DISENTANGLING LINKAGE AND POPULATION STRUCTURE IN ASSOCIATION MAPPING

HANBIN LEE[†,‡] AND MOOHYUK LEE[†]

[†]*Department of Medicine, Seoul National University College of Medicine*
Seoul, Republic of Korea
[‡]*Department of Mathematical Sciences, Seoul National University*
Seoul, Republic of Korea

Abstract. Genome-wide association study (GWAS) tests single nucleotide polymorphism (SNP) markers across the genome to localize the underlying causal variant of a trait. Because causal variants are seldom observed directly, a surrogate model based on genotyped markers are widely considered. Although many methods estimating the parameters of the surrogate model have been proposed, the connection between the surrogate model and the true causal model is yet investigated. In this work, we establish the connection between the surrogate model and the true causal model. Our theory shows the importance of the underlying demographic model in understanding the signal in GWAS.

1. Introduction

Genome-wide association study (GWAS) identifies regions in the genome responsible for the variation in a trait through *single nucleotide polymorphisms* (SNPs) distributed across the genome. Each SNP is tested for its association with the trait (generally through regression models) and its regression coefficient (henceforth, marker effect size) is recorded. Although markers themselves may not be causal, a significant effect size hints a true causal variant nearby the marker being tested. This is due to linkage, the correlation between two physically proximal variants. Earlier studies have found that SNPs are dense and widespread across the genome [1, 2]. Therefore, testing SNPs across the genome enables the identification of genetic signals genome-wide even if some causal variants are missing which gives the name of GWAS.

For continuous traits, the *quantitative trait model* is widely adopted where the trait is an additive function respect to the observed SNPs. From now on, we call this model as the *marker-additive model* (MAM). The estimation of MAM parameters have been extensively studied and many methods have been proposed [3, 4, 5]. MAM, however, is a surrogate model where the parameters (the marker effect sizes) have no direct causal interpretation since SNP markers are seldom causal themselves. Furthermore, both empirical and theoretical work have shown that stratification of allele frequencies due to population structure can produce spurious signals without linkage [6].

In this work, we consider an additive model respect to the causal variants, the *causal-additive model* (CAM), and establish the connection with the MAM. CAM parameters have direct causal interpretation because it directly involves often unobserved causal variants of the trait. We deliver a formula showing that marker effect size can be decomposed into three terms that reflect linkage and population structure separately. Best to our knowledge, our work is the first attempt to address the identification problem of covariate adjusted regression in GWAS.

E-mail address: hanbin973@snu.ac.kr.

2. Setup

2.1. **Notations and Models.** The data discussed in this work is $(Y_i, \mathbf{C}_i, \mathbf{M}_i, \mathbf{S}_i)_{i=1}^n$ where $Y_i \in \mathbb{R}$ is the trait, $\mathbf{C}_i \in \mathbb{R}^p$ is the causal variants, $\mathbf{M}_i \in \mathbb{R}^q$ is the markers and $\mathbf{S}_i$ is the population membership. In this study, we focus on diploid populations where elements of $\mathbf{C}_i$ and $\mathbf{M}_i$ take values $0, 1, 2$. $h = \mathbf{m}, \mathbf{p}$ denotes the haplotype index ($\mathbf{m}$ for maternal and $\mathbf{p}$ for paternal haplotype) so that $C_{ijh} = 0, 1$ and $M_{ikh} = 0, 1$, accordingly. The *marker-additive model* (MAM) is

$$Y_i = \mathbf{M}_i^T \boldsymbol{\beta} + \beta_0 + \delta_i \quad \text{and} \quad \mathbb{E}[\delta_i \mid \mathbf{M}_i] = 0 \tag{2.1}$$

where $\boldsymbol{\beta} \in \mathbb{R}^q$ is the marker effect size and δ_i is the noise variable. In most cases with very few exceptions, markers are tested marginally rather than as a whole. Without loss of generality, we write M_{i1} as the marker being tested which leads to the following regression.

$$Y_i = M_{i1}\beta_1 + \mathbf{M}_{i(-1)}^T \boldsymbol{\beta}_{(-1)} + \beta_0 + \delta_i \tag{2.2}$$

where (-1) in the subscript means that the first element is dropped from the original variable. Although model (2.1) contains all genotyped markers, in marginal testing, markers that are physically close to the variant being tested are removed from the regression. One such way is the *leave one chromosome out* (LOCO) approach where markers appearing in $\mathbf{M}_{i(-1)}$ are all sampled outside the chromosome which contains the variant being tested [7].

The *causal-additive model* (CAM), analogous to MAM, is

$$Y_i = \mathbf{C}_i^T \boldsymbol{\alpha} + \alpha_0 + \epsilon_i \quad \text{and} \quad \mathbb{E}[\epsilon_i \mid \mathbf{C}_i] = 0 \tag{2.3}$$

where $\boldsymbol{\alpha} \in \mathbb{R}^p$ is the causal effect size and ϵ_i is the noise variable. Since changes in $\mathbf{C}_i$ has direct influence on Y_i, $\boldsymbol{\alpha}$ has a straightforward causal interpretation.

