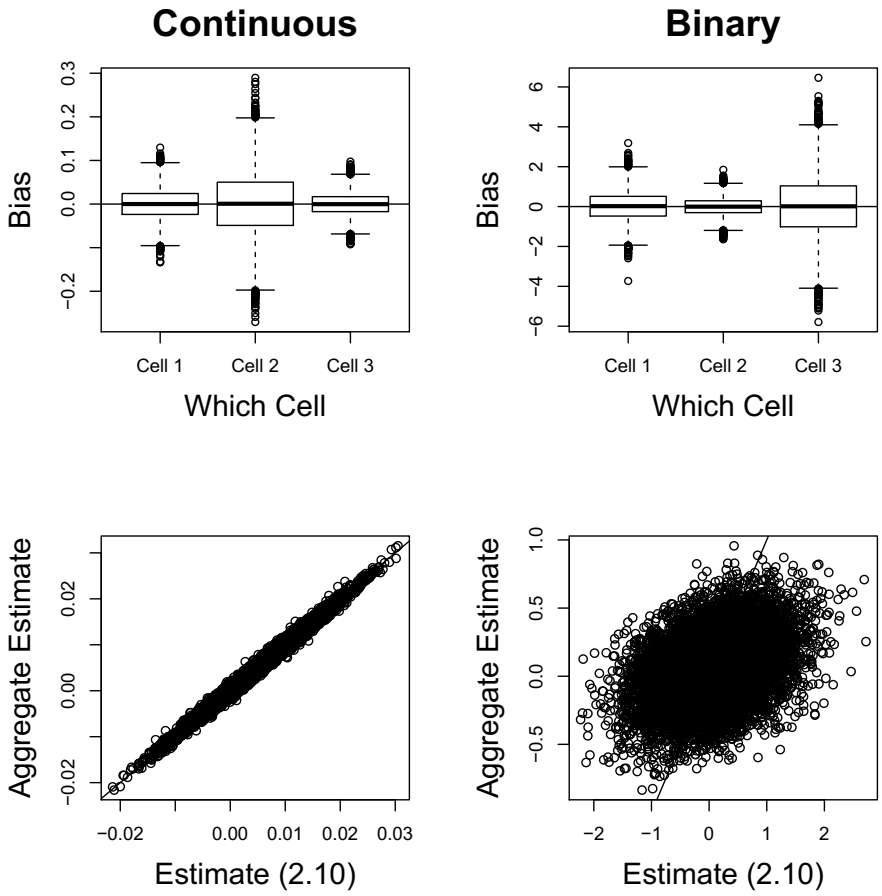

Figure 4.26: Results when X_2 confounds $\boldsymbol{\pi}$ distribution. Bias and relationship between aggregate estimates (2.6.1) 10^4 replications, true variance is exchangeable, assuming $\boldsymbol{D} = \sigma^2 \boldsymbol{I}_m$.

4.2.2 Supplementary Tables

Bias for other forms of the variance

Parameter	True Value	Mean D_1	Mean D_2	Empirical Var D_1	Empirical Var D_2	Mean V_{BC} D_1	Mean V_{BC} D_2
$\beta_{1,1}$	0.00	0.0004	0.0004	0.0018	0.0018	0.0018	0.0018
$\beta_{1,2}$	0.20	0.2003	0.2003	0.0428	0.0428	0.0433	0.0431
$\beta_{2,1}$	0.02	0.0209	0.0211	0.0062	0.0062	0.0064	0.0063
$\beta_{2,2}$	0.00	0.0025	0.0025	0.1460	0.1458	0.1496	0.1487
$\beta_{3,1}$	0.00	-0.0047	-0.0047	0.0893	0.0892	0.0911	0.0907
$\beta_{3,2}$	0.10	0.0925	0.0928	2.0402	2.0457	2.1501	2.1400

Table 4.7: Bias, empirical variance, and mean variance estimate of 10^4 simulations comparing when assume $D(\boldsymbol{\theta}) = \sigma^2 \boldsymbol{I}_m$ (D_1) vs $D(\boldsymbol{\theta}) = diag(\sigma_1^2, \cdots, \sigma_m^2)$ (D_2). For D_2, 955 of the simulations set $\sigma_2^2 = 0$ and 4759 set $\sigma_3^2 = 0$. True variance structure was independent.

