The effect of population kinship estimation bias in heritability estimation and genetic association

Zhuoran Hou[1], Alejandro Ochoa[1,2,*]

[1] Department of Biostatistics and Bioinformatics, Duke University, Durham, NC 27705, USA

[2] Duke Center for Statistical Genetics and Genomics, Duke University, Durham, NC 27705, USA

[*] Corresponding author: alejandro.ochoa@duke.edu

Abstract

Population kinship matrices are estimated for a variety of applications, including estimation of heritability and to control for population structure in genetic association studies. Recent work found that the most common family of kinship estimators can be severely biased. In this work, we investigate the effect of this kinship bias on the two downstream applications of heritability estimation and genetic association. We present a novel trait simulation strategy that accurately parametrizes heritability, even when utilizing real genotypes. Using these simulations, we find that heritability estimation becomes biased when using such biased kinship matrices. Remarkably however, this kinship bias does not affect genetic associations based on either Principal Components Analysis (PCA) or Linear Mixed-effects Models (LMM). Lastly, we explain our empirical observations using theory. In particular, the exact form of the bias of the standard kinship estimator is such that it is compensated for by fitting the intercept in both PCA and LMM approaches, which model population structure via covariates, suggesting that only downstream applications with this precise arrangement will be robust to this kinship bias.

1 Introduction

Kinship is utilized in principal components analyses and linear-mixed effects models to correct for structure in Genome-Wide Association Studies (GWAS) (Xie et al., 1998; Yu et al., 2006; Aulchenko et al., 2007; Price et al., 2006; Astle and Balding, 2009; Kang et al., 2008; Kang et al., 2010; Zhou and Stephens, 2012; Loh et al., 2015; Sul et al., 2018) and to estimate genome-wide heritability (Yang et al., 2010; Yang et al., 2011; Speed et al., 2012). The most commonly-used kinship estimator (Price et al., 2006; Astle and Balding, 2009; Rakovski and Stram, 2009; Thornton and McPeek, 2010; Yang et al., 2010; Yang et al., 2011; Zhou and Stephens, 2012; Speed et al., 2012; Speed and Balding, 2015; Loh et al., 2015; Wang et al., 2017; Sul et al., 2018) was recently determined to have a complex bias (Ochoa and Storey, 2021; Weir and Goudet, 2017).

Heritability has long been estimated from the trait correlation between close relatives, particularly identical and fraternal twins, or siblings (Falconer and Mackay, 1996). These initial estimation approaches were based on a regression model (Falconer and Mackay, 1996). A different model was needed to achieve accurate estimates when estimating on more distant relatives, such as the variance component model of the SOLAR approach (Almasy and Blangero, 1998), which is based on estimating kinship from pedigrees. This idea was extended from pedigrees to populations with the GCTA approach (Yang et al., 2010; Yang et al., 2011), which has spawned much research (Visscher et al., 2010; Speed et al., 2012; Krishna Kumar et al., 2016; Yang et al., 2016; Jiang et al., 2016). However, we previously found that the most common family of kinship estimators employed by these approaches for populations can be severely biased (Ochoa and Storey, 2021). Speed et al. (2012) focused on some biases related to other assumptions of the multivariate normal heritability model, but biases in kinship estimation were not known or included in these evaluations.

There are alternative approaches for estimating heritability, such as LD Score Regression (Bulik-Sullivan et al., 2015) and SumHer (Speed and Balding, 2019), which we do not consider in this work since they are not based on kinship matrices.

2 Methods

2.1 Genetic model

Suppose there are m biallelic loci and n diploid individuals. The genotype $x_{ij} \in \{0, 1, 2\}$ at a locus i of individual j is encoded as the number of reference alleles, for a preselected but otherwise arbitrary reference allele per locus. These genotypes can be treated as random variables structured according to relatedness. If φ_{jk} is the kinship coefficient of two individuals j and k, and p_i is the ancestral allele frequency at locus i, then under the kinship model (Ochoa and Storey, 2016; Ochoa and Storey, 2021) the expectation and covariance are given by

$$\mathrm{E}[\mathbf{X}] = 2\mathbf{p}\mathbf{1}_n^\intercal, \qquad \mathrm{Cov}(\mathbf{x}_i) = 4p_i(1 - p_i)\mathbf{\Phi},$$

where $\mathbf{x}_i$ is the length-n column vector of genotypes at locus i, $\mathbf{X} = (\mathbf{x}_i^\intercal)$ is the complete $m \times n$ genotype matrix, $\mathbf{\Phi} = (\varphi_{jk})$ is the $n \times n$ kinship matrix, $\mathbf{p} = (p_i)$ is a length-m column vector of ancestral allele frequencies, $\mathbf{1}_n = (1)$ is a length-n column vector where every element is 1, and the $\intercal$ superscript denotes matrix transposition. Both kinship ($\mathbf{\Phi}$) and ancestral allele frequencies ($\mathbf{p}$) are parameters that depend on the choice of ancestral population, for which the Most Recent Common Ancestor (MRCA) population is the most sensible choice (Ochoa and Storey, 2016; Ochoa and Storey, 2021). In this work, to simplify notation, we omit cumbersome notation that marks this dependence of parameters on the choice of ancestral population, not do we explicitly condition on the ancestral population when calculating expectations and covariances as done in previous work, although it is done implicitly. This and later notation is summarized in Table 1.

The length-n quantitative trait vector $\mathbf{y}$ for all individuals is assumed to follow a linear polygenic model,

$$\mathbf{y} = \mathbf{1}_n\alpha + \mathbf{X}^\intercal\beta + \epsilon, \tag{1}$$

where α is the intercept coefficient, $\beta = (\beta_i)$ is a length-m column vector of effect size coefficients for each locus i (which may be zero), and ϵ is a length-n column vector of non-genetic effects. To analyze the covariance structure of the trait, we shall treat α and β are fixed parameters, while $\mathbf{X}$ and ϵ are random. The non-genetic effects are assumed to be independent with variance $(1 - h^2)\sigma^2$

Table 1: **Mathematical notation.**

Variable	Dimensions	Description
m	1	Number of loci
n	1	Number of individuals
i	1	Locus (variant) index
j, k	1	Individual indexes
μ	1	Trait mean
σ^2	1	Trait variance scale
h^2	1	(Narrow-sense) Heritability
$\mathbf{X} = (x_{ij})$	$m \times n$	Genotype matrix
$\mathbf{x}_i = (x_{ij})$	$n \times 1$	Genotype vector at locus i
$\mathbf{y}$	$n \times 1$	Trait vetor
α	1	Intercept
$\beta = (\beta_i)$	$m \times 1$	Effect size coefficients
ϵ	$n \times 1$	Non-genetic random effect
$\mathbf{p} = (p_i)$	$m \times 1$	Ancestral allele frequencies
$\mathbf{\Phi} = (\varphi_{jk})$	$n \times n$	Kinship matrix
$\mathbf{1}_n$	$n \times 1$	Vector of ones
$\mathbf{I}_n$	$n \times n$	Identity matrix

given by the total trait variance scale σ^2 and the narrow-sense heritability h^2:

$$\mathrm{E}[\epsilon] = \mathbf{0}_n, \qquad \mathrm{Cov}(\epsilon) = (1-h^2)\sigma^2\mathbf{I}_n,$$

where $\mathbf{0}_n$ is a length-n column vector of zeroes. The expectation of the trait is therefore

$$\mathrm{E}[\mathbf{y}] = \alpha\mathbf{1}_n + \mathrm{E}\left[\mathbf{X}^\intercal\right]\beta + \mathrm{E}[\epsilon] = \mu\mathbf{1}_n, \qquad \text{where} \qquad \mu = \alpha + 2\mathbf{p}^\intercal\beta.$$

