Fixed and mixed-effect genetic association models are robust to common population kinship estimation biases

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

Genetic association studies for structured populations model the correlation structure between individuals, most often using population kinship matrices. The two most common association models, Principal Components Analysis (PCA) association and the Linear Mixed-effects Model (LMM), are expressed in terms of kinship matrices. However, the most common kinship estimators can have severe biases that were only recently characterized. The goal of this work is to characterize the effect of these kinship biases on genetic association. We employed a large simulated admixed family and data from the 1000 Genomes Project, both with simulated traits, to evaluate the performance of each of PCA and LMM when a variety of kinship matrices is used. We tested several common kinship estimators, and in the simulated admixed family we also included the true kinship matrix and the the theoretical limits of the biased estimators (as the number of loci approaches infinity). Remarkably, we find negligible differences in association statistics and overall performance between all kinship matrices tested. Our empirical observations led us to the hypothesis that the association tests are invariant to whether the true kinship matrix or two of the biased limits are employed. We prove, using linear algebra, that our hypothesis holds exactly for LMM and approximately for PCA. Our constructive proof shows that both biased kinship matrices result in regression covariates (PCs in PCA, random effects in LMM) whose biases are compensated for by fitting the intercept, suggesting that only models with this precise arrangement will be robust to these kinship biases.

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; Yang et al., 2011; Zhou and Stephens, 2012; Loh et al., 2015; Sul et al., 2018). 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 (Weir and Goudet, 2017; Ochoa and Storey, 2021).

WG has a simpler, uniform bias (Weir and Goudet, 2017; Ochoa and Storey, 2021).

popkin is unbiased (Ochoa and Storey, 2021).

2 Results

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

To quantify the effect of the various kinship matrix estimators, and their limiting biases, we simulated genotypes and a trait, and calculated association p-values for all kinship variants of the PCA and LMM methods. We first simulated an admixed population with $K = 3$ ancestries, then simulated a 20-generation random pedigree from the admixed population as founders. This high-dimensinal admixed family scenario results in a large difference in performance between PCA and LMM approaches (Yao and Ochoa, 2019). To further differentiate methods, we increase noise in kinship estimation by purposefully simulating a small number of loci: $m = 10,000$.

The kinship estimates on this simulation, and the theoretical limits of these estimators, are shown in Fig. 1. The true kinship matrix (Fig. 1A) shows the family relatedness as high values concentrated near the diagonal and the ancestry-driven population structure as the broad patterns off-diagonal. This simulation illustrates the severity of the biases of these estimators, which were previously characterized theoretically (Ochoa and Storey, 2021). Only the Popkin estimator is unbiased (Fig. 1B), the rest presenting large negative biases which in turn result in abundant negative estimates. The Standard kinship estimator (the mean-of-ratios, or MOR, version; Fig. 1E), which is the most common estimator in these applications, and the closely related ratio-of-means (ROM) estimator (Fig. 1D), both have severe biases that are not only overall downwardly biased, but these biases also vary for every pair of individuals. The limit of the Standard ROM estimator (Fig. 1C), calculated from the true kinship matrix and its known functional form (see Methods), closely matches the previously mentioned estimates obtained from genotypes. The GCTA estimator equals the Standard MOR estimator except for the diagonal values, which also have a different bias (Fig. 1F-G). Lastly, the Weir-Goudet estimator has a uniform downward bias (Fig. 1H-I).

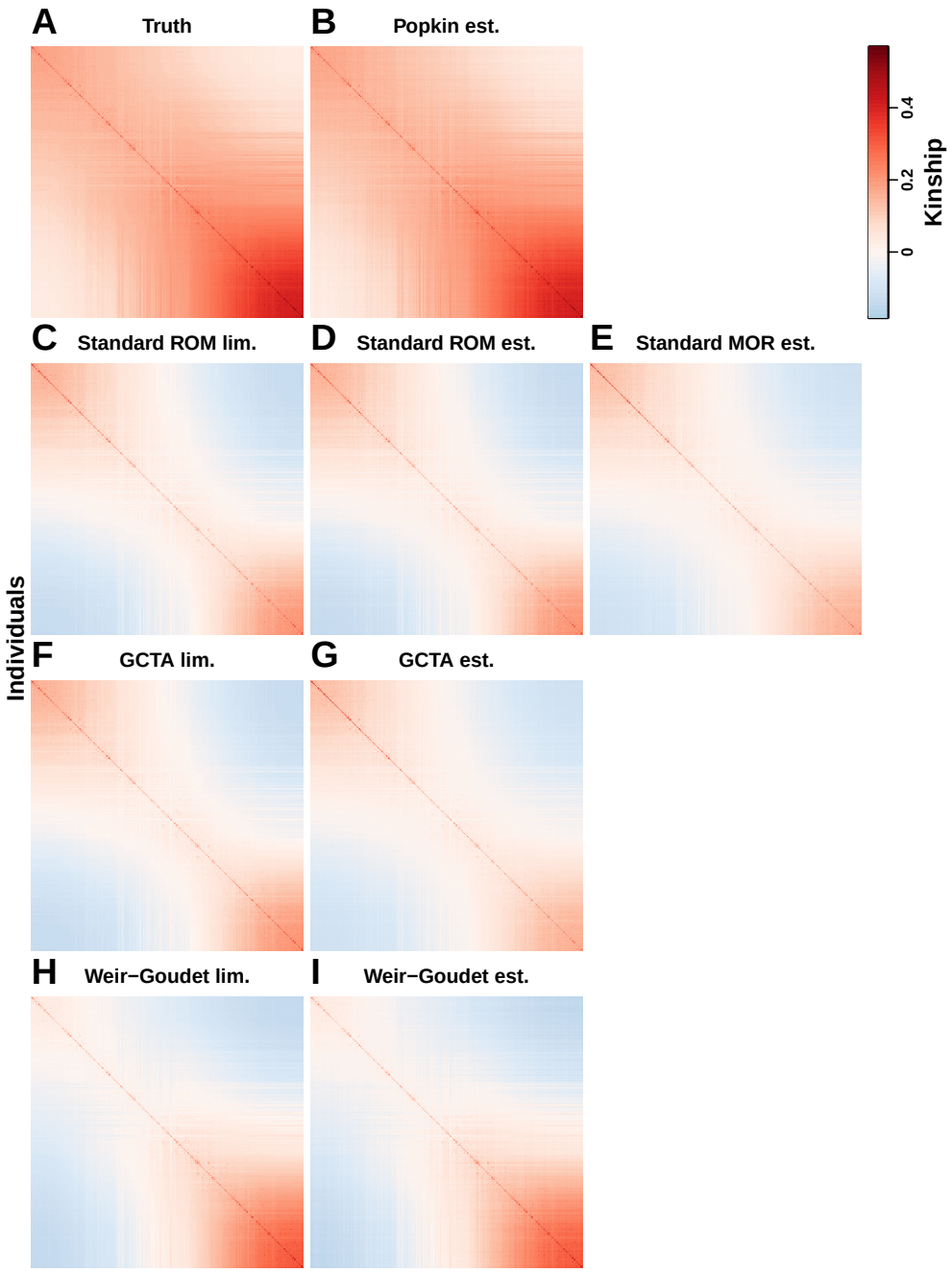

Figure 1: **Kinship estimates and their limits on the admixed family simulation.** Each panel represents a kinship matrix as a heatmap, with each of the $n = 1000$ individuals along both x and y axes, and the kinship value presented as color: positive estimates are in red, negative estimates in blue. The estimators considered are Popkin, Standard, GCTA, and Weir-Goudet. Each estimator has its limit in the first column (the limit of Popkin is Truth). Standard has two variants: ROM (ratio of means) and MOR (mean of ratios). Standard MOR does not have a closed-form limit, but in practice matches that of Standard ROM.