Several population structure models are discussed in this work. When the sample is obtained from L discrete populations indexed by l, $\mathbf{S}_i \in \mathbb{R}^L$ where

$$\mathbf{S}_i = (0, \ldots, 0, \ldots, 0) \text{ if the } i\text{-th individual is from population } l = 1$$

or

$$\mathbf{S}_i = (0, \ldots, \underset{\substack{\uparrow \\ l-\text{th}}}{1}, \ldots, 0) \text{ if the } i\text{-th individual is from population } l \neq 1$$

The *Pritchard-Stephens-Donnelly* (PSD) model is a more general model that describes population structure incorporating *admixture*, where the genetic materials of an individual originates from multiple ancestries (indexed by l with total L of them) rather than a single one [8]. In this case, $\mathbf{S}_i$ is often called the *admixture proportion* where

$$\mathbf{S}_i = (a_{i1}, \ldots, a_{il}, \ldots, a_{iL})$$

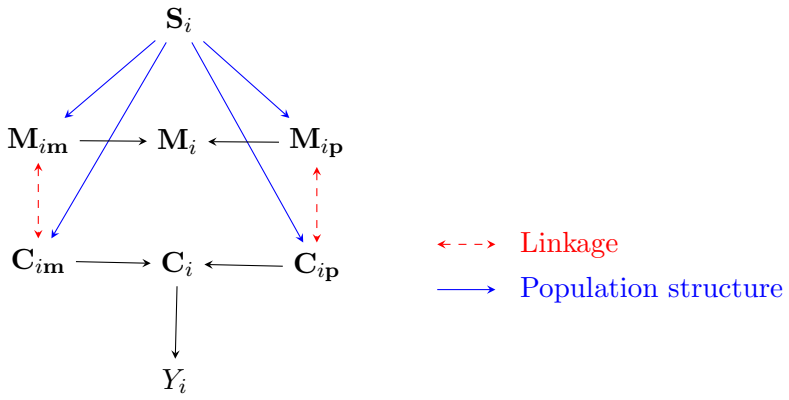
with $\sum_{l=1}^L a_{il} = 1$. The l-th element is the proportion of genome of an individual that originates from the l-th ancestral population. If one restricts the elements of $\mathbf{S}_i$ to take values 0 or 1, the PSD model reduces to the discrete populations model. The PSD model further assumes that the *Hardy-Weinberg Equilibrium* (HWE) holds conditional on $\mathbf{S}_i$. That being said, random mating is assumed and as a consequence, an individual receives its genome by randomly selecting a haplotype from the parent generation.

We define haplotype frequencies and *linkage disequilibrium* (LD) parameters. $f_j = \mathbb{P}(C_{ijh} = 1)$, $g_k = \mathbb{P}(M_{ikh} = 1)$ and $h_{jk} = \mathbb{P}(C_{ijh} = 1, M_{ikh} = 1)$ are allele frequencies and joint allele frequency of causal variant j, marker variant k and both variant j and variant k. LD covariance $D_{jk} = h_{jk} - f_j h_k$

is the covariance of variant j and k. When considering population membership, population specific analogues should be defined. $f_j^{\mathbf{s}} = \mathbb{P}(C_{ijh} = 1 \mid \mathbf{S}_i = \mathbf{s})$, $g_k^{\mathbf{s}} = \mathbb{P}(M_{ikh} = 1 \mid \mathbf{S}_i = \mathbf{s})$ and $h_{jk}^{\mathbf{s}} = \mathbb{P}(C_{ijh} = 1, M_{ikh} = 1 \mid \mathbf{S}_i = \mathbf{s})$ are the population specific allele frequencies and joint allele frequency. Population specific LD covariance is defined accordingly $D_{jk}^{\mathbf{s}} = h_{jk}^{\mathbf{s}} - f_j^{\mathbf{s}} h_k^{\mathbf{s}}$.

Finally, linear projection of X respect to Y, $\mathrm{L}[X \mid Y]$, will be frequently used throughout [9]. Note that the variables appearing on the right side the vertical sign (|) only contain the random components, and the intercept 1 is always included but ignored in the notation for convenience.

2.2. **The Causal Structure.** The conditional independence implied by the causal structure plays a pivotal role in this work. The causal structure is summarized in the following causal diagram.



The key condition we later use implied by the diagram is $\mathbf{M}_{i\mathbf{m}}, \mathbf{C}_{i\mathbf{m}} \perp\!\!\!\perp \mathbf{M}_{i\mathbf{p}}, \mathbf{C}_{i\mathbf{p}} \mid \mathbf{S}_i$. In other words, the genotype at different haplotypes are conditionally independent which is also implied by the conditional HWE assumption.

The diagram shows the two sources of association between the marker ($\mathbf{M}_i$) and the trait (Y_i):

(1) $\mathbf{M}_{ih} \dashleftarrow\!\dashrightarrow \mathbf{C}_{ih} \longrightarrow Y_i$: Association due to linkage.

(2) $\mathbf{M}_{ih} \longleftarrow \mathbf{S}_i \longrightarrow \mathbf{C}_{ih} \longrightarrow Y_i$: Association due to population structure.

In GWAS, we only want the first association to be present. This can be achieved by conditioning on $\mathbf{S}_i$, but it is generally not observed in population samples. We later show that this problem is partially, but not fully, accounted by including many markers to the regression as in (2.2).