Parameter	True Value	Mean D_1	Mean D_2	Empirical Var D_1	Empirical Var D_2	Mean V_{BC} D_1	Mean V_{BC} D_2
$\beta_{1,1}$	0.00	-0.0001	-0.0001	0.0028	0.0028	0.0027	0.0027
$\beta_{1,2}$	0.20	0.2051	0.2053	0.0642	0.0646	0.0647	0.0642
$\beta_{2,1}$	0.02	0.0185	0.0185	0.0098	0.0098	0.0098	0.0097
$\beta_{2,2}$	0.00	-0.0162	-0.0168	0.2267	0.2277	0.2305	0.2287
$\beta_{3,1}$	0.00	0.0036	0.0040	0.1416	0.1421	0.1398	0.1387
$\beta_{3,2}$	0.10	0.1147	0.1149	3.2775	3.2812	3.2937	3.2693

Table 4.8: Bias, empirical variance, and mean variance estimate of 10^4 simulations comparing when assume $D(\boldsymbol{\theta}) = \sigma^2 \boldsymbol{I}_m$ (D_1) vs $D(\boldsymbol{\theta}) = diag(\sigma_1^2, \cdots, \sigma_m^2)$ (D_2). For D_2, 24 of the simulations set $\sigma_2^2 = 0$ and 2936 set $\sigma_3^2 = 0$. True variance structure was exchangeable.

Type I error results for other forms of the variance

	X_1				X_2			
Level α	Overall	Cell 1	Cell 2	Cell 3	Overall	Cell 1	Cell 2	Cell 3
0.05	0.04949	0.05008	0.04926	0.04920	0.04880	0.04790	0.04831	0.04845
0.01	0.00982	0.00971	0.00941	0.00965	0.00942	0.00910	0.00920	0.00931
10^{-3}	0.001	0.00097	0.001	0.00093	0.00098	0.00082	0.00096	0.00094
10^{-5}	0.00001	0.00001	0.00002	0.00003	0.00001	0.00001	0.00003	0.00001

Table 4.9: Type I error results for continuous and binary variable when the true covariance is independent. Each value represents the proportion of 2x10^5 p-values smaller than α.

	X_1				X_2			
Level α	Overall	Cell 1	Cell 2	Cell 3	Overall	Cell 1	Cell 2	Cell 3
0.05	0.04924	0.04870	0.04850	0.04827	0.04924	0.04893	0.04738	0.04804
0.01	0.00979	0.00949	0.00957	0.00924	0.00950	0.00960	0.00887	0.00931
10^{-3}	0.00103	0.00098	0.00102	0.00102	0.00096	0.00097	0.00089	0.00094
10^{-5}	0.00001	0.00003	0.00001	0.00003	0.00001	0.00002	0.00001	0.00001

Table 4.10: Type I error results for continuous and binary variable when the true covariance is exchangeable. Each value represents the proportion of 2x10^5 p-values smaller than α.

Bias for other forms of the variance when X confounds π

Parameter	True Value	Mean D_1	Empirical D_1
$\beta_{1,1}$	0.00	0.0000	0.0008
$\beta_{1,2}$	0.20	0.1994	0.2743
$\beta_{2,1}$	0.02	0.0197	0.0030
$\beta_{2,2}$	0.00	-0.0008	0.1213
$\beta_{3,1}$	0.00	0.0002	0.0004
$\beta_{3,2}$	0.10	0.0888	1.4792

Table 4.11: Parameter estimates and empirical variance when X_2 determines $\boldsymbol{\pi}$. 10^4 simulations when assume $D(\boldsymbol{\theta}) = \sigma^2 \boldsymbol{I}_m$ (D_1).True covariance is independent

Parameter	True Value	Mean D_1	Empirical D_1
$\beta_{1,1}$	0.00	-0.0001	0.0012
$\beta_{1,2}$	0.20	0.2161	0.5297
$\beta_{2,1}$	0.02	0.0203	0.0054
$\beta_{2,2}$	0.00	-0.0051	0.1933
$\beta_{3,1}$	0.00	-0.0001	0.0007
$\beta_{3,2}$	0.10	0.1030	2.2967