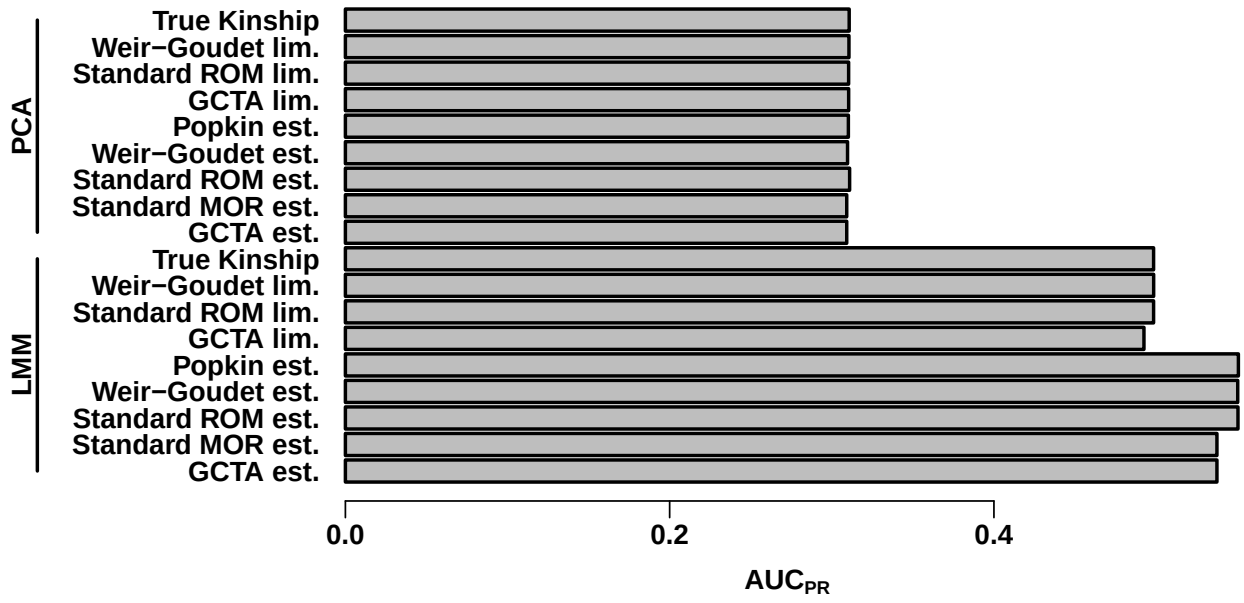

Figure 2: **Area Under the Precision-Recall Curves ($\mathbf{AUC_{PR}}$) for every combination of association model and kinship estimate.** Results based on a single replicate of the random genotype matrix and trait vector. Approaches appear to cluster primarily by association model (LMM vs PCA) and whether a kinship estimate was used or not, and do not depend much at all on the form of the bias.

We then performed both LMM and PCA association tests in order to determine if the kinship biases carry over to association biases. Surprisingly, we found that none of these kinship biases have discernible effects on association performance, as summarized by the Area Under the Precision-Recall Curve (AUC_{PR}; Fig. 2). The largest difference in performance is explained by the association model used (LMM vs PCA), as expected due to our use of a family simulation, where PCA is expected to perform less well than LMM. For PCA only, there are no clear differences between the performance of any of the kinship matrices. In contrast, the LMM results are relatively more sensitive to the use of noisy kinship estimates versus the (noiseless) limits of these estimates, a difference certainly increased by our deliberate use of a small number of simulated loci. Among the kinship estimator limits, use of the true kinship matrix, the limit of the Weir-Goudet estimator, or the limit of the Standard ROM estimator all result in practically the same performance, while the limit of the GCTA estimator shows a difference in performance. Similarly, among the estimates, the popkin, Weir-Goudet, and Standard ROM estimates result in the same performance, whereas the Standard MOR and GCTA estimates (which equals Standard MOR off-diagonal) have a slightly different performance.

To better understand the degree of agreement between the association results of the various kinship matrices, we next measured the agreement between methods at the level of the individual association p-values, summarized using pearson correlation coefficients (ρ; Fig. 3). We found that all the kinship matrices yield highly correlated p-values when applied to the same association model (LMM vs PCA), and otherwise mirroring our previous observations based on AUC_{PR}. The minimum correlation among PCA methods was 0.86, and among LMMs it was 0.84. Within PCA, the cluster that excludes the true kinship matrix and the popkin estimate has nearly identical p-values ($\rho > 0.99$), and inclusion of popkin lowers the minimum correlation to $\rho > 0.93$. Among LMMs, the cluster that includes the true kinship and the Weir-Goudet and Standard ROM limits also results in practically identical p-values ($\rho \approx 1$), while including the GCTA limit lowers that to $\rho > 0.97$. The LMM cluster that includes the Popkin, Weir-Goudet, and Standard ROM estimators also has $\rho \approx 1$, and separately the cluster with the Standard MOR and GCTA estimators has $\rho \approx 1$, while the LMM cluster that includes all estimates has $\rho > 0.96$. Thus, several sets of kinship matrices with different biases result in identical association statistics, within both LMMs and PCA models, while differences are largely driven by the association model used and whether the kinship matrices were estimates or not.