2.3. **The Goal.** Our goal is to characterize the estimand of the regression (2.2) under the CAM (2.3). Following the practice of excluding variants that are near to the marker being tested, we assume that $M_{i1} \perp\!\!\!\perp \mathbf{M}_{i(-1)} \mid \mathbf{S}_i$. Due to the causal structure, this is equivalent to saying that M_{i1} is unlinked to the rest of the variants in the regression. Hence, conditional on $\mathbf{S}_i$, the following regression and regression (2.2) has the same estimand.

$$Y_i = M_{i1}\beta_{1,\text{nocov}} + \beta_{0,\text{nocov}} + \delta_{i,\text{nocov}} \tag{2.4}$$

i.e. $\beta_1 = \beta_{1,\text{nocov}}$ when the regression is restricted to a fixed $\mathbf{S}_i$. Because the regression is conditioned on $\mathbf{S}_i$, β_1 only reflects route (1) in section (2.2). To distinguish it from the regression coefficient obtained from the whole sample without restricting $\mathbf{S}_i$, we write this estimand as $\beta_1(\mathbf{S}_i)$. It can be shown that

$$\begin{aligned}\beta_1(\mathbf{S}_i = \mathbf{s}) &= \sum_j \frac{D^{\mathbf{s}}_{j1}}{g^{\mathbf{s}}_1(1-g^{\mathbf{s}}_1)}\alpha_j \\ &= \sum_j \beta_{1j}(\mathbf{S}_i = \mathbf{s})\end{aligned} \tag{2.5}$$

by the formula of univariate linear regression. Given that $\mathbb{E}[C_{ij} \mid \mathbf{S}_i]$ and $\mathbb{E}[M_{ik} \mid \mathbf{S}_i]$ are linear functions respect to $\mathbf{S}_i$ (e.g. admixture), the regression applied to the whole population

$$Y_i = M_{1i}\beta_{1,\mathbf{s}} + \mathbf{S}_i\boldsymbol{\gamma}_{\mathbf{s}} + \beta_{0,\mathbf{s}} + \delta_{i,\mathbf{s}}$$

has the estimand

$$\begin{aligned}\beta_{1,\mathbf{s}} &= \sum_j \frac{\mathbb{E}_{\mathbf{S}}[\beta_{1j}(\mathbf{S}_i)\mathrm{Var}(M_{i1} \mid \mathbf{S}_i)]}{\mathbb{E}_{\mathbf{S}}[\mathrm{Var}(M_{i1} \mid \mathbf{S}_i)]} \\ &= \frac{\mathbb{E}_{\mathbf{S}}[\beta_1(\mathbf{S}_i)\mathrm{Var}(M_{i1} \mid \mathbf{S}_i)]}{\mathbb{E}_{\mathbf{S}}[\mathrm{Var}(M_{i1} \mid \mathbf{S}_i)]}\end{aligned} \tag{2.6}$$

similar to the case of OLS under *heterogeneous treatment effect* (HTE) [10]. Equation (2.6) is a weighted average over $\beta_1(\mathbf{S}_i)$ with weights that sum up to 1.

In the whole sample, two regressions (2.2) and (2.4) have different estimands because $M_{i1} \not\perp \mathbf{M}_{i(-1)}$ in general. We aim to understand how β_1, the marker effect size obtained from regression (2.2) applied to the whole population, is related to $\beta_1(\mathbf{S}_i)$, the desired estimand obtained from ideal populations each in HWE.

3. Main Results

In GWAS, we are only interested in the contribution of linkage with a physically proximal causal variant. It should be desirable if (1) in section (2.2) can be separated from (2). However, we show that this is only partially achievable under the population-based design. Nevertheless, it is achievable under the within-sibship design [11].

3.1. Population-based GWAS.

We begin by stating the result for a simple regression (2.4) without any covariates.

Proposition 3.1. *The estimand of regression (2.4) on the whole population is*

$$\begin{aligned}\beta_{1,\mathrm{nocov}} &= \sum_j \underbrace{\frac{\mathbb{E}_{\mathbf{S}}[\beta_{1j}(\mathbf{S}_i)\mathrm{Var}(M_{i1} \mid \mathbf{S}_i)]}{\mathrm{Var}[M_{i1}]}}_{\textit{linkage}} + \sum_j \underbrace{\frac{\mathrm{Cov}_{\mathbf{S}}[\mathbb{E}(C_{ij} \mid \mathbf{S}_i), \mathbb{E}(M_{i1} \mid \mathbf{S}_i)]}{\mathrm{Var}[M_{i1}]} \cdot \alpha_j}_{\textit{population structure}} \\ &= \frac{\mathbb{E}_{\mathbf{S}}[\beta_1(\mathbf{S}_i)\mathrm{Var}(M_{i1} \mid \mathbf{S}_i)]}{\mathrm{Var}[M_{i1}]} + \sum_j \frac{\mathrm{Cov}_{\mathbf{S}}[\mathbb{E}(C_{ij} \mid \mathbf{S}_i), \mathbb{E}(M_{i1} \mid \mathbf{S}_i)]}{\mathrm{Var}[M_{i1}]} \cdot \alpha_j\end{aligned}$$

Remark 1 (The Wahlund effect)**.** *The first term can be viewed as a average of $\beta_1(\mathbf{S}_i)$ weighted by $\frac{\mathrm{Var}(M_{i1}|\mathbf{S}_i)\mathbb{P}(\mathbf{S}_i)}{\mathrm{Var}[M_{i1}]}$. An important observation is that the weights sum up to a value smaller than* 1 *as long as there is allele frequency stratification of M_{i1} across $\mathbf{S}_i$. To see this,*

$$\mathrm{Var}[M_{i1}] = \mathbb{E}_{\mathbf{S}}[\mathrm{Var}(M_{i1} \mid \mathbf{S}_i)] + \mathrm{Var}_{\mathbf{S}}[\mathbb{E}_{\mathbf{S}}(M_{i1} \mid \mathbf{S}_i)]$$

which is due to the law of total variance. This was first reported in 1928 by Wahlund [12]. *The sum over the weights can also be expressed in terms of Wright's F-statistics* [13] *which is a popular measure of population structure.*

$$\frac{\mathbb{E}_{\mathbf{S}}[\mathrm{Var}(M_{i1} \mid \mathbf{S}_i)]}{\mathrm{Var}[M_{i1}]} = 1 - F_{\mathrm{ST}}$$

A straightforward consequence of the Wahlund effect is that the desired effect $\beta_1(\mathbf{S}_i)$ *is attenuated and the attenuation is proportional to the strength of population structure, namely,* F_{ST}.