The covariance matrix of the trait is

$$\mathrm{Cov}(\mathbf{y}) = \left(\sum_{i=1}^{m}\mathrm{Cov}(\mathbf{x}_i)\beta_i^2\right) + \mathrm{Cov}(\epsilon) = \mathbf{\Phi}\left(\sum_{i=1}^{m}4p_i(1-p_i)\beta_i^2\right) + (1-h^2)\sigma^2\mathbf{I}_n.$$

Therefore, we can write the covariance in terms of the heritability and the overall variance scale, in a formulation that matches previous work (Yang et al., 2010; Yang et al., 2011; Speed et al., 2012):

$$\mathrm{Cov}(\mathbf{y}) = \sigma^2\left(h^2 2\mathbf{\Phi} + (1-h^2)\mathbf{I}_n\right), \qquad \text{where} \qquad \sigma^2h^2 = \sum_{i=1}^{m}2p_i(1-p_i)\beta_i^2.$$

This last equation suggests a clear definition for the single-locus i contribution to the heritability, given by

$$h_i^2 = 2p_i(1-p_i)\beta_i^2\sigma^{-2}, \tag{2}$$

which is such that the total heritability is a sum of these single-locus heritabilities: $h^2 = \sum_{i=1}^{m}h_i^2$. Since the earlier expectations and covariances are conditioned on the choice of ancestral population, and given in terms of parameters that depend on it (p_i and $\mathbf{\Phi}$), then the parameters μ, σ^2, h^2 are all also dependent on the choice of ancestral population.

The parametrization of our model is equivalent to setting separate absolute scales to the genetic and environment variance components, as $\sigma_G^2 = h^2\sigma^2$ and $\sigma_E^2 = (1-h^2)\sigma^2$, respectively, which results in $\sigma_G^2 + \sigma_E^2 = \sigma^2$ and $\sigma_G^2/(\sigma_G^2+\sigma_E^2) = h^2$, as desired.

The factor of two in front of $\mathbf{\Phi}$ is traditionally there so that for an unstructured population $2\mathbf{\Phi} = \mathbf{I}_n$, in which case the trait covariance simplifies to $\mathrm{Cov}(\mathbf{y}) = \sigma^2\mathbf{I}_n$ for any value of h^2. More broadly, the variance of the trait for any outbred individual is σ^2 under this parametrization. In many previous presentations this factor of 2 does not appear there explicitly, but instead the kinship matrix is defined as $2\mathbf{\Phi}$ (Yang et al., 2011; Speed et al., 2012).

2.2 Multivariate Normal estimation model

Here we focus on heritability estimation from the Multivariate Normal (MVN) variance component model, namely (Speed et al., 2012; Yang et al., 2011)

$$\mathbf{y} \sim \mathrm{MVN}\left(\mu\mathbf{1}, \sigma^2\left(2h^2\mathbf{\Phi} + (1-h^2)\mathbf{I}_n\right)\right). \tag{3}$$

Note that, by construction, the above trait vector has the same mean and covariance matrix of the genetic trait model in Eq. (1). However, the trait in Eq. (3) is not drawn from genotypes anymore,

but solely from the kinship matrix; for this reason we refer to this as a *MVN* trait, in contrast to the *genetic* trait model in Eq. (1).

This MVN trait model (from Eq. (3)) is justified as arising from the genetic trait model (from Eq. (1)) in the limiting case of infinite causal locus effects, each with equal per-locus variance (heritability) according to Eq. (2) (equal $p_i(1-p_i)\beta_i^2$ for all i), the condition under which the genetic trait approaches the MVN distribution as a consequence of the multivariate central limit theorem. However, convergence may be slower in practice since both effect sizes (β_i) and ancestral allele frequencies (p_i) vary for realistic causal loci, so the product $p_i(1-p_i)\beta_i^2$ is not expected to be the same at all causal loci i. Another way of stating the difference is that Eq. (1) and Eq. (3) agree in the first two moments, but disagree in higher moments.

2.3 Kinship estimation

2.3.1 Standard kinship estimator

The "standard" kinship estimator is the most common estimator employed across various applications for population structure (Astle and Balding, 2009; Speed and Balding, 2015; Wang et al., 2017), including heritability estimation (Speed et al., 2012; Speed and Balding, 2015; Speed et al., 2017) and genetic association tests based on PCA (Price et al., 2006), LMMs (Astle and Balding, 2009; Zhou and Stephens, 2012; Loh et al., 2015; Sul et al., 2018) and other models (Rakovski and Stram, 2009; Thornton and McPeek, 2010). The popular heritability estimation approach GCTA (Yang et al., 2010; Yang et al., 2011) employs a variant of this estimator detailed in the next paragraph. The standard estimator, and its almost sure limit as the number of loci m and number of individuals n go to infinity (Ochoa and Storey, 2021), are given by

$$\hat{\varphi}_{jk}^{\text{std}} = \frac{1}{m}\sum_{i=1}^{m}\frac{(x_{ij}-2\hat{p}_i)\,(x_{ik}-2\hat{p}_i)}{4\hat{p}_i\,(1-\hat{p}_i)} \tag{4}$$

$$\xrightarrow[n,m\to\infty]{\text{a.s.}} \frac{\varphi_{jk}-\bar{\varphi}_j-\bar{\varphi}_k+\bar{\varphi}}{1-\bar{\varphi}}, \tag{5}$$

where $\hat{\varphi}_{jk}^{\text{std}}$ is the estimated kinship of individuals j and k, $\hat{p}_i = \frac{1}{2n}\sum_{j=1}^{n} x_{ij}$ is the standard ancestral allele frequency estimator, $\bar{\varphi}_j = \frac{1}{n}\sum_{k=1}^{n}\varphi_{jk}$ is the mean kinship of individual j with all others, and $\bar{\varphi} = \frac{1}{n^2}\sum_{j=1}^{n}\sum_{k=1}^{n}\varphi_{jk}$ is the overall mean kinship. This is a complex bias that varies for every pair of individuals, and which is on average a downward bias. (Note that the mean estimate, or $\frac{1}{n^2}\sum_{j=1}^{n}\sum_{k=1}^{n}\hat{\varphi}_{jk}^{\text{std}}$, is algebraically zero, regardless of the true value of the mean kinship.)

Variants of this approach that weigh loci according to linkage disequilibrium (Speed et al., 2017; Wang et al., 2017) do not alter the bias calculated above, since the same bias is present in each individual locus (Ochoa and Storey, 2021). Our previous work also considered a more general form where the ancestral allele frequency estimator $\hat{p}_i = \frac{1}{2}\sum_{j=1}^{n} w_j x_{ij}$ is calculated with weights w_j per individual j (such that $\sum_{j=1}^{n} w_j = 1$), and found that these weights alter the values of the bias terms $\bar{\varphi}_j$ and $\bar{\varphi}$ to be weighted averages, but no choice of weights eliminates these biases (Ochoa and Storey,

2021). Such weighted $\hat{p}_i$ estimates encompass the best unbiased linear estimator (Astle and Balding, 2009; Thornton and McPeek, 2010), with weights corresponding to $\mathbf{w} = (\mathbf{1}^\intercal \mathbf{\Phi}^{-1} \mathbf{1})^{-1} \mathbf{1}^\intercal \mathbf{\Phi}^{-1}$.