2.2 Theoretical justification of empirical observations under standard kinship bias

Our empirical observations suggested that the biases of the Standard kinship estimator do not alter association statistics whatsoever; here we provide proof that this is indeed the case. To eliminate random estimation noise from the analysis (which our empirical evaluations suggest play a minor

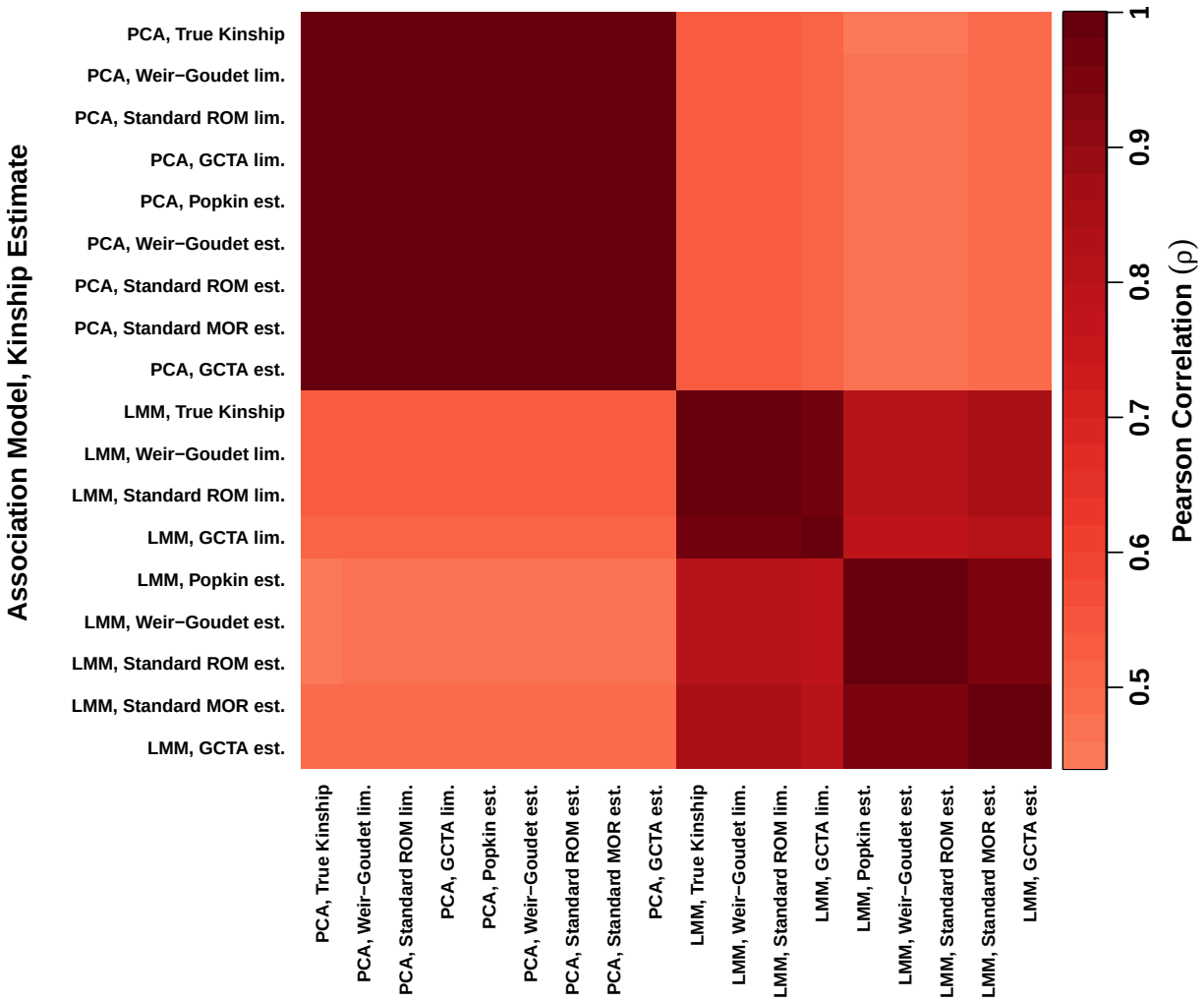

Figure 3: **Correlation between p-values of different association models and kinship estimates.** The association p-value vector (one value per tested locus) produced by each combination of association model (LMM vs PCA) and kinship matrix (x and y axes) was used to compute Pearson correlations (color). P-values between all kinship matrices using the same association model (PCA or LMM) are highly correlated (including biased and unbiased estimates), and often the maximum correlation of 1 is achieved.

role), we shall focus exclusively on the limiting bias of the Standard (ROM) kinship estimator. Our constructive proof identifies the exact conditions that allows the kinship bias to be compensated for, namely that population structure is modeled as a (fixed or random effect) covariate and an intercept term is fit jointly. More specifically, we find that the coefficients of the intercept and the random effect or PCs (for LMM and PCA, respectively) adapt to the bias, and no other coefficients change. This is fortunate, as the intercept and random effects or PCs coefficients are nuisance parameters that are most often unreported, while the focal genetic association coefficient and its p-value are completely unchanged by this precise bias.

Our theoretical results only consider the true kinship matrix $\mathbf{\Phi}$ and the limit of the standard kinship estimator (see **Methods**, Eq. (9)), which can be stated in matrix notation as

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

where $\mathbf{1}$ is a length-n column vector of ones, $\mathbf{J} = \mathbf{1}\mathbf{1}^\intercal$ is the $n \times n$ matrix full of ones, $\boldsymbol{\varphi} = \frac{1}{n}\mathbf{\Phi}\mathbf{1}$ is a length-n vector of per-row mean kinship values, and $\bar{\varphi} = \frac{1}{n^2}\mathbf{1}^\intercal\mathbf{\Phi}\mathbf{1}$ is the overall mean kinship (scalar). The two kinship matrices are related more succinctly using the $n \times n$ centering matrix,

$$\mathbf{C} = \mathbf{I} - \frac{1}{n}\mathbf{J},$$

where $\mathbf{I}$ is the $n \times n$ identity matrix. The limit of the standard kinship estimator is given in terms of a transformation of the true kinship matrix by

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

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

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

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

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

Moreover, any vector $\mathbf{v}$ orthogonal to $\mathbf{1}$ 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 $\mathbf{1}$.

This centering matrix provides the key insight as to why LMM and PCA approaches are robust to this specific kinship bias, namely that the bias in the random effects (for LMM) or eigenvectors (for PCA) of $\hat{\mathbf{\Phi}}^{\text{std-lim}}$ results in removing the mean values of these covariates only, so fitting the intercept term $\mathbf{1}\alpha$ compensates exactly for this bias. In the remaining sections we detail this argument, where we construct equivalent solutions under the true and biased kinship matrices.

2.2.1 Kinship matrix square root

Here we shall consider decompositions of positive semidefinite matrices of the form $\mathbf{\Sigma} = \mathbf{B}\mathbf{B}^\intercal$, which are guaranteed to exist. We denote such a $\mathbf{B}$ as a square root of $\mathbf{\Sigma}$, or in short $\mathbf{B} = \mathbf{\Sigma}^{\frac{1}{2}}$. These square roots of $\mathbf{\Sigma}$ are not unique, which is not a problem for our following argument; any such square root will work. (Note that there are alternate definitions of matrix square roots, such as $\mathbf{\Sigma} = \mathbf{B}\mathbf{B}$, but due to its connection to covariance, the definition $\mathbf{\Sigma} = \mathbf{B}\mathbf{B}^\intercal$ we adopted is most natural for this work and the notation $\mathbf{B} = \mathbf{\Sigma}^{\frac{1}{2}}$ simplifies our arguments.)