The second term due to population structure in **Proposition 3.1** has a more familiar interpretation. The numerator is non-zero if and only if the causal variant and the marker variant allele frequencies align simultaneously across $\mathbf{S}_i$. In sum, **Proposition 3.1** shows that population structure affects $\beta_{1,\mathrm{nocov}}$ in two folds. It attenuates the true signal and puts undesirable signals into the estimand.

Now we propose our main theorem which delivers the estimand of regression (2.2).

Theorem 3.2. *Let* $\widetilde{M}_{i1} = M_{i1} - \mathrm{L}[M_{i1} \mid \mathbf{M}_{i(-1)}]$. *The estimand of regression (2.2) is*

$$\begin{aligned}\beta_1 = &\sum_j \underbrace{\frac{\mathbb{E}_{\mathbf{S}}[\beta_{1j}(\mathbf{S}_i)\mathrm{Var}(M_{i1} \mid \mathbf{S}_i)]}{\mathrm{Var}[\widetilde{M}_{i1}]}}_{\textit{linkage}} \left(= \frac{\mathbb{E}_{\mathbf{S}}[\beta_1(\mathbf{S}_i)\mathrm{Var}(M_{i1} \mid \mathbf{S}_i)]}{\mathrm{Var}[\widetilde{M}_{i1}]} \right) \\ &+ \sum_j \underbrace{\frac{\mathbb{E}[C_{ij}(\mathbb{E}[M_{i1} \mid \mathbf{S}_i] - \mathbb{E}_{\mathbf{S}}[\mathbb{E}(M_{i1} \mid \mathbf{S}_i) \mid \mathbf{M}_{i(-1)}])]}{\mathrm{Var}[\widetilde{M}_{i1}]} \cdot \alpha_j}_{\textit{prediction error}} \\ &+ \sum_j \underbrace{\frac{\mathbb{E}[C_{ij}(\mathbb{E}[M_{i1} \mid \mathbf{M}_{i(-1)}] - \mathrm{L}[M_{i1} \mid \mathbf{M}_{i(-1)}])]}{\mathrm{Var}[\widetilde{M}_{i1}]} \cdot \alpha_j}_{\textit{functional misspecification}}\end{aligned}$$

Proposition 3.3. *The weights of the linkage term in* ***Theorem 3.2*** *sum up to a value smaller than* 1.

$$\frac{\mathbb{E}_{\mathbf{S}}[\mathrm{Var}(M_{i1} \mid \mathbf{S}_i)]}{\mathrm{Var}[\widetilde{M}_{i1}]} \leq 1$$

where the equality holds if and only if $\mathbb{E}[M_{i1} \mid \mathbf{M}_{i(-1)}] = \mathrm{L}[M_{i1} \mid \mathbf{M}_{i(-1)}]$.

Theorem 3.4. *Let* $\mathbf{M}^{(q)}_{i(-1)}$ *be a sequence of markers other than* 1 *with length* q. *Assume that* $\mathbf{M}^{(q)}_{i(-1)}$ *defines a monotonically increasing filteration* $\sigma\langle\mathbf{M}^{(q)}_{i(-1)}\rangle$ *respect to* q. *Under the casual structure of section (2.2),*

$$\mathbb{E}_{\mathbf{S}}[\mathbb{E}(M_{i1} \mid \mathbf{S}_i) \mid \mathbf{M}_{i(-1)}] \xrightarrow[q\to\infty]{} \mathbb{E}[M_{i1} \mid \mathbf{S}_i = \mathbf{s}]$$

in $L^1(\mathbb{P})$.

Theorem 3.4 tells us that with sufficiently large number of markers as covariates, the prediction error becomes negligible. This is very likely in modern GWAS since millions of variants are genotyped. We defer the detailed proof to the **Appendix** but provide a brief sketch here.

To prove the theorem, we resort ourselves to a more general problem where we consider the convergence of distributions, namely, $\mathbb{P}(\,\cdot \mid \mathbf{M}_{i(-1)}) \to \mathbb{P}(\,\cdot \mid \mathbf{S}_i = \mathbf{s})$ with $\mathbf{S}_i = \mathbf{s}$ fixed for an individual. If we view $\mathbf{s}$ as the true parameter, $\mathbb{P}(\mathbf{S}_i)$ as the prior distribution and $\mathbb{P}(\mathbf{S}_i \mid \mathbf{M}_{i(-1)})$ (where $\mathbf{M}_{i(-1)}$ is the data) as the posterior distribution, the convergence is simply a *posterior consistency* problem in Bayesian statistics [14, 15]. Loosely speaking, posterior consistency means that the posterior distribution degenerates to a point mass as the number of data increases. Hence, we evoke the Doob's theorem to prove our statement [15]. The theorem tells us that we can exactly predict the population membership $\mathbf{S}_i$ with large number of markers.

