The effect of population kinship estimation bias in structured genetic association studies

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 control for population structure in genetic association studies. Recent work found that the most common kinship estimators can be severely biased. In this work, we investigate the effect of this kinship bias on the downstream application of genetic association. Remarkably, kinship bias does not affect genetic associations based on either Principal Components Analysis (PCA) or Linear Mixed-effects Models (LMM). We present empirical observations using simulated data, then explain these observations theoretically using linear algebra. In particular, the exact form of the kinship bias is compensated for by fitting the intercept in both PCA and LMM approaches, which model population structure via covariates, suggesting that only models with this precise arrangement will be robust to this kinship bias. [TODO: does it change for WG?]

Contents

1 TO DO

Plan for finishing paper. Will pursue these items in this order:

- Prove WG equivalence just like for the standard kinship estimator. NOTE: I couldn't come up with a transformation like the centering matrix that carries over easily to the matrix square root, so it may be challenging. But this makes me think that it won't perform the same.

- Plot precision-recall curves for methods (or AUCs only), to show which methods actually perform.

- Switch from admixture-simulated genotypes to real genotypes (1000 Genomes or HDGP), because there are huge differences between LMM and PCA in real data (but not in admixture simulation; could also use a family simulation where there's also LMM-PCA gap). This also simplifies the methods (for real only, not family sim).

    - Only risk is that PCA "approximation" may fail in this real dataset (so that the true PCA and biased PCA models are no longer equivalent). In this case it might be worth showing both the admixture simulation and the real genotypes results.

Done in code, missing in paper:

- Add MOR (mean-of-ratios) version of standard kinship. See Eq. (6) and Eq. (8).

- Add another kinship estimator (Weir-Goudet: WG). See Eqs. (12) and (13).

2 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).

3 Results

3.1 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 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).

[NOTE: current results only include ROM (ratio-of-means) version of standard kinship estimator, which converges best to limit but is not what most people use.]

We found that all the methods yield highly correlated values (for p-values in Fig. 1, effect sizes $\hat{\beta}$ in Fig. 2). 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).

3.2 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.

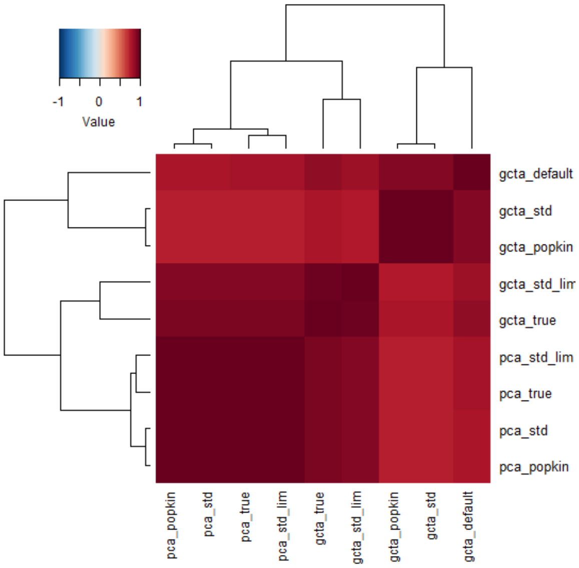

Figure 1: **Heatmap and dendrogram for p-values of GCTA and PCA with different kinship matrices ($m = 10,000$).**

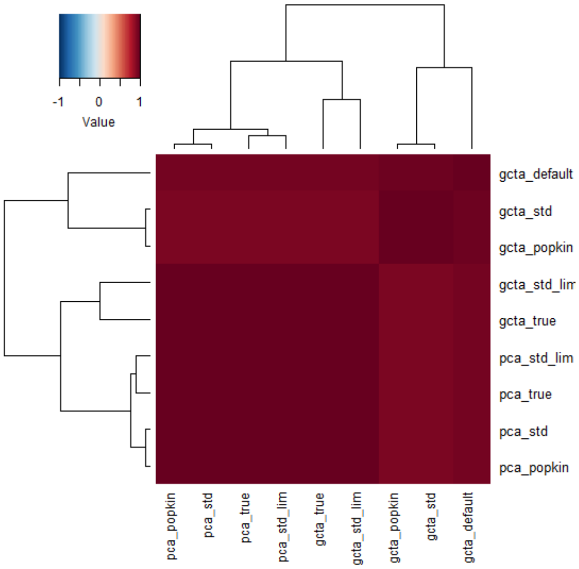

Figure 2: **Heatmap and dendrogram for $\hat{\beta}$ of GCTA and PCA with different kinship matrices ($m = 10,000$).**

Therefore, our theoretical results only consider the true kinship matrix $\mathbf{\Phi}$ and the limit of the standard kinship estimator (see **Methods**, Eq. (7)), which can be stated in matrix notation as

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

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, $\varphi = \frac{1}{n}\mathbf{1}^\intercal \mathbf{\Phi}$ 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.

3.2.1 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

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

It can be directly verified that this matrix square root indeed satisfies $\left(\hat{\mathbf{\Phi}}^{\text{std-lim}}\right)^{\frac{1}{2}} \left(\left(\hat{\mathbf{\Phi}}^{\text{std-lim}}\right)^{\frac{1}{2}}\right)^{\intercal} = \hat{\mathbf{\Phi}}^{\text{std-lim}}$ as given in Eq. (1).