Given a square root of the true kinship matrix, $\mathbf{\Phi}^{\frac{1}{2}}$, we can construct a square root of the limit of the standard kinship estimator as

$$\left(\hat{\mathbf{\Phi}}^{\text{std-lim}}\right)^{\frac{1}{2}} = \frac{1}{\sqrt{1-\bar{\varphi}}}\mathbf{C}\mathbf{\Phi}^{\frac{1}{2}}. \tag{2}$$

This matrix square root can be verified to satisfy $\left(\hat{\mathbf{\Phi}}^{\text{std-lim}}\right)^{\frac{1}{2}}\left(\left(\hat{\mathbf{\Phi}}^{\text{std-lim}}\right)^{\frac{1}{2}}\right)^\intercal = \hat{\mathbf{\Phi}}^{\text{std-lim}}$ as given in Eq. (1).

2.2.2 The LMM genetic association model

In genetic association we are given data for n individuals, namely a length-n vector of trait values $\mathbf{y}$, which correspond to a quantitative trait, and a length-n vector $\mathbf{x}_i$ of genotypes at each locus i. These genotypes are encoded as dosages for a given risk allele that varies for each locus i, so it takes on the values of 0, 1, or 2 for diploid individuals. The goal is to determine if there is a significant association between the trait and the genotype vectors. Most genetic association models are formulated as, or are equivalent to, regression models.

The LMM regression model is given by

$$\mathbf{y} = \alpha\mathbf{1} + \beta\mathbf{x}_i + \mathbf{s} + \boldsymbol{\epsilon}, \tag{3}$$

where α is the intercept coefficient, β is the genetic effect coefficient, $\boldsymbol{\epsilon}$ are random independent residuals ($\boldsymbol{\epsilon} \sim \text{Normal}(\mathbf{0}, \sigma_\epsilon^2\mathbf{I})$ for some σ_ϵ^2 fit to the data), and the random effect satisfies (Sul et al., 2018)

$$\mathbf{s} \sim \text{Normal}\left(\mathbf{0}, \sigma^2\mathbf{\Phi}\right),$$

where σ^2 is also fit to the data. The dependence of the model on $\mathbf{\Phi}$ is clearer by writing it equivalently as

$$\mathbf{y} = \alpha\mathbf{1} + \beta\mathbf{x}_i + \sigma\mathbf{\Phi}^{\frac{1}{2}}\mathbf{r} + \boldsymbol{\epsilon}, \tag{4}$$

where $\mathbf{r} \sim \text{Normal}\left(\mathbf{0}, \mathbf{I}\right)$. The equivalence of Eq. (3) and Eq. (4) follows since $\mathbf{r}$ being Multivariate Normal implies that the affine transformation $\mathbf{s} = \sigma\mathbf{\Phi}^{\frac{1}{2}}\mathbf{r}$ is also Multivariate Normal, with matching

mean and covariance of the desired $\mathbf{s}$, namely

$$\begin{aligned}\mathrm{E}[\mathbf{s}] &= \sigma\boldsymbol{\Phi}^{\frac{1}{2}}\,\mathrm{E}[\mathbf{r}] = \mathbf{0},\\ \mathrm{Cov}(\mathbf{s}) &= \left(\sigma\boldsymbol{\Phi}^{\frac{1}{2}}\right)\mathrm{Cov}(\mathbf{r})\left(\sigma\boldsymbol{\Phi}^{\frac{1}{2}}\right)^{\mathsf{T}} = \sigma^2\boldsymbol{\Phi}.\end{aligned}$$

Note that the equivalence holds for every square root of $\boldsymbol{\Phi}$ (all $\mathbf{s} = \sigma\boldsymbol{\Phi}^{\frac{1}{2}}\mathbf{r}$ are equal in distribution), since the only requirement for equivalence, $\left(\boldsymbol{\Phi}^{\frac{1}{2}}\right)\left(\boldsymbol{\Phi}^{\frac{1}{2}}\right)^{\mathsf{T}} = \boldsymbol{\Phi}$, is satisfied by hypothesis.

2.2.3 Equivalent LMM fit under standard kinship bias

In the LMM of Eq. (4), $\mathbf{y}$, $\mathbf{x}_i$, and $\boldsymbol{\Phi}$ are given, while the coefficients α, β, σ, and the random effects $\mathbf{r}$ and $\boldsymbol{\epsilon}$ are fit to this data. This data is typically fit using maximum likelihood (ML) or restricted maximum likelihood (REML) (Kang et al., 2008); our argument covers any likelihood-based approach, since we will establish a parameter map between both models that results in equal likelihoods for every parameter set in the map.

We shall consider two model fits, one where $\boldsymbol{\Phi}$ is given, while in the other we provide $\boldsymbol{\Phi}' = \hat{\boldsymbol{\Phi}}^{\text{std-lim}}$ instead. The first fit will result in the unprimed variables below, while we distinguish the second fit using primed variables, namely:

$$\begin{aligned}\mathbf{y} &= \alpha\mathbf{1} + \beta\mathbf{x}_i + \sigma\boldsymbol{\Phi}^{\frac{1}{2}}\mathbf{r} + \boldsymbol{\epsilon}\\ &= \alpha'\mathbf{1} + \beta'\mathbf{x}_i + \sigma'\left(\boldsymbol{\Phi}'\right)^{\frac{1}{2}}\mathbf{r}' + \boldsymbol{\epsilon}'.\end{aligned}$$

Now we shall construct coefficients for the second model that result in the same fit to the data, including the same likelihood, as the first model. We achieve this by first setting $\beta' = \beta$, $\boldsymbol{\epsilon}' = \boldsymbol{\epsilon}$, and $\mathbf{r}' = \mathbf{r}$. Note that the previous equal random effects (including residuals) immediately results in the same likelihood for both models. The only parameters left to construct are α' and σ', which must satisfy

$$\alpha\mathbf{1} + \sigma\boldsymbol{\Phi}^{\frac{1}{2}}\mathbf{r} = \alpha'\mathbf{1} + \sigma'\left(\boldsymbol{\Phi}'\right)^{\frac{1}{2}}\mathbf{r}.$$

Next we substitute $\boldsymbol{\Phi}' = \hat{\boldsymbol{\Phi}}^{\text{std-lim}}$ using the matrix square root determined in terms of $\boldsymbol{\Phi}$ in Eq. (2), which results in

$$\alpha\mathbf{1} + \sigma\boldsymbol{\Phi}^{\frac{1}{2}}\mathbf{r} = \alpha'\mathbf{1} + \sigma'\frac{1}{\sqrt{1-\bar{\varphi}}}\mathbf{C}\boldsymbol{\Phi}^{\frac{1}{2}}\mathbf{r}.$$