Remark 2. *The functional misspecification* $\mathbb{E}[M_{i1} \mid \mathbf{M}_{i(-1)}] - \mathrm{L}[M_{i1} \mid \mathbf{M}_{i(-1)}]$ *is generally nonzero. One concrete example is the two discrete population model. A closed-form analysis can be delivered assuming that elements of* $\mathbf{M}_{i(-1)}$ *are mutually unlinked. By the Bayes theorem,*

$$\mathbb{P}(\mathbf{S}_i \mid \mathbf{M}_{i(-1)}) = \frac{\mathbb{P}(\mathbf{M}_{i(-1)} \mid \mathbf{S}_i)\, \mathbb{P}(\mathbf{S}_i)}{\mathbb{P}(\mathbf{M}_{i(-1)})}$$

This gives

$$\begin{aligned}\frac{\mathbb{P}(\mathbf{S}_i = 1 \mid \mathbf{M}_{i(-1)})}{1 - \mathbb{P}(\mathbf{S}_i = 1 \mid \mathbf{M}_{i(-1)})} &= \frac{\mathbb{P}(\mathbf{S}_i = 1 \mid \mathbf{M}_{i(-1)})}{\mathbb{P}(\mathbf{S}_i = 0 \mid \mathbf{M}_{i(-1)})} \\ &= \frac{\mathbb{P}(\mathbf{M}_{i(-1)} \mid \mathbf{S}_i = 1)}{\mathbb{P}(\mathbf{M}_{i(-1)} \mid \mathbf{S}_i = 0)} \cdot \frac{\mathbb{P}(\mathbf{S}_i = 1)}{\mathbb{P}(\mathbf{S}_i = 0)}\end{aligned}$$

On the other hand, $\mathbb{P}(\mathbf{M}_{i(-1)} \mid \mathbf{S}_i)$ *is factorized due to the conditional independence* $M_{ik} \perp\!\!\!\perp M_{ik'} \mid \mathbf{S}_i$.

$$\begin{aligned}\mathbb{P}(\mathbf{M}_{i(-1)} \mid \mathbf{S}_i = \mathbf{s}) &= \prod_{k \neq 1} \mathbb{P}(M_{ik} \mid \mathbf{S}_i) \\ &= \prod_{k \neq 1} \binom{2}{M_{ik}} (g_k^{\mathbf{s}})^{M_{ik}} (1 - g_k^{\mathbf{s}})^{2 - M_{ik}}\end{aligned}$$

Finally, applying $\log(\,\cdot\,)$ *on both sides of the equation shows that* $\mathbb{E}[\mathbf{S}_i \mid \mathbf{M}_{i(-1)}]$ *is a logit-linear function respect to* $\mathbf{M}_{i(-1)}$.

$$\operatorname{logit}(\mathbb{E}[\mathbf{S}_i \mid M_{i(-1)}]) = \sum_{k \neq 1} \left[M_{ik} \log\left(\frac{g_k^1}{g_k^0}\right) + (2 - M_{ik}) \log\left(\frac{1 - g_k^1}{1 - g_k^0}\right) \right] + \log\left(\frac{\mathbb{P}(\mathbf{S}_i = 1)}{\mathbb{P}(\mathbf{S}_i = 0)}\right)$$

In sum, even with handful of covariates, the non-vanishing functional misspecification attenuates the true signal according to **Proposition 3.3** and leaves an additional confounding term according to **Theorem 3.2**.

3.2. **Within-sibship GWAS.** In within-sibship GWAS, we observe the family membership of each individual. Denote the family membership of individual i as $\mathbf{F}_i$. The families are index by f.

Marginal testing of markers are performed using the following regression.

$$Y_i = M_{i1}\beta_{1,\mathbf{f}} + \sum_f \mathbf{1}(\mathbf{F}_i = f)\gamma_f + \beta_0 + \delta_i \tag{3.1}$$

The OLS estimator of regression (3.1) is equivalent to the famous sibling difference regression when there are only two siblings per family [16].

$$Y_i - Y_{i'} = (M_{i1} - M_{i'1})\beta_{1,\text{sd}} + (\delta_i - \delta_{i'}) \quad \text{where} \quad \mathbf{F}_i = \mathbf{F}_{i'} \tag{3.2}$$

The proof of the algebraic equivalence can be found in Wooldrdige (2021) [17].

The estimand of regression (3.1) is

$$\begin{aligned} \beta_{1,\mathbf{f}} &= \sum_j \frac{\mathbb{E}_{\mathbf{F}}[\beta_{1j}(\mathbf{F}_i)\text{Var}(M_{i1} \mid \mathbf{F}_i)]}{\mathbb{E}_{\mathbf{F}}[\text{Var}(M_{i1} \mid \mathbf{F}_i)]} \\ &= \frac{\mathbb{E}_{\mathbf{F}}[\beta_1(\mathbf{F}_i)\text{Var}(M_{i1} \mid \mathbf{F}_i)]}{\mathbb{E}_{\mathbf{F}}[\text{Var}(M_{i1} \mid \mathbf{F}_i)]} \end{aligned} \tag{3.3}$$

where $\beta_1(\mathbf{F}_i = f)$ is the estimand of regression (3.1) restricted to family $\mathbf{F}_i = f$. By combining the facts (1) regressions (3.1) and (3.2) have algebraically identical OLS estimators and (2) OLS consistently estimates $\beta_{1,\mathbf{f}}$, we show that the estimand $\beta_{1,\mathbf{f}}$ is equal to $\beta_{1,\mathbf{s}}$, the population estimand when $\mathbf{S}_i$ is known.