2.3.2 GCTA kinship estimator

The GCTA software (Yang et al., 2011) estimate what they refer to as a Genetic Relatedness Matrix (GRM), which is evidently twice a kinship matrix estimate due to the similarity to Eq. (4) and due to the historical connection between kinship and heritability in Eq. (3) and in previous work (Falconer and Mackay, 1996). In fact, the GCTA kinship estimates for two different individuals is identical to Eq. (4) (after taking into account the factor of 2 in Eq. (3)):

$$\hat{\varphi}_{jk}^{\text{GCTA}} = \hat{\varphi}_{jk}^{\text{std}} \qquad \text{for} \qquad j \neq k.$$

The GCTA kinship estimator differs from the standard estimator only for $j = k$ (Yang et al., 2011), where the estimator and its limit are instead given by (Ochoa and Storey, 2021):

$$\hat{\varphi}_{jj}^{\text{GCTA}} = \frac{1}{2} + \frac{1}{m} \sum_{i=1}^{m} \frac{x_{ij}^2 - (1 + 2\hat{p}_i)\, x_{ij} + 2\hat{p}_i^2}{4\hat{p}_i\,(1 - \hat{p}_i)} \tag{6}$$

$$\xrightarrow[n,m\to\infty]{\text{a.s.}} \frac{\varphi_{jj} - \bar{\varphi}_j}{1 - \bar{\varphi}}. \tag{7}$$

2.3.3 Popkin kinship estimator

The popkin (population kinship) estimator is given by (Ochoa and Storey, 2021)

$$\begin{aligned} A_{jk} &= \frac{1}{m} \sum_{i=1}^{m} (x_{ij} - 1)(x_{ik} - 1) - 1, \\ \hat{A}_{\min} &= \min_{u \neq v} \frac{1}{|S_u||S_v|} \sum_{j \in S_u} \sum_{k \in S_v} A_{jk}, \\ \hat{\varphi}_{jk}^{\text{new}} &= 1 - \frac{A_{jk}}{\hat{A}_{\min}}, \end{aligned} \tag{8}$$

where S_u are subpopulations that partition individuals. This estimator is accurate ($\hat{\varphi}_{jk}^{\text{new}} \xrightarrow[m\to\infty]{\text{a.s.}} \varphi_{jk}$) under the assumption that $\hat{A}_{\min}$ is calculated over individual pairs whose true kinship is zero. In other words, the two subpopulations S_u and S_v with the minimum mean A_{jk} value should have its true mean kinship value φ_{jk} be zero.

2.4 Heritability and genetic association software

TODO: state versions, download links, etc.

Outline:

- SOLAR-Eclipse (herit only) (Almasy and Blangero, 1998).

- GCTA (both heritability and genetic association) (Yang et al., 2011).
- PCA: used plink2 (genetic association only) (Chang et al., 2015).

2.5 Simulations

2.5.1 Trait simulation algorithm

Suppose the genotype matrix $\mathbf{X}$ is available, and we have fixed values for the number of causal loci m_1, the trait mean, variance scale, and heritability (μ, σ^2, h^2). The goal is to choose the intercept α and draw random effect sizes β that result in the desired trait parameters. First we randomly select m_1 loci to be causal, and subset the genotype matrix $\mathbf{X}$ and ancestral allele frequency vector $\mathbf{p}$ so that from this point on they contain only those causal loci (they now have dimensions $m_1 \times n$ and length m_1, respectively).

Below we divide the algorithm into two steps: (1) scaling the effect sizes, and (2) centering the trait. Each step forks into two cases: whether the true ancestral allele frequencies $\mathbf{p}$ are known or not (the latter requires a known kinship matrix $\mathbf{\Phi}$).

Scaling effect sizes. The initial effect sizes β_i are drawn independently from a standard normal distribution:

$$\beta_i \sim \mathrm{N}(0, 1).$$

First we consider the simpler case of known ancestral allele frequencies $\mathbf{p} = (p_i)$. The initial genetic variance scale is

$$\sigma_0^2 = \sum_{i=1}^{m_1} 2p_i(1 - p_i)\beta_i^2.$$

We obtain the desired variance by dividing each β_i by σ_0 (which results in a variance of 1) and then multiply by $h\sigma$ (which results in the desired variance of $h^2\sigma^2$). Combining both steps, the update is

$$\beta \leftarrow \beta \frac{h\sigma}{\sigma_0}.$$

Now we consider the case of unknown ancestral allele frequencies but known kinship matrix. First, sample estimates $\hat{\mathbf{p}} = (\hat{p}_i)$ of the ancestral allele frequencies are constructed from the genotype data as

$$\hat{p}_i = \frac{1}{2n} \mathbf{1}_n^\intercal \mathbf{x}_i.$$

Although this estimator is unbiased ($\mathrm{E}[\hat{\mathbf{p}}] = \mathbf{p}$), the resulting variance estimates of interest $\hat{p}_i\,(1 - \hat{p}_i)$ are downwardly biased (Ochoa and Storey, 2021):

$$\mathrm{E}\left[\hat{p}_i\,(1 - \hat{p}_i)\right] = p_i(1 - p_i)(1 - \bar{\varphi}),$$

where $\bar{\varphi} = \frac{1}{n^2}\mathbf{1}_n^\intercal \mathbf{\Phi} \mathbf{1}_n$ is the mean kinship coefficient in the data. Therefore the initial genetic variance scale, estimated as

$$\hat{\sigma}_0^2 = \sum_{i=1}^{m_1} 2\hat{p}_i(1-\hat{p}_i)\beta_i^2,$$

has an expectation of

$$\mathrm{E}\left[\hat{\sigma}_0^2\right] = \sigma_0^2(1-\bar{\varphi}).$$

Therefore, assuming that this additional factor $(1-\bar{\varphi})$ is known, the update

$$\beta \leftarrow \beta \frac{h\sigma\sqrt{1-\bar{\varphi}}}{\hat{\sigma}_0}$$

results in the desired variance.

Centering the trait. Here we consider the problem of selecting the intercept coefficient α that, together with the previous effect size coefficient vector β, result in the desired trait mean μ.

When ancestral allele frequencies are known, the trait can be centered precisely. Given our model, we obtain the desired overall trait mean μ by choosing the intercept coefficient to be

$$\alpha = \mu - 2\mathbf{p}^\intercal\beta.$$

When ancestral allele frequencies are unknown, the solution is to choose the intercept coefficient

$$\alpha = \mu - 2\hat{\bar{p}}\mathbf{1}_{m_1}^\intercal\beta, \qquad \hat{\bar{p}} = \frac{1}{m_1}\mathbf{1}_{m_1}^\intercal\hat{\mathbf{p}} = \frac{1}{2m_1 n}\mathbf{1}_{m_1}^\intercal\mathbf{X}^\intercal\mathbf{1}_n = \frac{1}{2}\bar{X},$$

where $\mathbf{1}_{m_1}$ is a length-m_1 column vector of ones. Note that this overal mean allele frequency $\hat{\bar{p}}$ is computed among causal loci only. This works very well in practice since β is drawn randomly, so it is uncorrelated to $\mathbf{p}$ and therefore

$$\frac{1}{m_1}\mathbf{p}^\intercal\beta = \frac{1}{m_1}\sum_{i=1}^{m_1} p_i\beta_i \approx \left(\frac{1}{m_1}\sum_{i=1}^{m_1} p_i\right)\left(\frac{1}{m_1}\sum_{i=1}^{m_1}\beta_i\right) = \frac{1}{m_1}\bar{p}\mathbf{1}_{m_1}^\intercal\beta$$

is a good approximation.

Now we discuss why the more obvious naive approach, which would be to center the trait using estimated ancestral allele frequencies as $\alpha = \mu - 2\hat{\mathbf{p}}^\intercal\beta$, does not work. This approach is equivalent to centering genotypes at each locus as

$$\mathbf{y} = \alpha\mathbf{1}_n + \sum_{i=1}^{m_1}(\mathbf{x}_i - 2\hat{p}_i\mathbf{1}_n)\beta_i + \epsilon.$$

However, this operation introduces a distortion in the covariance of the genotypes (Ochoa and Storey, 2021):

$$\mathrm{Cov}\left(\mathbf{x}_i - 2\hat{p}_i\mathbf{1}_n\right) = p_i(1-p_i)\left(\mathbf{\Phi} + \bar{\varphi}\mathbf{1}_n\mathbf{1}_n^\intercal - \varphi\mathbf{1}_n^\intercal - \mathbf{1}_n\varphi^\intercal\right),$$

where $\bar{\varphi}$ is the overall mean kinship, as before, and $\varphi = \frac{1}{n}\mathbf{\Phi}\mathbf{1}_n$ is a length-n column vector of per-row mean kinship values. These undesireable distortions propagate to the trait, which we confirmed in simulations (not shown). Note that the intercept version we chose instead does not induce this genotype centering, which prevents the undesireable distortions in the trait covariance.

