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

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

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

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

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

**Abstract**

Population kinship matrices are estimated for a variety of applications, including estimation of heritability and to control for population structure in genetic association studies. Recent work found that the most common family of kinship estimators can be severely biased. In this work, we investigate the effect of this kinship bias on the two downstream applications of heritability estimation and genetic association. We present a novel trait simulation strategy that accurately parametrizes heritability, even when utilizing real genotypes. Using these simulations, we find that heritability estimation becomes biased when using such biased kinship matrices. Remarkably however, this kinship bias does not affect genetic associations based on either Principal Components Analysis (PCA) or Linear Mixed-effects Models (LMM). Lastly, we explain our empirical observations using theory. In particular, the exact form of the bias of the standard kinship estimator is such that it is compensated for by fitting the intercept in both PCA and LMM approaches, which model population structure via covariates, suggesting that downstream applications without this precise arrangement will not be robust to this kinship bias.

# 1 Introduction

We previously found that the most common family of kinship estimators is biased (Ochoa and Storey, 2016b).

GCTA estimated the heritability of human height from population data (Yang et al., 2010). GCTA is based on the biased kinship estimator (Yang et al., 2010; Yang et al., 2011).

# 2 Methods

## 2.1 Genetic model

Suppose there are $m$ biallelic loci and $n$ diploid individuals. The genotype $x_{ij}$ 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 $\varphi_{jk}$ 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, 2016a; Ochoa and Storey, 2016b) the expectation and covariance are given by

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

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

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

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

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

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

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

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

The covariance matrix of the trait is

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

Therefore, we can write the covariance in terms of the heritability and the overall variance scale, in a formulation that matches previous work [TODO: add citations]:

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

Since the above expectation and covariance is conditioned on the choice of ancestral population, and given in terms of parameters that depend on it ($p_i$ and $\boldsymbol{\Phi}$), then the parameters $\mu, \sigma^2, h^2$ are all also dependent on the choice of ancestral population.

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

Table 1: **Mathematical notation.**

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

The factor of two in front of $\mathbf{\Phi}$ is traditionally there so that for an unstructured population $2\mathbf{\Phi} = \mathbf{I}_n$, in which case the trait covariance simplifies to $\mathrm{Cov}(\mathbf{y}) = \sigma^2\mathbf{I}_n$ for any value of $h^2$. More broadly, the variance of the trait for any outbred individual is $\sigma^2$ under this parametrization.

## 2.2 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 $(\mu, \sigma^2, h^2)$. The goal is to choose the intercept $\alpha$ and draw random effect sizes $\beta$ 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 $\mathbf{\Phi}$).

### 2.2.1 Scaling effect sizes

The initial effect sizes $\beta_i$ 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 $\beta_i$ by $\sigma_0$ (which results in a variance of 1) and then multiply by $h\sigma$ (which results in the desired variance of $h^2\sigma^2$). Combining both steps, the update is

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

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

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

Although this estimator is unbiased ($\mathrm{E}[\hat{\mathbf{p}}] = \mathbf{p}$), the resulting variance estimates of interest $\hat{p}_i\,(1-\hat{p}_i)$ are downwardly biased (Ochoa and Storey, 2016b):

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

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

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

has an expectation of

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

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

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

results in the desired variance.

### 2.2.2 Centering the trait

Here we consider the problem of selecting the intercept coefficient $\alpha$ that, together with the previous effect size coefficient vector $\beta$, result in the desired trait mean $\mu$.

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

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

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

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

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

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

is a good approximation.

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

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

However, this operation introduces a distortion in the covariance of the genotypes (Ochoa and Storey, 2016b):

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

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

### 2.3 Admixture simulation for genotype matrices

TODO: describe the BNPSD simulation.

### 2.4 Kinship estimation

TODO: Present our popkin estimator, standard estimator (MOR and ROM versions), and GCTA estimator (like standard-MOR but with diff diagonal), with limits.

### 2.5 Heritability and genetic association software

TODO: state versions, download links, etc.

Outline:

- SOLAR-Eclipse (herit only) (Almasy and Blangero, 1998).
- GCTA (both heritability and genetic association) (Yang et al., 2011).
- PCA: our optimized R implementation [TODO: cite Yao and Ochoa]. Genetic association only.

## 3 Results

### 3.1 Empirical demonstration of biases in heritability estimation

TODO: add results using SOLAR and GCTA, varying the input kinship matrices (estimates and limits from methods).

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

TODO: add results showing that p-values are highly correlated under all variations of these kinship matrices, for PCA and LMM (under simple admixture simulation without family structure).

### 3.3 Theoretical justification of empirical observations

Here, to eliminate random estimation noise from the analysis (which our empirical evaluations suggest play a minor role), we shall focus on the limiting bias of the standard kinship estimator. Therefore, our theoretical results only consider the true kinship matrix $\mathbf{\Phi}$ and the limit of the standard kinship estimator, given by

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

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

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

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

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

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

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

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

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

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

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

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

□

Now we may prove the desired theorem.

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

*Proof.* Since $\mathbf{1}_n$ is in both rowspaces, it suffices to consider vectors $\mathbf{v}$ orthogonal to $\mathbf{1}_n$, which satisfy $\mathbf{C}\mathbf{v} = \mathbf{v}$. We shall prove below that any such vector is in the nullspace of $\mathbf{\Phi}$ if and only if it is in the nullspace of $\hat{\mathbf{\Phi}}^{\text{std-lim}}$. Then, since the nullspace of $\mathbf{\Phi}$ and $\mathbf{1}_n$ is the same as the nullspace of $\hat{\mathbf{\Phi}}^{\text{std-lim}}$ and $\mathbf{1}_n$, it follows from the fundamental theorem of linear algebra that their rowspaces are also the same.

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

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

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

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

### 3.3.1 Theoretical justification for PCA genetic association

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

$$\mathbf{y} = \mathbf{1}_n\alpha + \mathbf{x}_i\beta_i + \mathbf{U}_r\gamma_r + \epsilon, \tag{2}$$

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

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

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

is only approximately true. In other words, centering the kinship matrix first and then approximating to the top $r$ dimensions is not exactly equal to first approximating to $r$ dimensions and then centering, although in simulations we found it to be a very good approximation (especially in the absence of family structure, which is where use of PCA is most appropriate anyway [TODO: cite

Yao, other papers]). Therefore, in this case the rowspace of $\mathbf{\Phi}_r$ and $\mathbf{1}_n$ only approximately equals the rowspace of $\hat{\mathbf{\Phi}}_r^{\text{std-lim}}$ and $\mathbf{1}_n$.

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

### 3.3.2 Theoretical justification for LMM genetic association

...

## 4 Discussion

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

...

## References


Almasy, Laura and John Blangero (1, 1998). "Multipoint Quantitative-Trait Linkage Analysis in General Pedigrees". *The American Journal of Human Genetics* 62(5), pp. 1198–1211.

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

— (2016b). "$F_{\text{ST}}$ and kinship for arbitrary population structures II: Method of moments estimators". Submitted, preprint at `http://biorxiv.org/content/early/2016/10/27/083923`.

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. (7, 2011). "GCTA: a tool for genome-wide complex trait analysis". *Am. J. Hum. Genet.* 88(1), pp. 76–82.