Theorem 3.5. *If we ignore recombination,*

$$\frac{\mathbb{E}_{\mathbf{F}}[\beta_1(\mathbf{F}_i)\text{Var}(M_{i1} \mid \mathbf{F}_i)]}{\mathbb{E}_{\mathbf{F}}[\text{Var}(M_{i1} \mid \mathbf{F}_i)]} = \frac{\mathbb{E}_{\mathbf{S}}[\beta_1(\mathbf{S}_i)\text{Var}(M_{i1} \mid \mathbf{S}_i)]}{\mathbb{E}_{\mathbf{S}}[\text{Var}(M_{i1} \mid \mathbf{S}_i)]}$$

3.3. **Non-genetic confounding.** We initially assumed that $\mathbb{E}[\epsilon_i \mid \mathbf{C}_i, \mathbf{M}_i] = 0$ in the CAM model (2.3). This can be relaxed so that $\epsilon_i = \nu_i + U_i$ where $\mathbb{E}[\nu_i \mid \mathbf{C}_i, \mathbf{M}_i] = 0$ and $\mathbb{E}[U_i] = 0$. U_i may contain non-genetic random variables.

Equation (2.6) still holds provided

(1) $\mathbf{M}_i, \mathbf{C}_i \longleftarrow \mathbf{S}_i \longrightarrow U_i$: U_i is correlated with the genetic variables only through $\mathbf{S}_i$.
(2) $\mathbb{E}[U_i \mid \mathbf{S}_i]$ is linear respect to $\mathbf{S}_i$.

which are identical to the conditions in which genetic confounding is resolved with linear regression.

The estimand using regression (2.2) is also similar to **Theorem 3.2** with two additional terms.

$$\begin{aligned} &+ \underbrace{\frac{\mathbb{E}[U_i(\mathbb{E}[M_{i1} \mid \mathbf{S}_i] - \mathbb{E}_{\mathbf{S}}[\mathbb{E}(M_{i1} \mid \mathbf{S}_i) \mid \mathbf{M}_{i(-1)}])]}{\text{Var}[\widetilde{M}_{i1}]}}_{\text{prediction error}} \\ &+ \underbrace{\frac{\mathbb{E}[U_i(\mathbb{E}[M_{i1} \mid \mathbf{M}_{i(-1)}] - \text{L}[M_{i1} \mid \mathbf{M}_{i(-1)}])]}{\text{Var}[\widetilde{M}_{i1}]}}_{\text{functional misspecification}} \end{aligned} \tag{3.4}$$

Prediction error vanishes and the functional misspecification term is generally non-zero as before. The derivation is essentially the same as in **Theorem 3.2** in which C_{ij} is simply replaced by U_i. Note that the result does not depend on the two conditions in the previous paragraph. The formula shows how non-genetic confounding is partially resolved with genetic markers in linear regression: genetic markers indirectly controls the non-genetic variables by predicting $\mathbf{S}_i$.

4. Conclusions

In this work, we aimed to solve the identification problem of GWAS of quantitative traits using linear regression in a structured population consisting of subpopulations that are conditionally in HWE. We established the connection between the CAM and the MAM which provides a closed-form formula for what is being estimated using linear regression. The formula shows that under the popular population design, population structure exhibits a two-fold effect in which it induces

an additive confounding term together with an attenuation of the true effect of a causal variant. As expected, within-sibship design can overcome this problem due to direct access to family membership.

We expect our framework to be extended to incorporate other important evolutionary processes such as assortative mating and inbreeding. As the (conditional) independence between the two haplotypes of an individual plays an important role in our results and proofs, haplotype dependence induced by such evolutionary processes is likely to have an non-trivial impact on GWAS estimands.

One shortcoming of our work is that it only deals with identification and tells little about the estimation process. Among popular methods, only *linear mixed models* (LMMs) exactly conform to equation (2.2). Since PC correction applies *principal component analysis* (PCA) to $\mathbf{M}_{i(-1)}$ prior to regression, it is a biased estimator for β_1 [18].

This leaves an interesting conjecture in the light of our theory together with previous works on PCA which show that it asymptotically recovers admixture proportion correctly [19, 20]. Hence, depending on the model of population structure, PC correction may deliver the correct estimand while LMMs do not: PC correction does twice wrong (the wrong estimand and the wrong estimator), accidentally reaching the right answer while LMMs are wrong once (the wrong estimand but the right estimator), reaching a wrong answer in the end, contradicting common wisdom [21].

Acknowledgements

We thank the following colleagues who provided helpful feedbacks after reading an early version of the draft. Our work would have been impossible without them. Doc Edge (University of Southern California, US) gave important comments on a population genetic perspective. Qingyuan Zhao (University of Cambridge, UK) suggested recent literature in statistics and probability related to our work.

Proofs

Proof of ***Proposition 3.1****.* It follows from **Theorem 3.2** by setting $\mathbf{M}_{i(-1)}$ empty. □

Proof of ***Theorem 3.2****.* By the Frisch-Waugh-Lovell (FWL) theorem [9], the estimand of regression (2.2) is

$$\beta_1 = \frac{\mathbb{E}[Y_i \widetilde{M}_{i1}]}{\mathbb{E}[\widetilde{M}_{i1}^2]}$$

and the estimand of the regression restricted to $\mathbf{S}_i = \mathbf{s}$ is

$$\beta_1(\mathbf{S}_i = \mathbf{s}) = \frac{\mathbb{E}[Y_i(M_{i1} - \mathbb{E}[M_{i1} \mid \mathbf{S}_i = \mathbf{s}])]}{\mathrm{Var}[M_{i1} \mid \mathbf{S}_i]}$$