The unknowns are solved for by left-multiplying, in turns, by $\mathbf{C}$ (which makes terms with $\mathbf{1}$ vanish) and by $\mathbf{1}^{\mathsf{T}}$ (which makes the term with $\mathbf{C}$ vanish). In the first case, left-multiplying by $\mathbf{C}$ results in

$$\sigma\mathbf{C}\boldsymbol{\Phi}^{\frac{1}{2}}\mathbf{r} = \sigma'\frac{1}{\sqrt{1-\bar{\varphi}}}\mathbf{C}\boldsymbol{\Phi}^{\frac{1}{2}}\mathbf{r},$$

so the only value of the scalar σ' that satisfies this equation is

$$\sigma' = \sigma\sqrt{1-\bar{\varphi}}.$$

In the second case, left-multiplying by $\mathbf{1}^\intercal$, while noting that $\mathbf{1}^\intercal\mathbf{1} = n$, and solving for α' results in

$$\alpha' = \alpha + \sigma\frac{1}{n}\mathbf{1}^\intercal\mathbf{\Phi}^{\frac{1}{2}}\mathbf{r}.$$

Note that the last equation can also be written as

$$\alpha' = \alpha + \eta, \qquad \eta \sim \text{Normal}(0, \sigma^2\bar{\varphi}),$$

since $\eta = \sigma\frac{1}{n}\mathbf{1}^\intercal\mathbf{\Phi}^{\frac{1}{2}}\mathbf{r}$ is a scalar random Normal variable (because it is a sum of marginals of a Multivariate Normal, each of which are Normal) with mean zero (because $\mathbf{r}$ had mean zero) and the stated variance (since $\text{Var}(\sigma\frac{1}{n}\mathbf{1}^\intercal\mathbf{\Phi}^{\frac{1}{2}}\mathbf{r}) = (\sigma\frac{1}{n}\mathbf{1}^\intercal\mathbf{\Phi}^{\frac{1}{2}})\,\text{Cov}(\mathbf{r})(\sigma\frac{1}{n}\mathbf{1}^\intercal\mathbf{\Phi}^{\frac{1}{2}})^\intercal$, $\text{Cov}(\mathbf{r}) = \mathbf{I}$, and $\frac{1}{n^2}\mathbf{1}^\intercal\mathbf{\Phi}\mathbf{1} = \bar{\varphi}$).

We just determined that every solution in terms of the true kinship matrix (including every combination of fixed coefficients and random effects) has a corresponding solution in terms of the limit of the standard kinship estimator, which has equal likelihood and equal fit to the data. This includes the optimal solution, whether determined by the ML or REML criteria. In both cases, the coefficient for the genetic effect is identical ($\beta' = \beta$ above), and because the fit to the data is also equal (in terms of likelihood and/or residuals), the association p-value is also equal (whether determined from the likelihood or from residuals). Thus, while two coefficients (the intercept α and the random effect variance scale σ^2) vary depending on whether the true or the limit of the biased standard kinship estimator are used, these are both nuisance parameters as far as the association test is concerned. The focal genetic effect coefficient and significance statistic are both invariant under this particular kinship bias compared to using the true kinship matrix.

2.2.4 The PCA genetic association model

The argument we just presented for LMM equivalence can be made with small changes for the PCA regression model, since these two models are so similar. Here we shall state the PCA model and elaborate on its strong connection to the LMM model, which has been presented before in similar forms (Astle and Balding, 2009; Hoffman, 2013).

The PCA regression model is

$$\mathbf{y} = \alpha\mathbf{1} + \beta\mathbf{x}_i + \mathbf{U}_d\boldsymbol{\gamma}_d + \boldsymbol{\epsilon}, \tag{5}$$

where d is the number of principal components, $\mathbf{U}_d$ is the $n \times d$ matrix of top eigenvectors (often refered to as "principal components" in genetics), and $\boldsymbol{\gamma}_d$ is a length-d vector of coefficients for each eigenvector. Note that the only difference from the LMM model (Eq. (4)) is the replacement of $\sigma\mathbf{\Phi}^{\frac{1}{2}}\mathbf{r}$ with $\mathbf{U}_d\boldsymbol{\gamma}_d$ here.

Before proceeding with our proof for invariance under the PCA model, to enhance our intuition of these two models, we present the relationship between eigendecomposition and matrix square

roots, which helps us connect the PCA model firmly to the LMM. Denote the eigendecomposition of the true kinship matrix as

$$\mathbf{\Phi} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\intercal,$$

where $\mathbf{U}$ is the complete $n \times n$ matrix of eigenvectors, and $\mathbf{\Lambda}$ is the $n \times n$ diagonal matrix of eigenvalues. Therefore, one square root of $\mathbf{\Phi}$ is given by

$$\mathbf{\Phi}^{\frac{1}{2}} = \mathbf{U}\mathbf{\Lambda}^{\frac{1}{2}},$$

where $\mathbf{\Lambda}^{\frac{1}{2}}$ contains the square roots of each eigenvalue along the diagonal. This equation reveals that the LMM model in Eq. (4) can be written in terms of the eigendecomposition, and thus resemble the PCA model even more closely, since

$$\sigma\mathbf{\Phi}^{\frac{1}{2}}\mathbf{r} = \sigma\mathbf{U}\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{r} = \mathbf{U}\boldsymbol{\gamma},$$

so that the length-n vector $\boldsymbol{\gamma}$ of coefficients for all the n eigenvectors is given by $\boldsymbol{\gamma} = \sigma\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{r}$. Thus, the PCA model attempts to fit coefficients only for the top d eigenvectors, whereas the LMM model effectively fits all of these coefficients by constraining them to a distribution.

2.2.5 Approximately equivalent PCA fit under standard kinship bias

We shall again consider two alternate model fits, here based on the PCA model of Eq. (5), one where the eigenvector matrix $\mathbf{U}_d$ corresponds to the true kinship matrix, and in the other $\mathbf{U}'_d$ corresponds to the biased limit of the standard kinship estimator. They key approximation is that

$$\mathbf{U}'_d \approx \mathbf{C}\mathbf{U}_d,$$

which is not strictly equal (since $\mathbf{CU}$ is not orthogonal, as eigenvectors must be), but we have found it to be a good approximation in practice.

The two model fits we are considering are

$$\begin{aligned}\mathbf{y} &= \alpha\mathbf{1} + \beta\mathbf{x}_i + \mathbf{U}_d\boldsymbol{\gamma}_d + \boldsymbol{\epsilon} \\ &= \alpha'\mathbf{1} + \beta'\mathbf{x}_i + \mathbf{U}'_d\boldsymbol{\gamma}'_d + \boldsymbol{\epsilon}',\end{aligned}$$

and we again assume that the focal parameter $\beta' = \beta$ and the residuals $\boldsymbol{\epsilon}' = \boldsymbol{\epsilon}$ are equal. Eliminating the resulting shared terms, and replacing $\mathbf{U}'_d = \mathbf{C}\mathbf{U}_d$ (assuming that our approximation holds exactly) results in the remaining coefficients having to satisfy

$$\alpha\mathbf{1} + \mathbf{U}_d\boldsymbol{\gamma}_d = \alpha'\mathbf{1} + \mathbf{C}\mathbf{U}_d\boldsymbol{\gamma}'_d.$$

We solve for the missing coefficients by left-multiplying by $\mathbf{C}$ and $\mathbf{1}^\intercal$ as before, which here ultimately results in

$$\boldsymbol{\gamma}'_d = \boldsymbol{\gamma}_d, \qquad \alpha' = \alpha + \frac{1}{n}\mathbf{1}^\intercal\mathbf{U}_d\boldsymbol{\gamma}_d.$$

Thus, as for LMM, here the nuisance intercept coefficient compensates for the bias in the eigenvectors.