3.2.2 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

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

where α is the intercept coefficient, β is the genetic effect coefficient, ϵ are random independent residuals ($\epsilon \sim \text{Normal}(\mathbf{0}, \mathbf{I}\sigma^2_\epsilon)$ for some σ^2_ϵ), 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

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

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

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

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

3.2.3 Equivalent LMM fit under standard kinship bias

In the LMM of Eq. (4), $\mathbf{y}$, $\mathbf{x}_i$, and $\mathbf{\Phi}$ are given, while the coefficients α, β, σ, and the random effects $\mathbf{r}$ and ϵ 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 $\mathbf{\Phi}$ is given, while in the other we provide $\mathbf{\Phi}' = \hat{\mathbf{\Phi}}^{\text{std-lim}}$ instead. The first fit will result in the unprimed variables below, while we distinguish the second fit using primed variables, namely:

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

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 $\beta' = \beta$, $\epsilon' = \epsilon$, 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

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

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

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

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

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

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

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

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.

3.2.4 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

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

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 γ_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 $\mathbf{U}_d\gamma_d$ 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 $\mathbf{\Lambda}^{\frac{1}{2}}$ simply 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

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

so that the length-n vector γ of coefficients for all the n eigenvectors is given by $\gamma = \sigma\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{r}$. 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.

3.2.5 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

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

and we again assume that the focal parameter $\beta' = \beta$ and the residuals $\epsilon' = \epsilon$ 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

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

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

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

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.

4 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 estiator, so we expect an approximately similar performance between those two and using the true kinship matrix.

5 Methods

5.1 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

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

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.

5.2 Kinship estimation

5.2.1 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

$$\hat{\varphi}_{jk}^{\text{std-rom}} = \frac{\sum\limits_{i=1}^{m} \left(x_{ij} - 2\hat{p}_i\right)\left(x_{ik} - 2\hat{p}_i\right)}{\sum\limits_{i=1}^{m} 4\hat{p}_i\left(1 - \hat{p}_i\right)} \tag{6}$$

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

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, or $\frac{1}{n^2}\sum_{j=1}^{n}\sum_{k=1}^{n} \hat{\varphi}_{jk}^{\text{std-rom}}$, is algebraically zero, regardless of the true value of the mean kinship.)

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

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

This estimator does not have closed-form limit, but it is well approximated by Eq. (7) 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. (7), 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 $\mathbf{w} = (\mathbf{1}^\intercal \mathbf{\Phi}^{-1}\mathbf{1})^{-1}\mathbf{1}^\intercal \mathbf{\Phi}^{-1}$.

5.2.2 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 Eq. (8). In fact, the GCTA kinship estimates for two different individuals is identical to Eq. (8) (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):

$$\hat{\varphi}_{jj}^{\text{GCTA}} = \frac{1}{2} + \frac{1}{m} \sum_{i=1}^{m} \frac{x_{ij}^2 - (1 + 2\hat{p}_i)\, x_{ij} + 2\hat{p}_i^2}{4\hat{p}_i\,(1 - \hat{p}_i)} \tag{9}$$

$$\xrightarrow[n,m\to\infty]{\text{a.s.}} \frac{\varphi_{jj} - \bar{\varphi}_j}{1 - \bar{\varphi}}. \tag{10}$$

5.2.3 Popkin kinship estimator

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

$$\begin{aligned} A_{jk} &= \frac{1}{m} \sum_{i=1}^{m} (x_{ij} - 1)(x_{ik} - 1) - 1, \\ \hat{A}_{\min} &= \min_{u \neq v} \frac{1}{|S_u||S_v|} \sum_{j \in S_u} \sum_{k \in S_v} A_{jk}, \\ \hat{\varphi}_{jk}^{\text{new}} &= 1 - \frac{A_{jk}}{\hat{A}_{\min}}, \end{aligned} \tag{11}$$

where S_u are subpopulations that partition individuals. This 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. 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.

5.2.4 Weir-Goudet kinship estimator

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

$$\hat{\varphi}_{jk}^{\text{WG}} = 1 - \frac{A_{jk}}{\hat{A}_{\text{avg}}} \tag{12}$$

$$\xrightarrow[m\to\infty]{\text{a.s.}} \frac{\varphi_{jk} - \tilde{\varphi}}{1 - \tilde{\varphi}}, \tag{13}$$

where A_{jk} is as in Eq. (11),

$$\hat{A}_{\text{avg}} = \frac{2}{n(n-1)} \sum_{j=2}^{n} \sum_{k=1}^{j-1} A_{jk},$$

$$\tilde{\varphi} = \frac{2}{n(n-1)} \sum_{j=2}^{n} \sum_{k=1}^{j-1} \varphi_{jk}.$$

Thus, the WG estimator resembles the popkin estimator in Eq. (11), 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).

5.3 Software

TODO: state versions, download links, etc.

GCTA (Yang et al., 2011).

PCA: implemented regression in R (function lm). p-values based on F-test.

popkin is estimated with the popkin R package.

All other estimators are calculated using the popkinsuppl R package.

5.4 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.]

5.4.1 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

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

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

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

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

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

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

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

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

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

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

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

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

5.4.2 Admixture simulation for genotype matrices

TODO: describe the BNPSD simulation.

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.

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”. Submitted, preprint at `http://biorxiv.org/content/early/2016/10/27/083915`.

Ochoa, Alejandro and John D. Storey (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.

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.