Now we expand the numerator of β_1. Substituting Y_i with the CAM model (2.3) gives

$$\begin{aligned}
\mathbb{E}[Y_i \widetilde{M}_{i1}] &= \sum_j \mathbb{E}[C_{ij} \widetilde{M}_{i1}] \alpha_j + \mathbb{E}[\epsilon_i \widetilde{M}_{i1}] \\
&= \sum_j \mathbb{E}[C_{ij} \widetilde{M}_{i1}] \alpha_j \\
&= \sum_j \Big(\mathbb{E}[C_{ij} \dot{M}_{i1}] + \mathbb{E}[C_{ij} (\mathbb{E}[M_{i1} \mid \mathbf{S}_i] - \mathrm{L}[M_{i1} \mid \mathbf{M}_{i(-1)}])] \Big) \cdot \alpha_j
\end{aligned}$$

The first term $\mathbb{E}[C_{ij} \dot{M}_{i1}] \alpha_j$ can be expressed in terms of $\beta_{1j}(\mathbf{S}_i = \mathbf{s})$ by substituting the expression of it in the first paragraph of the proof.

$$\begin{aligned}\mathbb{E}[C_{ij}\dot{M}_{i1}]\alpha_j &= \mathbb{E}_{\mathbf{S}}[\mathbb{E}(C_{ij}\dot{M}_{i1} \mid \mathbf{S}_i)\alpha_j] \\ &= \mathbb{E}_{\mathbf{S}}[\beta_{1j}(\mathbf{S}_i)\mathrm{Var}(M_{i1} \mid \mathbf{S}_i)]\end{aligned}$$

The second term is expanded by adding and subtracting $E[M_{i1} \mid \mathbf{M}_{i(-1)}]$.

$$\begin{aligned}&\mathbb{E}[C_{ij}(\mathbb{E}[M_{i1} \mid \mathbf{S}_i] - \mathrm{L}[M_{i1} \mid \mathbf{M}_{i(-1)})] \\ &= \mathbb{E}[C_{ij}(\mathbb{E}[M_{i1} \mid \mathbf{S}_i] - \mathbb{E}[M_{i1} \mid \mathbf{M}_{i(-1)}])] + \mathbb{E}[C_{ij}(\mathbb{E}[M_{i1} \mid \mathbf{M}_{i(-1)}] - \mathrm{L}[M_{i1} \mid \mathbf{M}_{i(-1)}])]\end{aligned}$$

Finally, the following reformulation of $\mathbb{E}[M_{i1} \mid \mathbf{M}_{i(-1)}]$ completes the proof.

$$\begin{aligned}\mathbb{E}[M_{i1} \mid \mathbf{M}_{i(-1)}] &= \mathbb{E}_{\mathbf{S}}[\mathbb{E}(M_{i1} \mid \mathbf{M}_{i(-1)}, \mathbf{S}_i) \mid \mathbf{M}_{i(-1)}] \\ &= \mathbb{E}_{\mathbf{S}}[\mathbb{E}(M_{i1} \mid \mathbf{S}_i) \mid \mathbf{M}_{i(-1)}]\end{aligned}$$

due to $M_{i1} \perp\!\!\!\perp \mathbf{M}_{i(-1)} \mid \mathbf{S}_i$.

□

Proof of ***Proposition 3.3****.* Let $\dot{M}_{i1} = M_{i1} - \mathbb{E}[M_{i1} \mid \mathbf{S}_i]$.

$$\frac{\mathbb{E}_{\mathbf{S}}[\mathrm{Var}(M_{i1} \mid \mathbf{S}_i)]}{\mathrm{Var}[\widetilde{M}_{i1}]} = \frac{\mathbb{E}_{\mathbf{S}}[\mathrm{Var}(M_{i1} \mid \mathbf{S}_i)]}{\mathrm{Var}[\dot{M}_{i1}]} \cdot \frac{\mathrm{Var}[\dot{M}_{i1}]}{\mathrm{Var}[\widetilde{M}_{i1}]}$$

The first term is smaller than 1 due to the law of total variance.

Now we show that the second term is smaller than 1.

$$\begin{aligned}\mathrm{Var}[\dot{M}_{i1}] &= \mathbb{E}\left[(M_{i1} - \mathbb{E}[M_{i1} \mid \mathbf{S}_i])^2\right] \\ &= \mathbb{E}\left[(M_{i1} - \mathbb{E}[M_{i1} \mid \mathbf{S}_i, \mathbf{M}_{i(-1)}])^2\right] \\ &\leq \mathbb{E}\left[(M_{i1} - \mathbb{E}[M_{i1} \mid \mathbf{M}_{i(-1)}])^2\right] \\ &\leq \mathbb{E}\left[(M_{i1} - \mathrm{L}[M_{i1} \mid \mathbf{M}_{i(-1)}])^2\right] \\ &= \mathrm{Var}[\widetilde{M}_{i1}]\end{aligned}$$

The second line follows from the conditional independence implied by the causal structure. The third line follows from the property of conditional expectation. The forth line follows from the fact that the conditional expectation minimizes the square norm.

The result can be informally inferred in a intuitive way using the causal graph. Explaining the variance of M_{i1} with $\mathbf{S}_i$ is more effective than with $\mathbf{M}_{i(-1)}$ because the path from $\mathbf{M}_{i(-1)}$ from M_{i1}, $\mathbf{M}_{i(-1)} \leftarrow \mathbf{S}_i \rightarrow M_{i1}$, must go through the path $\mathbf{S}_i \rightarrow M_{i1}$.