The PCA regression is an ordinary multiple linear regression, which is fit by minimizing the sum of square residuals. Since the residuals were equal in both models, then the optimal solution in one model maps to the optimal solution in the other model as well. The p-value of the genetic effect is usually calculated using a chi-squared test or an F-test, both of which depend only on the residuals and the degrees of freedom, all of which are invariant under the solution parameter map we constructed. Therefore, both the focal genetic effect coefficient β and its p-value are invariant under this particular kinship bias compared to using the true kinship matrix. However, the result for PCA relies on an approximation, whereas for LMM it was exact.

2.3 Proof that association based on Weir-Goudet and true kinship are equivalent

The limiting bias of the Weir-Goudet (WG) estimator (see Methods, Eq. (15)) is given in terms of the true kinship matrix by

$$\hat{\mathbf{\Phi}}^{\text{WG-lim}} = \frac{1}{1-\tilde{\varphi}}(\mathbf{\Phi} - \tilde{\varphi}\mathbf{J}). \tag{6}$$

Our strategy for proving the association equivalence under both kinship matrices is to construct a random effect structured as the true kinship matrix from a random effect structured as WG. Although we were unable to calculate a matrix square root for the WG limit in terms of the true kinship matrix, we found that this alternate derivation for LMMs proves the desired result and leads to strikingly analogous equations for the variance scale σ^2 and intercept α between both models, where the only difference ultimately is the replacement of the mean kinship $\bar{\varphi}$ in the standard estimator results with $\tilde{\varphi}$ for WG. We were unable to prove the PCA case with this strategy.

2.3.1 Constructing random effects with true kinship structure from Weir-Goudet structure

Solving for the true kinship matrix in Eq. (6), we obtain

$$\mathbf{\Phi} = (1-\tilde{\varphi})\hat{\mathbf{\Phi}}^{\text{WG-lim}} + \tilde{\varphi}\mathbf{J}. \tag{7}$$

Let the random effect from the LMM using the WG limit be

$$\mathbf{s}^{\text{WG}} \sim \text{Normal}\left(\mathbf{0}, \sigma^2_{\text{WG}}\hat{\mathbf{\Phi}}^{\text{WG-lim}}\right),$$

where σ^2_{WG} is a parameter fit to the data. The Normal distribution above requires that $\hat{\mathbf{\Phi}}^{\text{WG-lim}}$ be positive definite, a fact that is proven for arbitrary true kinship matrices (except for one degenerate case) in Appendix A. We desire to construct a random effect such that $\mathbf{s} \sim \text{Normal}\left(\mathbf{0}, \sigma^2\mathbf{\Phi}\right)$, where again σ^2 has been fit to the data. Eq. (7) suggest a relationship of the form

$$\mathbf{s} = \mathbf{s}^{\text{WG}} + \eta\mathbf{1},$$

where η is a random scalar drawn independently from a Normal distribution,

$$\eta \sim \text{Normal}(0, \sigma_\eta^2),$$

where σ_η^2 is yet to be determined. This constructed $\mathbf{s}$ is indeed Multivariate Normal since $\eta\mathbf{1}$ is a (degenerate) Multivariate Normal and the sum of two Multivariate Normal variables is itself a Multivariate Normal variable. All that is left is to match the mean of covariance of the desired $\mathbf{s}$, namely

$$\begin{aligned}
\text{E}[\mathbf{s}] &= \text{E}\left[\mathbf{s}^{\text{WG}}\right] + \text{E}[\eta]\mathbf{1} = \mathbf{0}, \\
\text{Cov}(\mathbf{s}) &= \text{Cov}\left(\mathbf{s}^{\text{WG}}\right) + \text{Var}(\eta)\mathbf{J} \\
&= \sigma^2_{\text{WG}}\hat{\mathbf{\Phi}}^{\text{WG-lim}} + \sigma_\eta^2\mathbf{J} = \sigma^2\mathbf{\Phi},
\end{aligned}$$

which is achieved with $\sigma^2 = \sigma^2_{\text{WG}}/(1-\tilde{\varphi})$ and $\sigma_\eta^2 = \tilde{\varphi}\sigma^2$.

2.3.2 Equivalent LMM fit under Weir-Goudet kinship bias

Similarly to the case for the standard kinship estimator, we consider two equivalent LMM fits, one with $\mathbf{s}$ from the true kinship matrix, the other with $\mathbf{s}' = \mathbf{s}^{\text{WG}}$ from the WG limit:

$$\mathbf{y} = \alpha\mathbf{1} + \beta\mathbf{x}_i + \mathbf{s} + \boldsymbol{\epsilon} = \alpha'\mathbf{1} + \beta'\mathbf{x}_i + \mathbf{s}' + \boldsymbol{\epsilon}'.$$

As before, our empirical results motivate that $\beta = \beta', \boldsymbol{\epsilon} = \boldsymbol{\epsilon}'$, and only α' is left, which must satisfy

$$\alpha\mathbf{1} + \mathbf{s} = \alpha'\mathbf{1} + \mathbf{s}'.$$

We now substitute our constructed $\mathbf{s} = \mathbf{s}' + \eta\mathbf{1}$, where η is as determined in the previous subsection, which after simplifying yields

$$\alpha\mathbf{1} + \eta\mathbf{1} = \alpha'\mathbf{1}.$$

Note $\mathbf{s}'$ drops out of the equation, which follows from having previously selected its variance appropriately to match that of the model for the true kinship matrix, namely

$$\sigma' = \sigma\sqrt{1-\tilde{\varphi}},$$

which agrees with the solution we found earlier for the standard kinship estimator, expect replacing $\bar{\varphi}$ with $\tilde{\varphi}$. Lastly, the relation between intercepts is

$$\alpha' = \alpha + \eta,$$

which again resembles the solution for the standard model, where both draw a random variable with mean zero, except the earlier variance of $\sigma^2\bar{\varphi}$ is replaced with $\sigma^2\tilde{\varphi}$ here. As before, these results imply that the association statistics are invariant to the choice between the true and WG limit kinship matrices.

3 Discussion

The biased kinship matrix may be more desireable in PCA, from the standoint of numerical stability, 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).

The GCTA kinship estimator was not analyzed theoretically, but it is very similar to the standard kinship estimator, so we expect an approximately similar performance between those two and using the true kinship matrix.

4 Methods

4.1 Genetic model

The following genetic model justifies the use of kinship matrices in association studies, and is the basis of all kinship estimation bias calculations that our theoretical work depends upon.