2.5.2 Admixture simulation for genotype matrices

TODO: describe the BNPSD simulation.

3 Results

3.1 Empirical demonstration of biases in heritability estimation

Here we consider the effect of kinship estimator bias in heritability estimation. We use simulated genotypes and traits, and either estimate kinship matrices from these genotypes or we show the effect of asymptotic estimator bias based on previously-calculated limits. In particular, we consider the true kinship matrix of the simulation, the unbiased popkin estimator in Eq. (8), the standard kinship estimator in Eq. (4) and its limit in Eq. (5), and the GCTA estimator in Eq. (6) and its limit in Eq. (7). Lastly, we also consider two simulated trait types, namely the genetic trait in Eq. (1) and the MVN trait (or infinitesimal approximation) in Eq. (3). Each scenario was replicated 10 times, in each case producing a new genotype matrix as well as a new genetic and MVN traits.

First we considered a setting where there is only population structure (absence of family structure). We find that all estimates based on the genetic trait are upwardly biased, including that based on the true kinship matrix of the simulation (Fig. 1). In this case (genetic trait) the GCTA kinship estimates result in heritability estimates without an apparent bias, though its variance is very high. For the MVN trait case, the true kinship matrix results in unbiased estimation, as expected since here the model is fully specified correctly (the MVN trait assumption holds, and the true kinship matrix is used). The rest of the estimators and limits are biased for MVN traits too, although the direction varies for different cases: all kinship estimates underestimate heritability, while the GCTA limit overestimates it. Many of the overestimates across kinship variants and trait simulation types result in values near or at one.

Since traditional heritability estimation is based on family structure, we decided to consider a simulation including both population and family structure. Simply adding siblings results in striking changes in estimation biases (Fig. 2). Most versions of this analysis yield estimates with lower variance and bias when siblings are included, and the proportion of estimates close to 1 is sharply reduced. Another large differences is for the GCTA kinship limit, which is near one in the absence of siblings, and are instead downwardly biased when siblings are present. We still observe that all estimates on the genetic trait are upwardly biased. Moreover, although the GCTA estimated kinship appeared to ahve low bias in the absence of siblings, the inclusion of siblings reduces their variance and reveals an updard bias in these estimates too. For the MVN trait, now all estimates are downwardly biased, except for the true kinship matrix (which is expected to be unbiased), and possibly the limit of the standard kinship estimator.

In all these cases (Figs. 1 and 2) we find that kinship estimates do not result in the same estimates, on average, as their limits (expected as the number of loci goes to infinity). In all but

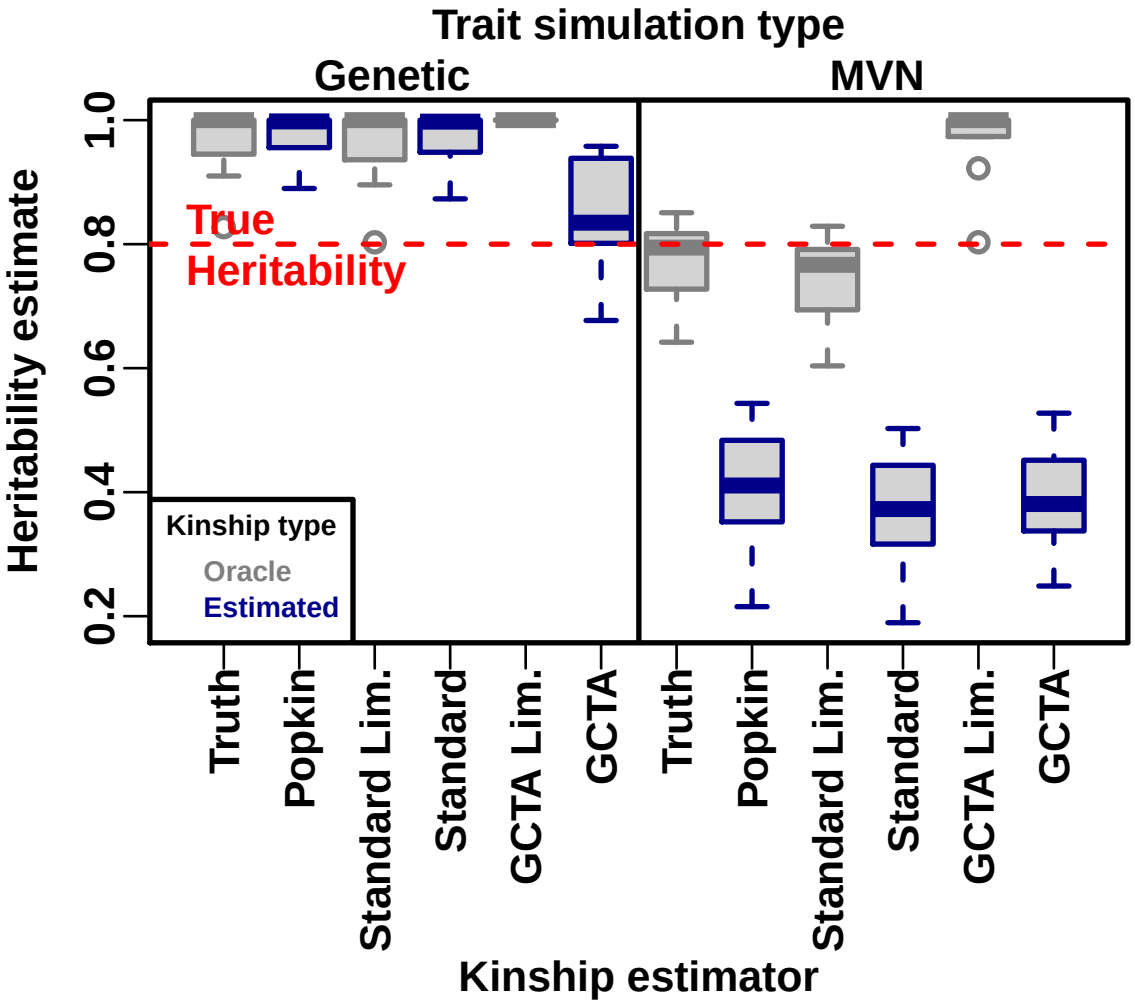

Figure 1: **Heritability estimation by GCTA with various kinship matrices and trait simulation types, admixture structure only.** Oracle kinship matrices rely on the true kinship matrix being known (unknown in practice), while estimated kinship matrices are obtained from observed genotypes and are what is used in practice. Genetic traits are simulated from genotypes, whereas MVN traits are simulated from the Multivariate Normal model, which is itself used to infer heritability but is unrealistic since it does not depend on any genotypes. Genotypes and traits were simulated for $n = 5000$ individuals, $m = 100000$ loci, for 10 replicates, from an admixture simulation ($K = 3$ ancestral subpopulations, $F_{\mathrm{ST}} = 0.3$ for admixed individuals, bias coefficient of $s = 0.5$), no family structure, and the traits had a true heritability of $h^2 = 0.8$ and $m_1 = 100$ causal loci).