□

Proof of ***Theorem 3.4****.* See **Theorem 6.9** of Ghosal and van der Vaart [15] (Doob's theorem). First, substitute the variables appearing in the theorem according to our notation. Replace $X^{(n)}$ to $\mathbf{M}_i^{(q)}$, $\sigma\langle X^{(1)}, X^{(2)}, \ldots\rangle$ to $\sigma\langle M_{i2}, M_{i3}, \ldots\rangle$, θ to $\mathbf{s}$ and Π to $\mathbb{P}(\mathbf{S}_i)$. Next, apply the theorem to $f(\mathbf{s}) = \mathbb{E}[M_{i1} \mid \mathbf{S}_i = \mathbf{s}]$ which gives the desired result. □

Proof of ***Theorem 3.5****.* The proof is based on Veller and Coop (2023) [cite]. For a fixed $\mathbf{S}_i = \mathbf{s}$, equation (7) of Veller and Coop shows that

$$\beta_{1,\mathbf{f}}(\mathbf{S}_i = \mathbf{s}) = \frac{2}{H_1(\mathbf{S}_i = \mathbf{s})} \sum_j D_{j1}^{\mathbf{s}} \alpha_j$$

where $H_1(\mathbf{S}_i = \mathbf{s}) = 2g_1^{\mathbf{s}}(1 - g_1^{\mathbf{s}})$ and $D_{j1}^{\mathbf{s}} = h_{j1}^{\mathbf{s}} - f_j^{\mathbf{s}} g_1^{\mathbf{s}}$. Therefore, $\beta_{1,\mathbf{f}}(\mathbf{S}_i = \mathbf{s}) = \beta_1(\mathbf{S}_i = \mathbf{s})$. Note that λ was replaced to 1 and l was replaced to j to match our notation. This implicitly assumes

that families are nested within populations. Finally, substituting the result to equation (3.1) gives the desired result.

□

References

[1] The International HapMap Consortium. A haplotype map of the human genome. *Nature*, 437(7063):1299–1320, oct 2005.

[2] The International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449(7164):851–861, oct 2007.

[3] Alkes L Price, Nick J Patterson, Robert M Plenge, et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*, 38(8):904–909, jul 2006.

[4] Hyun Min Kang, Jae Hoon Sul, Susan K Service, et al. Variance component model to account for sample structure in genome-wide association studies. *Nat Genet*, 42(4):348–354, mar 2010.

[5] Joelle Mbatchou, Leland Barnard, Joshua Backman, et al. Computationally efficient whole-genome regression for quantitative and binary traits. *Nat Genet*, 53(7):1097–1103, may 2021.

[6] Noah A Rosenberg and Magnus Nordborg. A general population-genetic model for the production by population structure of spurious genotype–phenotype associations in discrete, admixed or spatially distributed populations. *Genetics*, 173(3):1665–1678, jul 2006.

[7] Jian Yang, Noah A Zaitlen, Michael E Goddard, et al. Advantages and pitfalls in the application of mixed-model association methods. *Nat Genet*, 46(2):100–106, jan 2014.

[8] Jonathan K Pritchard, Matthew Stephens, and Peter Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959, jun 2000.

[9] Jeffrey M. Wooldridge. *Econometric Analysis of Cross Section and Panel Data, second edition*. The MIT Press, 10 2010.

[10] Guido W Imbens and Jeffrey M Wooldridge. Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*, 47(1):5–86, mar 2009.

[11] Laurence J. Howe, Michel G. Nivard, Tim T. Morris, et al. Within-sibship genome-wide association analyses decrease bias in estimates of direct genetic effects. *Nat Genet*, 54(5):581–592, may 2022.

[12] STEN WAHLUND. ZUSAMMENSETZUNG VON POPULATIONEN UND KORRELATIONSERSCHEINUNGEN VOM STANDPUNKT DER VERERBUNGSLEHRE AUS BETRACHTET. *Hereditas*, 11(1):65–106, jul 2010.

[13] Sewall Wright. ISOLATION BY DISTANCE. *Genetics*, 28(2):114–138, mar 1943.

[14] Subhashis Ghosal and Aad van der Vaart. Convergence rates of posterior distributions for noniid observations. *Ann. Statist.*, 35(1), feb 2007.

[15] Subhashis Ghosal and Aad van der Vaart. *Fundamentals of Nonparametric Bayesian Inference*. Cambridge University Press, 06 2017.

[16] Ben Brumpton, Eleanor Sanderson, Karl Heilbron, et al. Avoiding dynastic, assortative mating, and population stratification biases in mendelian randomization through within-family analyses. *Nat Commun*, 11(1), jul 2020.

[17] Jeffrey M. Wooldridge. Two-way fixed effects, the two-way mundlak regression, and difference-in-differences estimators. *SSRN Journal Electronic Journal*, 2021.

[18] The Tien Mai and Pierre Alquier. Understanding the population structure correction regression. In *Proceedings of the 4th International Conference on Statistics: Theory and Applications*. Avestia Publishing, aug 2022.

[19] Gil McVean. A genealogical interpretation of principal components analysis. *PLoS Genet Genetics*, 5(10):e1000686, oct 2009.

[20] Xiuwen Zheng and Bruce S. Weir. Eigenanalysis of SNP data with an identity by descent interpretation. *Theoretical Population Biology*, 107:65–76, feb 2016.

[21] Arthur Korte, Bjarni J Vilhjálmsson, Vincent Segura, et al. A mixed-model approach for genome-wide association studies of correlated traits in structured populations. *Nat Genet*, 44(9):1066–1071, aug 2012.