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

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

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, $\mathbf{1} = (1)$ 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, nor do we explicitly condition on the ancestral population (it is done implicitly) when calculating expectations and covariances as done in previous work.

4.2 Kinship estimation

4.2.1 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). GCTA (Yang et al., 2010; Yang et al., 2011) employs a variant of this estimator detailed in the next subsection.

There are two versions of this standard kinship estimator, namely the mean-of-ratios (MOR) and ratio-of-means (ROM) version (Ochoa and Storey, 2021). Most approaches implement the MOR version. However, the ROM version has more favorable convergence properties relevant to our overall theoretical argument.

The ROM version of the standard kinship estimator, and its almost sure limit as the number of loci m go to infinity (Ochoa and Storey, 2021), are given by

$$\hat{\varphi}_{jk}^{\text{std-rom}} = \frac{\sum_{i=1}^{m} (x_{ij} - 2\hat{p}_i)(x_{ik} - 2\hat{p}_i)}{\sum_{i=1}^{m} 4\hat{p}_i(1-\hat{p}_i)} \tag{8}$$

$$\xrightarrow[m\to\infty]{\text{a.s.}} \frac{\varphi_{jk} - \bar{\varphi}_j - \bar{\varphi}_k + \bar{\varphi}}{1 - \bar{\varphi}}, \tag{9}$$

where $\hat{\varphi}_{jk}^{\text{std-rom}}$ 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 per row and column, or $\frac{1}{n}\sum_{k=1}^{n} \hat{\varphi}_{jk}^{\text{std-rom}}$ for every j, is algebraically zero, regardless of the true value of the row mean kinship; the same is true for the MOR version below.)

The MOR version of the standard estimator, which again is the most common form of the estimator, is given by

$$\hat{\varphi}_{jk}^{\text{std-mor}} = \frac{1}{m}\sum_{i=1}^{m} \frac{(x_{ij} - 2\hat{p}_i)(x_{ik} - 2\hat{p}_i)}{4\hat{p}_i(1-\hat{p}_i)}. \tag{10}$$

This estimator does not have closed-form limit, but it is well approximated by Eq. (9) in practice, especially when loci with small minor allele frequencies are excluded prior to calculating this estimate.

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 in Eq. (9), 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 $\mathbf{w} = (\mathbf{1}^\intercal \mathbf{\Phi}^{-1}\mathbf{1})^{-1}\mathbf{1}^\intercal\mathbf{\Phi}^{-1}$.

4.2.2 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 the Standard MOR kinship estimator Eq. (10). In fact, the GCTA kinship estimates for two different individuals is identical to Eq. (10) (after taking into account the factor of 2):

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

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):

$$\hat{\varphi}_{jj}^{\text{GCTA}} = \frac{1}{2} + \frac{1}{m}\sum_{i=1}^{m} \frac{x_{ij}^2 - (1+2\hat{p}_i)\, x_{ij} + 2\hat{p}_i^2}{4\hat{p}_i\,(1-\hat{p}_i)} \tag{11}$$

$$\xrightarrow[n,m\to\infty]{\text{a.s.}} \frac{\varphi_{jj} - \bar{\varphi}_j}{1-\bar{\varphi}}. \tag{12}$$

4.2.3 Popkin kinship estimator

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

$$\begin{aligned} A_{jk} &= \frac{1}{m}\sum_{i=1}^{m}(x_{ij}-1)(x_{ik}-1) - 1, \\ \hat{\varphi}_{jk}^{\text{new}} &= 1 - \frac{A_{jk}}{\hat{A}_{\min}}, \end{aligned} \tag{13}$$

where $\hat{A}_{\min}$ is an estimate of the minimum of the limiting values of A_{jk}. Here we employ $\hat{A}_{\min} = \min_{j\neq k} A_{jk}$. This kinship estimator is accurate, namely by satisfying

$$\hat{\varphi}_{jk}^{\text{new}} \xrightarrow[m\to\infty]{\text{a.s.}} \varphi_{jk},$$

under the assumption that $\hat{A}_{\min}$ is calculated over individual pairs whose true kinship is zero.

4.2.4 Weir-Goudet kinship estimator

The Weir-Goudet (WG) kinship estimator and its limit are given by (Weir and Goudet, 2017; Ochoa and Storey, 2021)

$$\hat{\varphi}_{jk}^{\text{WG}} = 1 - \frac{A_{jk}}{\hat{A}_{\text{avg}}} \tag{14}$$

$$\xrightarrow[m\to\infty]{\text{a.s.}} \frac{\varphi_{jk} - \tilde{\varphi}}{1-\tilde{\varphi}}, \tag{15}$$

where A_{jk} is as in Eq. (13), and

$$\begin{aligned}\hat{A}_{\text{avg}} &= \frac{2}{n(n-1)} \sum_{j=2}^{n} \sum_{k=1}^{j-1} A_{jk}, \\ \tilde{\varphi} &= \frac{2}{n(n-1)} \sum_{j=2}^{n} \sum_{k=1}^{j-1} \varphi_{jk}.\end{aligned} \tag{16}$$

Thus, the WG estimator resembles the popkin estimator in Eq. (13), except it replaces $\hat{A}_{\text{min}}$ with $\hat{A}_{\text{avg}}$ and that results in a uniform downward bias given by $\tilde{\varphi}$, which is the mean kinship between all different individual pairs (it excludes the diagonal, or self-kinship values, compared to the $\bar{\varphi}$ that appears in the standard and GCTA estimators). In Appendix B we prove that

$$0 \le \tilde{\varphi} \le \bar{\varphi} \le \bar{d} \le 1,$$

where $\bar{d} = \frac{1}{n} \sum_{j=1}^{n} \varphi_{jj}$, and equalities are achieved if and only if all kinship values are equal.

4.3 Simulations

[TODO: heritability doesn't matter here, so we can probably shorten considerably. Trait simulation is as in Yiqi's paper, can defer to that.]

4.3.1 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 $\boldsymbol{\beta}$ 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 $\boldsymbol{\Phi}$).

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

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

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

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

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

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

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 $\bar{\varphi} = \frac{1}{n^2}\mathbf{1}^\intercal\boldsymbol{\Phi}\mathbf{1}$ 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

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

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 $\boldsymbol{\beta}$, 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

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

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

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

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 $\boldsymbol{\beta}$ is drawn randomly, so it is uncorrelated to $\mathbf{p}$ and therefore

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

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 $\alpha = \mu - 2\hat{\mathbf{p}}^\intercal\boldsymbol{\beta}$, does not work. This approach is equivalent to centering genotypes at each locus as

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

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

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

where $\bar{\varphi}$ is the overall mean kinship, as before, and $\boldsymbol{\varphi} = \frac{1}{n}\boldsymbol{\Phi}\mathbf{1}$ 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.

4.3.2 Admixture simulation for genotype matrices

TODO: describe the BNPSD simulation.