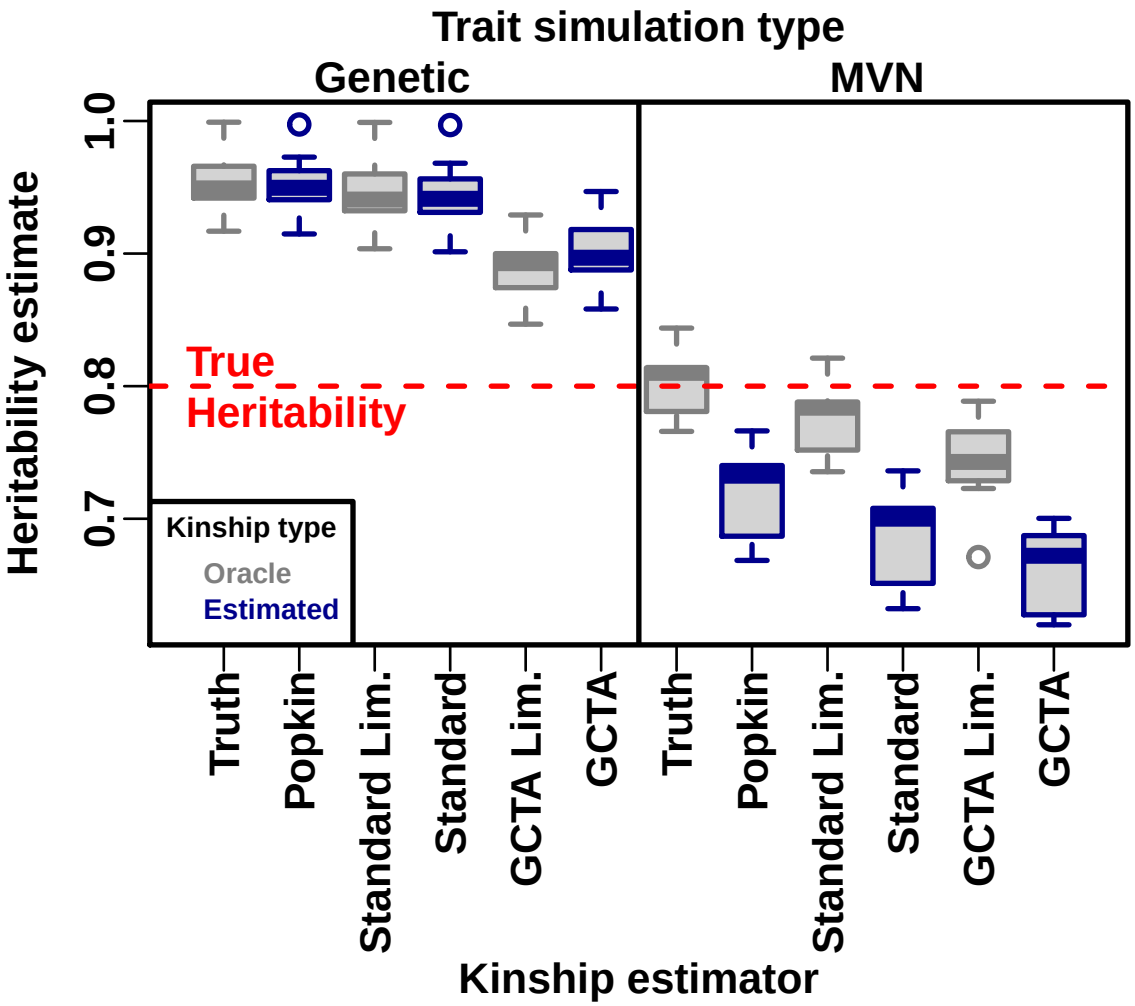

Figure 2: **Heritability estimation by GCTA with various kinship matrices and trait simulation types, admixture plus family structure (siblings).** Like Fig. 1, but proximal individuals (with the most similar ancestries) were mated to simulate siblings, and only the second generation (the siblings) were used to estimate heritability.

one of these cases, heritability estimates are downwardly biased when estimated kinship matrices are used compared to the limits of their respective kinship estimators. This suggests that kinship estimation variance results in an additional downward bias in heritability estimation, compared to the bias due to kinship estimator bias. We also find that estimation from proper, genetic traits, is biased under the MVN model on which GCTA's inference is based. Although MVN traits result in unbiased estimation under the true kinship matrix, this is an impractical scenario since the true kinship matrix is never known and true traits are never MVN.

Overall, since true traits are more like our genetic trait simulation, we expect there to be a heritability upward bias due to performing inference with the misspecified MVN model, and a separate downward bias due to the noise in kinship estimation. These biases must be further characterized in order to understand when they are problematic, and doing so may lead to new strategies for correction of these biases.

[INTERNAL NOTES:] However, these results are currently limited as only 10 replicates were simulated. These genotype simulations are expensive, but we should aim to speed them up to scale better as the number of individuals increases.

3.1.1 Next steps

1. The above suggests that heritability estimates from kinship estimates should approach those of their limits as the number of loci goes to infinity. We should come up with simulations where the number of loci is increased. However, the current code requires a lot of memory, so we may need to use the cluster and optimize the code. For simplicity, we could focus on a single estimator, such as "popkin", and can also add a step where kinship estimates are perfectly unbiased regardless of number of loci, to eliminate that confounding issue. This is because popkin is only unbiased as long as $\hat{A}_{\text{min}}$ in Eq. (8) is estimated perfectly, and I know for a fact that in our simulations there is a small bias in how $\hat{A}_{\text{min}}$, but that small bias may ultimately have an effect too and it'd be interesting to look at it separately. Anyway, the experiment could verify that kinship estimation noise does go to zero as the number of loci goes to infinity; that could help us determine the relationship between the number of loci and the bias, which could be used to correct the bias. Alternatively, should the experiment fail, we need to think about what other explanations there are for the observed biases (not sure what that would be yet).

2. Another prediction is that genetic traits should converge to MVN traits as the number of causal loci goes to infinity. More narrowly, heritability estimates from genetic traits (for simplicity, lets stick with the true kinship matrix) should approach the unbiased heritability estimates of MVN traits as the number of causal loci goes to infinity. A theoretical problem with this convergence is that right now the effect sizes, or per-locus heritabilities, vary randomly per causal locus, whereas convergence is only guaranteed when each per-locus heritability is equal.

So perhaps that is not happening here for that reason, and I can extend the simtrait code to simulate traits where effect sizes are exactly equal for every locus. This issue is related to the assumptions on distribution of effect sizes discussed by Speed et al. (2012). If the expected result is observed, then we can again figure out how the bias and the number of causal loci are related, and there may be other parameters that are important, such as the variance of effect sizes, though maybe it is not possible to fully characterize this bias. On the other hand, if the expected result is not observed, one potential explanation is that "simtrait" has bugs. I have performed several tests before that suggests that things are fine, but it remains a remote possibility.

3. Can you think of other interesting questions or tests we should make? Remember that our main simulation (based on the genetic trait, Eq. (1)) and the inference model (MVN, Eq. (3)) disagree, so the inference model is misspecified, and the use of estimated kinship matrices makes the model additionally misspecified, so there are several potential explanations for the biases. We should think more about what other explanations there are that I didn't list, or what other computational experiments can be informative towards understanding the problem better.

3.2 Empirical demonstration of robustness to kinship bias in PCA and LMM genetic assocation studies

To quantify the effect of the various kinship matrix estimates, and their limiting biases, we calculated effect size coefficients and p-values for all variants of the PCA and LMM methods, and calculated correlation coefficients of these vectors across methods. To further differentiate methods, we reduced the number of loci in the simulation to $m = 10,000$, which results in greater noise in all kinship matrix estimates (compared to a typical association study, which often contains millions of loci), which allows us to better distinguish that effect from the effect of bias (which remains in the limit of infinite loci).

We found that all the methods yield highly correlated values (for p-values in Fig. 3, effect sizes $\hat{\beta}$ in Fig. 4). However, while all PCA variants cluster together, LMM statistics do not cluster with each other. Instead, LMMs using the limiting kinship values cluster strongly with PCA, while LMMs that use estimated kinship matrices formed a separate clusted. This suggests that PCA is more robust to estimation noise than the LMM approach. Nevertheless, these effects are expected to be very small in real datasets, which typically have greater numbers of loci (recall we artificially reduced this number here to create greater differences between methods).

3.3 Theoretical justification of empirical observations

Here, to eliminate random estimation noise from the analysis (which our empirical evaluations suggest play a minor role), we shall focus on the limiting bias of the standard kinship estimator.