Table 4.12: Parameter estimates and empirical variance when X_2 determines $\boldsymbol{\pi}$. 10^4 simulations when assume $D(\boldsymbol{\theta}) = \sigma^2 \boldsymbol{I}_m$ (D_1).True covariance is exchangeable

4.3 Supplement for Paper 3

4.3.1 Supplementary Figures

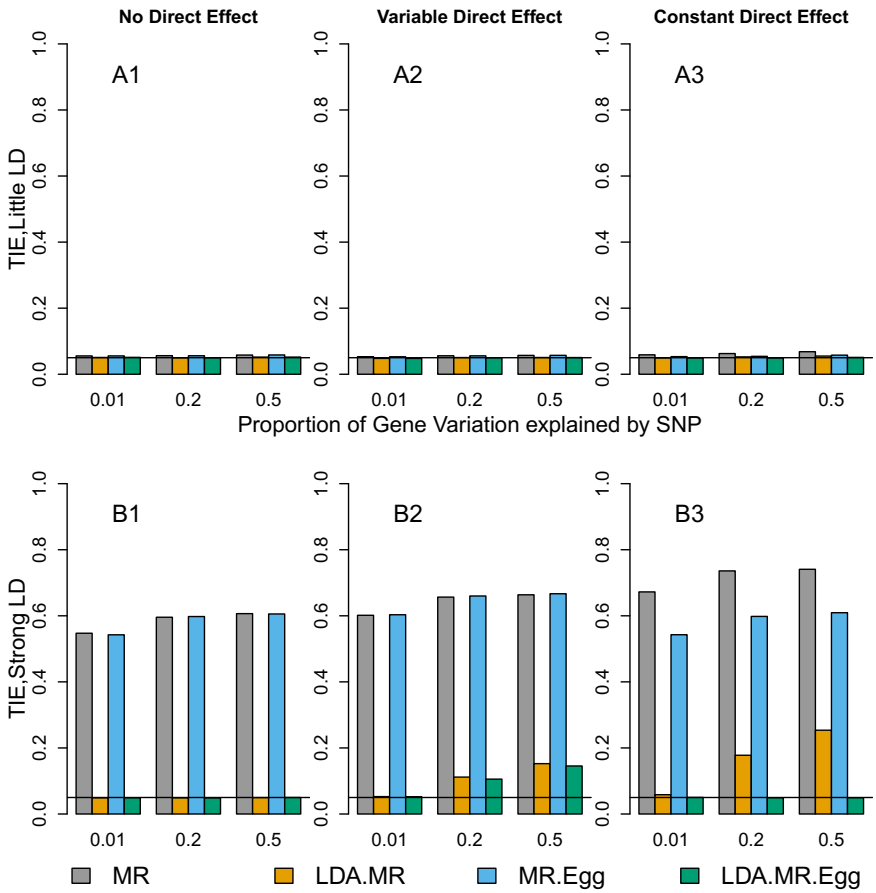

Figure 4.27: Type I error results when J=300. Each bar represents results over 5x10^4 simulations. Evaluated at $\alpha = 0.05$. First panel represent when low LD (plots with A). Second panel represents when strong LD (plots with B). From left to right correspond to: no direct effect, variable direct effect, and a constant direct effect.