Also describe family structure.

Describe 1000 Genomes.

4.4 Software

TODO: state versions, download links, etc.

GCTA (Yang et al., 2011). Used to calculate GCTA kinship estimates. We pass $2\boldsymbol{\Phi}$ for all kinship matrices tested.

PCA: implemented regression in R (function lm). p-values based on F-test. Used $r = k - 1 = 2$ for admixed family simulations, and $r = 10$ for 1000 Genomes project.

popkin is estimated with the popkin R package.

All other estimators and limits are calculated using the popkinsuppl R package.

References

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.

Hefferon, Jim (2020). "Linear Algebra Fourth Edition", pp. 144–152.

Hoffman, Gabriel E. (2013). "Correcting for population structure and kinship using the linear mixed model: theory and extensions". *PLoS ONE* 8(10), e75707.

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.

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". *bioRxiv* (10.1101/083915). `https://doi.org/10.1101/083915`.

— (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.

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.

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.

Yao, Yiqi and Alejandro Ochoa (2019). "Testing the effectiveness of principal components in adjusting for relatedness in genetic association studies". *bioRxiv*. Publisher: Cold Spring Harbor Laboratory Section: New Results, p. 858399.

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.

Appendices

A Proof that WG limits are always positive definite.

Starting from the fact that a true kinship matrix is positive definite by definition (it is a covariance matrix), we shall prove that the limit of the Weir-Goudet estimator is also positive definite, with the exception of one degenerate case. Recall from Eq. (6) that these two matrices are related by $\hat{\mathbf{\Phi}}^{\text{WG-lim}} = \frac{1}{1-\tilde{\varphi}}(\mathbf{\Phi} - \tilde{\varphi}\boldsymbol{J})$. We shall not consider $\mathbf{\Phi} = \mathbf{J}$ as a valid kinship matrix, which therefore ensures that $\tilde{\varphi} < 1$ as there is at least one kinship value with $\varphi_{jk} < 1$. Below we will consider two linear subspaces of $\mathbb{R}^n$, S_1 spanned by $\mathbf{1}$ and S_2 its complement (orthogonal to $\mathbf{1}$), and prove that $\hat{\mathbf{\Phi}}^{\text{WG-lim}}$ is positive definite in both subspaces, therefore it is positive-definite in the direct sum of the subspaces, which equals the entire space: $\mathbb{R}^n = S_1 \oplus S_2$. (This follows since vectors $\mathbf{v}$ for which $\hat{\mathbf{\Phi}}^{\text{WG-lim}}$ is not positive definite, if they exist, span a linear subspace, but its intersection to S_1, S_2, and therefore $S_1 \oplus S_2$, is trivial (Hefferon, 2020).) In both subspaces we will prove that $\mathbf{v} \in S_i$ and $\mathbf{v} \neq \mathbf{0}$ implies $\mathbf{v}^\intercal \hat{\mathbf{\Phi}}^{\text{WG-lim}} \mathbf{v} > 0$ which proves that $\hat{\mathbf{\Phi}}^{\text{WG-lim}}$ is positive definite in that subspace.

We begin by considering $\mathbf{v} \in S_2$, which satisfy $\mathbf{1}^\intercal \mathbf{v} = \mathbf{0}$, and by hypothesis $\mathbf{v} \neq \mathbf{0}$. Therefore $\mathbf{v}^\intercal \mathbf{J} \mathbf{v} = 0$ in this subspace, which results in

$$\mathbf{v}^\intercal \hat{\mathbf{\Phi}}^{\text{WG-lim}} \mathbf{v} = \frac{1}{1-\tilde{\varphi}} \mathbf{v}^\intercal \mathbf{\Phi} \mathbf{v} > 0,$$

where the final inequality follows since the true kinship matrix is positive definite and $1 - \tilde{\varphi} > 0$.

Lastly, we consider $\mathbf{v} \in S_1$, which are necessarily of the form $\mathbf{v} = \beta\mathbf{1}$, and by hypothesis $\beta \neq 0$. Therefore

$$\mathbf{v}^\intercal \hat{\mathbf{\Phi}}^{\text{WG-lim}} \mathbf{v} = \frac{\beta^2}{1-\tilde{\varphi}} (\mathbf{1}^\intercal \mathbf{\Phi} \mathbf{1} - \tilde{\varphi} n^2) = \frac{\beta^2 n^2}{1-\tilde{\varphi}} (\bar{\varphi} - \tilde{\varphi}),$$

where $\bar{\varphi}$ is the overall mean kinship value, while $\tilde{\varphi}$ is the mean of the off-diagonal kinship values only (Eq. (16)). Note that $\beta^2, n^2, 1 - \tilde{\varphi} > 0$, so the desired result follows if $\tilde{\varphi} < \bar{\varphi}$, which is proven in Appendix B. In general it is true that $\tilde{\varphi} \leq \bar{\varphi}$, and $\tilde{\varphi} = \bar{\varphi}$ occurs if and only if the true kinship

matrix has the degenerate form $\mathbf{\Phi} = \bar{\varphi}\mathbf{J}$, which is a singular matrix not expected to be observed in any real scenarios (in this case $\hat{\mathbf{\Phi}}^{\text{WG-lim}}$ is a matrix full of zeroes).

B Mean kinship inequalities

Denote the mean of the diagonal kinship terms as $\bar{d} = \frac{1}{n}\sum_{j=1}^{n}\varphi_{jj}$. Here we prove that

$$0 \leq \tilde{\varphi} \leq \bar{\varphi} \leq \bar{d} \leq 1,$$

with each of $\tilde{\varphi} = \bar{\varphi}$ and $\bar{\varphi} = \bar{d}$ if and only if all kinship values are equal.

The inequalities $0 \leq \bar{\varphi} \leq \bar{d} \leq 1$ follow directly from previous work, applied to a kinship matrix rather than a coancestry matrix as done originally, as the proof required solely a covariance matrix with values between 0 and 1 (Ochoa and Storey, 2021). Recall that $\tilde{\varphi}$ is defined in Eq. (16). The lower bound $0 \leq \tilde{\varphi}$ follows since every kinship value is non-negative. Note that $\bar{\varphi}$ and $\tilde{\varphi}$ are related by

$$\bar{\varphi} = \frac{\tilde{\varphi}(n-1) + \bar{d}}{n}. \tag{17}$$

Applying $\bar{\varphi} \leq \bar{d}$ to Eq. (17) and simplifying yields $\tilde{\varphi} \leq \bar{d}$. Lastly, since $\bar{\varphi} - \tilde{\varphi} = (\bar{d} - \tilde{\varphi})/n$ (from rearranging Eq. (17)), it also follows that $\tilde{\varphi} \leq \bar{\varphi}$, as desired. Furthermore, $\tilde{\varphi} = \bar{\varphi}$ holds if and only if all $\varphi_{jk} = \bar{d}$, since that is necessary and sufficient for $\bar{\varphi} = \bar{d}$.