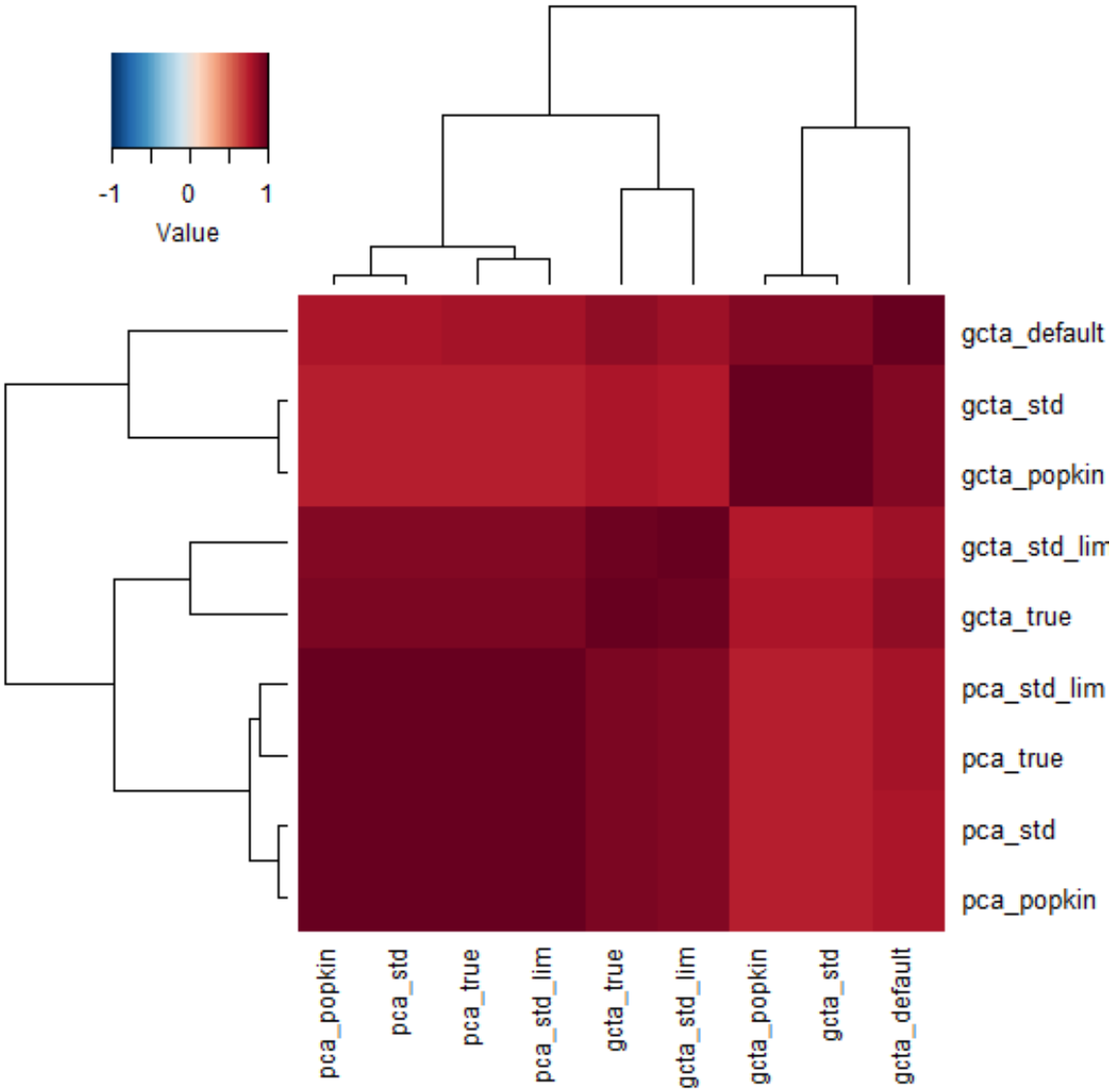

Figure 3: **Heatmap and dendrogram for p-values of GCTA and PCA with different kinship matrices ($m = 10,000$).**

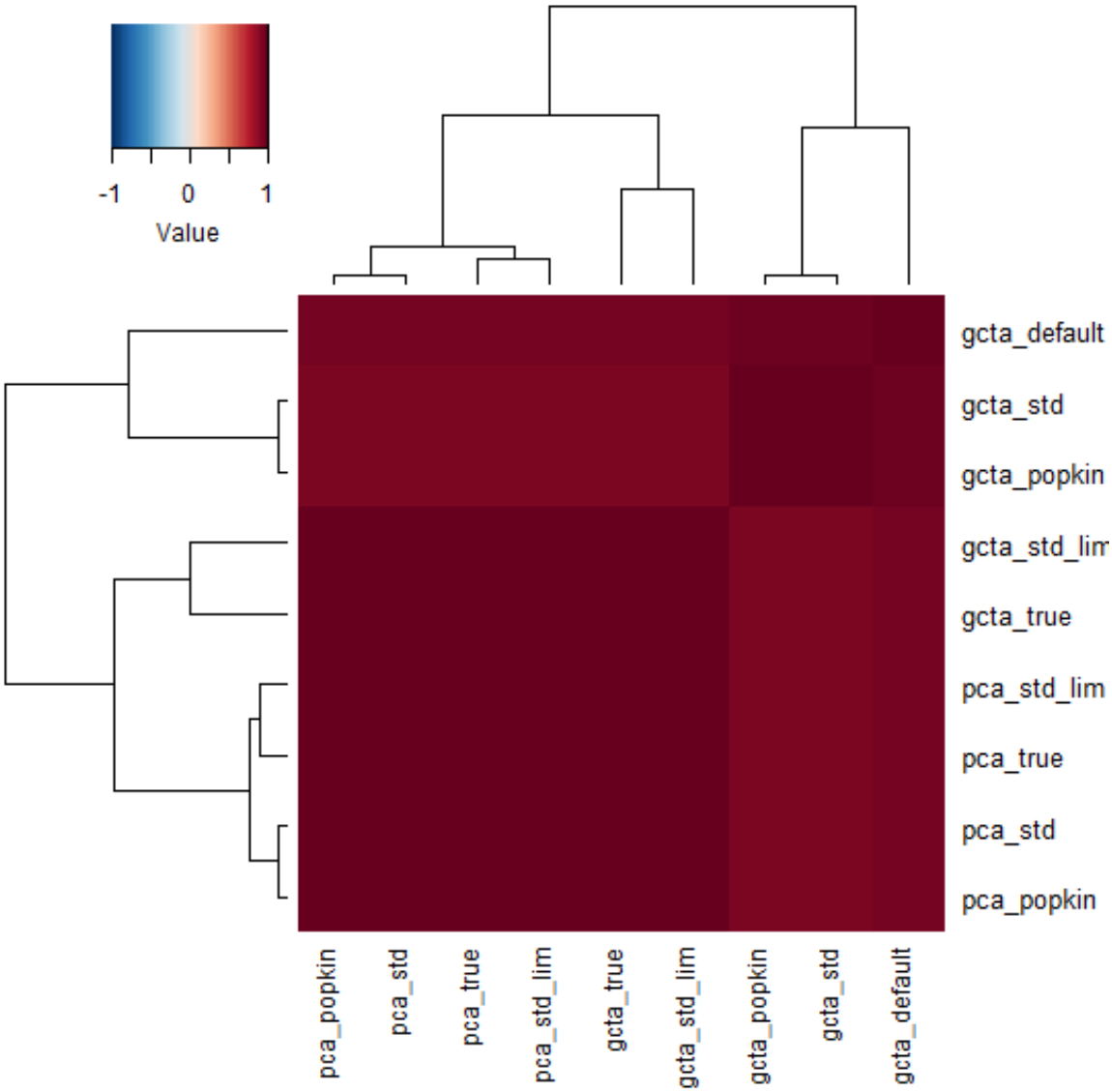

Figure 4: **Heatmap and dendrogram for $\hat{\beta}$ of GCTA and PCA with different kinship matrices ($m = 10,000$).**