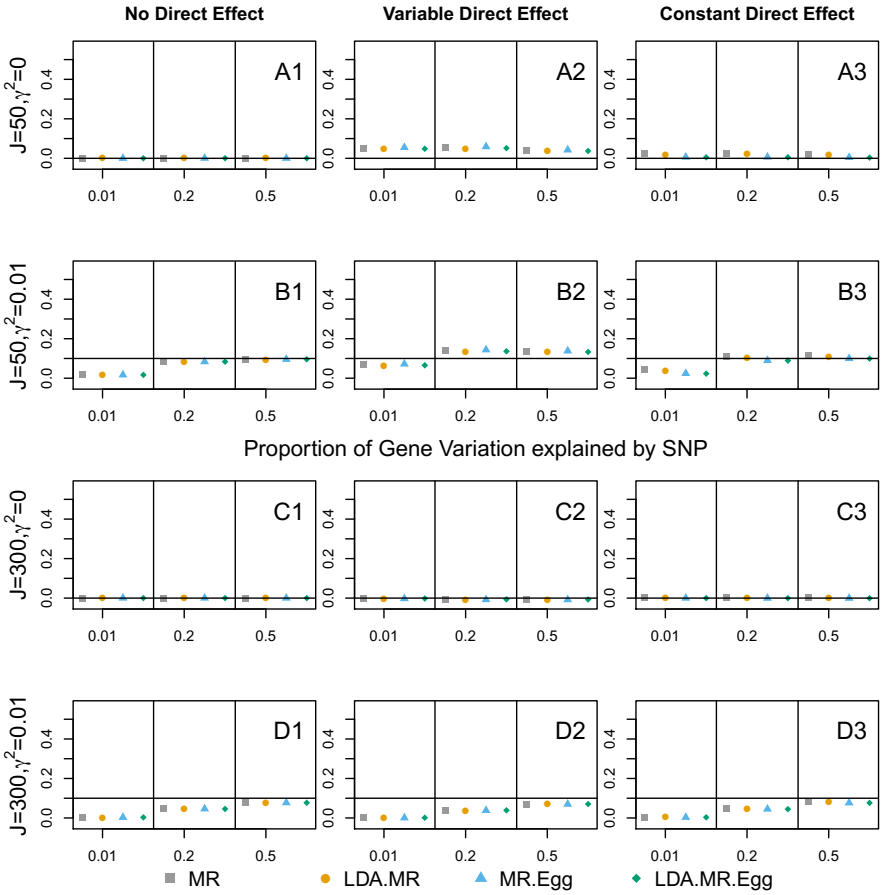

Figure 4.28: Bias plots for when there is low LD in the SNP set. First row corresponds to $J = 50$, $\gamma = 0$ (plots with A). Second panel (plots with B) when $J = 50$ and $\gamma^2 = 0.01$. Third panel (plots with C) when $J = 300$ and $\gamma = 0$. Final panel (plots with D) $J = 300$ and $\gamma^2 = 0.01$. From left to right: no direct effect, variable direct effect, constant direct effect.

4.3.2 Supplementary Tables

Gene	LDA MR	LDA MR-Egger	TWAS
PILRA	3.222e-07	1.669e-07	7.672e-06
PILRB	3.962e-08	2.753e-08	2.871e-06
ATG10	3.611e-27	2.143e-08	1.844e-10
PIDD	1.269e-06	2.074e-05	1.639e-09
L3MBTL3	1.075e-08	4.204e-09	3.671e-12
ATP6AP1L	4.407e-23	2.043e-10	4.207e-07
RP11-73O6.3	1.285e-06	7.399e-08	1.779e-11
LRRC37A2	4.797e-11	8.343e-09	3.403e-07
AP006621.6	2.109e-05	2.989e-05	1.261e-07
AP006621.5	3.119e-05	1.510e-05	6.626e-08
STAG3L5P-PVRIG2P-PILRB	3.071e-08	7.310e-09	3.174e-06

Table 4.13: Genes Bonferroni significant by TWAS, LDA MR, and LDA MR Egger.

Gene	LDA MR	LDA MR-Egger	TWAS
NUP107	1.000e-05	0.0008259	1.701e-05
LRRC37A	1.536e-10	0.0016355	1.963e-09
CPNE1	5.921e-07	0.0070146	3.795e-05
KANSL1-AS1	2.754e-05	0.0675970	2.963e-10

Table 4.14: Genes Bonferroni significant by LDA MR and TWAS but not LDA MR-Egger.

Gene	LDA MR	LDA MR-Egger	TWAS
AP006621.1	0.0001245	7.029e-05	2.646e-08
LINC00886	0.0001465	7.921e-06	1.776e-05

Table 4.15: Genes Bonferroni significant by TWAS and LDA MR-Egger but not LDA MR.