Therefore, our theoretical results only consider the true kinship matrix $\mathbf{\Phi}$ and the limit of the standard kinship estimator from Eq. (4), which can be restated in terms of matrix operations as

$$\hat{\mathbf{\Phi}}^{\text{std-lim}} = \frac{1}{1-\bar{\varphi}} \left(\mathbf{\Phi} + \bar{\varphi} \mathbf{1}_n \mathbf{1}_n^\intercal - \varphi \mathbf{1}_n^\intercal - \mathbf{1}_n \varphi^\intercal \right).$$

The two kinship matrices are related more succinctly using the centering matrix $\mathbf{C}$:

$$\hat{\mathbf{\Phi}}^{\text{std-lim}} = \frac{1}{1-\bar{\varphi}} \mathbf{C} \mathbf{\Phi} \mathbf{C}, \qquad \mathbf{C} = \mathbf{I}_n - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^\intercal.$$

The centering matrix has been well studied, and we review its properties here. For any length-n vector $\mathbf{v}$ we have

$$\mathbf{C}\mathbf{v} = \mathbf{v} - \mathbf{1}_n \bar{v},$$

where $\bar{v} = \frac{1}{n} \mathbf{1}_n^\intercal \mathbf{v}$ is the mean value of the elements of $\mathbf{v}$. Therefore, $\mathbf{v} = \mathbf{1}_n$ gets transformed to the zero vector, so it is an eigenvector with an eigenvalue of zero:

$$\mathbf{C}\mathbf{1}_n = \mathbf{0}_n.$$

Moreover, any vector $\mathbf{v}$ orthogonal to $\mathbf{1}_n$ has a zero mean element ($\bar{v} = 0$) by hypothesis and it is not altered by $\mathbf{C}$ ($\mathbf{C}\mathbf{v} = \mathbf{v}$). Therefore, the nullspace of $\mathbf{C}$ is spanned by $\mathbf{1}_n$.

This centering matrix provides the key insight as to why PCA and LMM approaches are robust to this specific kinship bias, namely that by fitting the intercept term $\alpha \mathbf{1}_n$ together with the eigenvectors (for PCA) or random effects (in the same rowspace as the eigenvectors; for LMM) of $\hat{\mathbf{\Phi}}^{\text{std-lim}}$, they complete the rowspace of $\hat{\mathbf{\Phi}}^{\text{std-lim}}$ to equal the rowspace of the true kinship matrix $\mathbf{\Phi}$ plus the intercept. First we show the following lemma.

Lemma. *$\mathbf{1}_n$ is in the nullspace of $\hat{\mathbf{\Phi}}^{\text{std-lim}}$ but not of $\mathbf{\Phi}$.*

Proof. The vector $\mathbf{1}_n$ is not in the nullspace of any true kinship matrix $\mathbf{\Phi}$, since $\mathbf{\Phi}\mathbf{1}_n \neq \mathbf{0}_n$, which follows since all kinship values are non-negative and the diagonal of the kinship matrix is strictly positive (it has a minimum value of $\frac{1}{2}$). However, $\mathbf{1}_n$ is in the nullspace of $\hat{\mathbf{\Phi}}^{\text{std-lim}}$ since $\mathbf{C}\mathbf{1}_n = \mathbf{0}_n$:

$$\hat{\mathbf{\Phi}}^{\text{std-lim}} \mathbf{1}_n = \frac{1}{1-\bar{\varphi}} \mathbf{C} \mathbf{\Phi} \mathbf{C} \mathbf{1}_n = \mathbf{0}_n.$$

□

Now we may prove the desired theorem.

Theorem. *The rowspace of $\mathbf{\Phi}$ and $\mathbf{1}_n$ equals rowspace of $\hat{\mathbf{\Phi}}^{\text{std-lim}}$ and $\mathbf{1}_n$.*

Proof. Since $\mathbf{1}_n$ is in both rowspaces, it suffices to consider vectors $\mathbf{v}$ orthogonal to $\mathbf{1}_n$, which satisfy $\mathbf{C}\mathbf{v} = \mathbf{v}$. We shall prove below that any such vector is in the nullspace of $\mathbf{\Phi}$ if and only if it is in the

nullspace of $\hat{\mathbf{\Phi}}^{\text{std-lim}}$. Then, since the nullspace of $\mathbf{\Phi}$ and $\mathbf{1}_n$ is the same as the nullspace of $\hat{\mathbf{\Phi}}^{\text{std-lim}}$ and $\mathbf{1}_n$, it follows from the fundamental theorem of linear algebra that their rowspaces are also the same.

If $\mathbf{v}$ is in the nullspace of $\mathbf{\Phi}$, then $\mathbf{\Phi v} = \mathbf{0}_n$. It follows that

$$\hat{\mathbf{\Phi}}^{\text{std-lim}}\mathbf{v} = \frac{1}{1-\bar{\varphi}}\mathbf{C\Phi C v} = \frac{1}{1-\bar{\varphi}}\mathbf{C\Phi v} = \mathbf{0}_n,$$

so $\mathbf{v}$ is also in the nullspace of $\hat{\mathbf{\Phi}}^{\text{std-lim}}$.

Conversely, if $\mathbf{v}$ is in the nullspace of $\hat{\mathbf{\Phi}}^{\text{std-lim}}$, then $\hat{\mathbf{\Phi}}^{\text{std-lim}}\mathbf{v} = \mathbf{0}_n$, which implies that $\mathbf{C\Phi C v} = \mathbf{C\Phi v} = \mathbf{0}_n$. Left multiplying by $\mathbf{v}^\intercal$ results in $\mathbf{v}^\intercal\mathbf{C\Phi v} = \mathbf{v}^\intercal\mathbf{\Phi v} = 0$, which implies that $\mathbf{v}$ is also in the nullspace of the positive-semidefinite matrix $\mathbf{\Phi}$, as desired. If $\mathbf{\Phi}$ were positive definite, then no such $\mathbf{v} \neq \mathbf{0}_n$ would exist ($\mathbf{\Phi}$ would have the trivial nullspace $\{\mathbf{0}_n\}$). □

3.3.1 Theoretical justification for PCA genetic association

In PCA-based genetic association, the desired result follows from the previous theorem, as follows. Here the goal is to fit the trait $\mathbf{y}$ using a model simiar to our main model in Eq. (1), namely

$$\mathbf{y} = \mathbf{1}_n\alpha + \mathbf{x}_i\beta_i + \mathbf{U}_r\gamma_r + \epsilon, \tag{9}$$

where instead of including the whole genotype matrix $\mathbf{X}$ as we did in Eq. (1), here the genotype vector $\mathbf{x}_i$ at a single locus i is fit, and the effect of the rest of the genome is approxiated using the top r eigenvectors of the kinship matrix, contained in the $n \times r$ matrix $\mathbf{U}_r$ and its length-r vector of coefficients γ_r. At each locus i the coefficients α, β_i, and γ_r are fit to minimize the squared error between the observed trait and the model, and the residuals and possibly the degrees of freedom are used to evaluate the significance of the fit for the genotype under the null hypothesis that $\beta_i = 0$.

Here the kinship matrix in question is not the full kinship matrix $\mathbf{\Phi}$, but its r-dimensional approximation $\mathbf{\Phi}_r = \mathbf{U}_r\mathbf{\Lambda}_r\mathbf{U}_r^\intercal$, where $\mathbf{\Lambda}_r$ is an $r \times r$ diagonal matrix of the top r eigenvalues. One key requirement in need of verification is that $\mathbf{1}_n$ is not in the nullspace of $\mathbf{\Phi}_r$, which certainly holds for reasonable approximations as $\mathbf{1}_n^\intercal\mathbf{\Phi}_r\mathbf{1}_n$ estimates the mean kinship of the data, which is non-zero except for completely unstructured populations. The other key assumption,

$$\hat{\mathbf{\Phi}}_r^{\text{std-lim}}(1-\bar{\varphi}) = (\mathbf{C\Phi C})_r \approx \mathbf{C\Phi}_r\mathbf{C},$$

is only approximately true. In other words, centering the kinship matrix first and then approximating to the top r dimensions is not exactly equal to first approximating to r dimensions and then centering, although in simulations we found it to be a very good approximation (especially in the absence of family structure, which is where use of PCA is most appropriate anyway [TODO: cite Yao, other papers]). Therefore, in this case the rowspace of $\mathbf{\Phi}_r$ and $\mathbf{1}_n$ only approximately equals the rowspace of $\hat{\mathbf{\Phi}}_r^{\text{std-lim}}$ and $\mathbf{1}_n$.

The rowspace of $\mathbf{U}_r$ is the same as the rowspace of the r-dimensional kinship matrix it is derived from, so either $\mathbf{\Phi}_r$ or $\hat{\mathbf{\Phi}}_r^{\text{std-lim}}$. Thus, when fitting the model, the space spanned by $\mathbf{1}_n\alpha + \mathbf{U}_r\gamma_r$ is approximately the same whether the eigenvectors $\mathbf{U}_r$ are derived from the true kinship matrix $\mathbf{\Phi}$ or the limit of the biased estimator $\hat{\mathbf{\Phi}}^{\text{std-lim}}$, and this is so because the intercept vector $\mathbf{1}_n$ is present in the model. That implies that, while the coefficients α and γ_r may change when fit based on $\mathbf{\Phi}$ or $\hat{\mathbf{\Phi}}^{\text{std-lim}}$, the sum of the term $\mathbf{1}_n\alpha + \mathbf{U}_r\gamma_r$ will be approximately the same, as it is being chosen to minimize the least square error and the subspace is approximately the same, so the value of this solution must be approximately the same. Thus, the fit of the focal coefficient β_i is approximately the same either way, and so is its evaluation of significance (the sum of square errors is also approximately the same). Therefore, the genetic association test is approximately unchanged when the eigenvectors of either $\mathbf{\Phi}$ or $\hat{\mathbf{\Phi}}^{\text{std-lim}}$ are included as covariates.

3.3.2 Theoretical justification for LMM genetic association

...

4 Discussion

The biased kinship matrix may be more desireable in PCA, from a numerical standoint, as the resulting eigenvectors are not only orthogonal to each other but also to the intercept (whereas the eigenvectors of the true kinship matrix are not orthogonal to the intercept; see our Lemma).

...

References

Almasy, L. and J. Blangero (1998). "Multipoint quantitative-trait linkage analysis in general pedigrees". *Am. J. Hum. Genet.* 62(5), pp. 1198–1211.

Astle, William and David J. Balding (2009). "Population Structure and Cryptic Relatedness in Genetic Association Studies". *Statist. Sci.* 24(4). Mathematical Reviews number (MathSciNet): MR2779337, pp. 451–471.

Aulchenko, Yurii S., Dirk-Jan de Koning, and Chris Haley (2007). "Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis". *Genetics* 177(1), pp. 577–585.

Bulik-Sullivan, Brendan K. et al. (2015). "LD Score regression distinguishes confounding from polygenicity in genome-wide association studies". *Nat. Genet.* 47(3), pp. 291–295.

Chang, Christopher C. et al. (2015). "Second-generation PLINK: rising to the challenge of larger and richer datasets". *GigaScience* 4(1), p. 7.

Falconer, Douglas S. and Trudy F. C. Mackay (1996). *Introduction to Quantitative Genetics.* 4 edition. Harlow: Pearson.

Jiang, Jiming et al. (2016). "On high-dimensional misspecified mixed model analysis in genome-wide association study". *Ann. Statist.* 44(5), pp. 2127–2160.

Kang, Hyun Min et al. (2008). "Efficient control of population structure in model organism association mapping". *Genetics* 178(3), pp. 1709–1723.

Kang, Hyun Min et al. (2010). "Variance component model to account for sample structure in genome-wide association studies". *Nat. Genet.* 42(4), pp. 348–354.

Krishna Kumar, Siddharth et al. (2016). "Limitations of GCTA as a solution to the missing heritability problem". *Proc. Natl. Acad. Sci. U.S.A.* 113(1), E61–70.

Loh, Po-Ru et al. (2015). "Efficient Bayesian mixed-model analysis increases association power in large cohorts". *Nat. Genet.* 47(3), pp. 284–290.

Ochoa, Alejandro and John D. Storey (2016). "F_{ST} and kinship for arbitrary population structures I: Generalized definitions". Submitted, preprint at `http://biorxiv.org/content/early/2016/10/27/083915`.

— (2021). "Estimating FST and kinship for arbitrary population structures". *PLoS Genet* 17(1), e1009241.

Price, Alkes L. et al. (2006). "Principal components analysis corrects for stratification in genome-wide association studies". *Nat. Genet.* 38(8), pp. 904–909.

Rakovski, Cyril S. and Daniel O. Stram (2009). "A kinship-based modification of the armitage trend test to address hidden population structure and small differential genotyping errors". *PLoS ONE* 4(6), e5825.

Speed, Doug and David J. Balding (2015). "Relatedness in the post-genomic era: is it still useful?" *Nat. Rev. Genet.* 16(1), pp. 33–44.

— (2019). "SumHer better estimates the SNP heritability of complex traits from summary statistics". *Nat. Genet.* 51(2), pp. 277–284.

Speed, Doug et al. (2012). "Improved heritability estimation from genome-wide SNPs". *Am. J. Hum. Genet.* 91(6), pp. 1011–1021.

Speed, Doug et al. (2017). "Reevaluation of SNP heritability in complex human traits". *Nat Genet* 49(7), pp. 986–992.

Sul, Jae Hoon, Lana S. Martin, and Eleazar Eskin (2018). "Population structure in genetic studies: Confounding factors and mixed models". *PLoS Genet.* 14(12), e1007309.

Thornton, Timothy and Mary Sara McPeek (2010). "ROADTRIPS: case-control association testing with partially or completely unknown population and pedigree structure". *Am. J. Hum. Genet.* 86(2), pp. 172–184.

Visscher, Peter M., Jian Yang, and Michael E. Goddard (2010). "A commentary on 'common SNPs explain a large proportion of the heritability for human height' by Yang et al. (2010)". *Twin Res Hum Genet* 13(6), pp. 517–524.

Wang, Bowen, Serge Sverdlov, and Elizabeth Thompson (2017). "Efficient Estimation of Realized Kinship from SNP Genotypes". *Genetics*, genetics.116.197004.

Weir, Bruce S. and Jérôme Goudet (2017). "A Unified Characterization of Population Structure and Relatedness". *Genetics* 206(4), pp. 2085–2103.

Xie, C., D. D. Gessler, and S. Xu (1998). "Combining different line crosses for mapping quantitative trait loci using the identical by descent-based variance component method". *Genetics* 149(2), pp. 1139–1146.

Yang, Jian et al. (2010). "Common SNPs explain a large proportion of the heritability for human height". *Nat. Genet.* 42(7), pp. 565–569.

Yang, Jian et al. (2011). "GCTA: a tool for genome-wide complex trait analysis". *Am. J. Hum. Genet.* 88(1), pp. 76–82.

Yang, Jian et al. (2016). "GCTA-GREML accounts for linkage disequilibrium when estimating genetic variance from genome-wide SNPs". *Proc. Natl. Acad. Sci. U.S.A.* 113(32), E4579–4580.

Yu, Jianming et al. (2006). "A unified mixed-model method for association mapping that accounts for multiple levels of relatedness". *Nat. Genet.* 38(2), pp. 203–208.

Zhou, Xiang and Matthew Stephens (2012). "Genome-wide efficient mixed-model analysis for association studies". *Nat. Genet.* 44(7), pp. 821–824.