Gene	LDA MR	LDA MR-Egger	TWAS
PPIE	5.732e-06	9.504e-06	1.987e-01
GNPAT	1.298e-05	1.483e-09	5.821e-02
EPB41L4A	2.131e-11	2.883e-11	5.589e-01
ZSCAN29	6.570e-12	7.653e-09	9.242e-05
SLC26A1	3.935e-05	1.302e-05	9.987e-03
CISD2	2.513e-10	1.063e-07	1.989e-01
C6orf57	1.080e-06	5.419e-07	1.870e-01
UVSSA	3.161e-06	1.735e-05	2.707e-01
BDH2	6.524e-10	4.554e-06	2.532e-03
CATSPER2	4.238e-07	7.263e-09	7.682e-01
LINC00476	7.271e-05	4.557e-05	7.214e-01
RP11-54C4.1	1.007e-08	1.148e-08	1.659e-01
PINLYP	2.425e-07	1.485e-06	3.744e-04
ZSCAN31	1.521e-08	1.309e-08	1.533e-01
U91328.19	6.373e-05	5.636e-05	4.634e-01
RP11-351D16.3	2.951e-05	7.349e-06	3.920e-01

Table 4.16: Genes Bonferroni significant by LDA MR and LDA MR-Egger but not TWAS.

Gene	LDA MR	LDA MR-Egger	TWAS
CASP8	0.0297766	0.0105158	2.805e-14
CRHR1	0.0002619	0.0018539	1.288e-10
MAN2C1	0.1970724	0.5777374	1.723e-07
TRIM4	0.0001380	0.0012471	2.302e-05
SETD9	0.0121179	0.2774525	1.721e-10
ELP5	0.1583336	0.2012524	5.266e-05
YBEY	0.0299639	0.0157095	1.427e-05
BTN3A2	0.0834988	0.2506026	3.239e-05
CRHR1-IT1	0.0004111	0.0025725	1.228e-10
LRRC37A4P	0.0195509	0.0202474	4.013e-09
RASA4DP	0.2267255	0.3900677	4.528e-07
GABPB1-AS1	0.0006807	0.0007000	5.885e-05
CTD-3110H11.1	0.0006901	0.0008831	2.974e-05
RP11-554A11.9	0.0341998	0.2405126	4.654e-06
CTD-2323K18.1	0.0087171	0.0161156	1.312e-05

Table 4.17: Genes Bonferroni significant by just TWAS.

Gene	LDA MR	LDA MR-Egger	TWAS
TRIP4	1.119e-06	5.806e-01	0.340119
RAD51C	1.022e-05	1.103e-03	0.013972
TRIM37	5.486e-06	7.485e-05	0.010489
NKTR	3.681e-08	1.037e-03	0.792067
ZNF391	6.315e-05	2.261e-01	0.344649
ADAL	5.732e-10	5.010e-01	0.270200
ZNF354A	4.028e-06	4.185e-04	0.372818
LYSMD4	6.395e-05	5.511e-01	0.008325
RP5-874C20.3	1.225e-07	6.628e-04	0.278504
ENTPD3-AS1	8.418e-09	3.029e-03	0.010911
GS1-124K5.4	4.845e-05	6.028e-01	0.915393
RP3-465N24.5	9.876e-06	1.934e-02	0.809040
CTC-459F4.1	4.477e-05	5.650e-02	0.483305
SUZ12P	3.085e-10	4.640e-01	0.226856
SLC25A1P5	2.683e-05	1.101e-02	0.809859
CTC-459F4.3	6.378e-05	1.512e-02	0.675142
RP11-502I4.3	1.315e-05	4.720e-01	0.010634
RP4-758J24.5	4.656e-08	6.600e-03	0.518255
CTA-390C10.10	1.009e-07	4.238e-02	0.012089

Table 4.18: Genes Bonferroni significant by just LDA MR.

	LDA MR	LDA MR-Egger	TWAS
FIGNL1	0.4853841	3.380e-05	0.37396
C12orf52	0.2034739	6.044e-07	0.54945
CSGALNACT2	0.0001389	3.235e-05	0.53816
ZNF780A	0.0267498	6.905e-05	0.36309
C6orf163	0.0016959	4.441e-08	0.13725
TYW1B	0.8428725	1.798e-05	0.03721
AC005614.5	0.0745431	1.036e-06	0.90817

Table 4.19: Genes Bonferroni significant by just LDA MR-Egger.