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

...

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 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, 2016a; Ochoa and Storey, 2016b) the expectation and covariance are given by

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

where $\mathbf{x}_i$ is the length-n column vector of genotypes at locus i, $\mathbf{X} = (\mathbf{x}_i^\mathsf{T})$ 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 T 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,

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

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

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

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

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

The covariance matrix of the trait is

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

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]:

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

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 $\mathbf{\Phi}$), then the parameters μ, σ^2, h^2 are all also dependent on the choice of ancestral population.

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

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

Table 1: **Mathematical notation.**

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

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 (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 $\boldsymbol{\Phi}$).

2.2.1 Scaling effect sizes

The initial effect sizes 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

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

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

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

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

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

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 $\bar{\varphi} = \frac{1}{n^2} \mathbf{1}_n^{\mathsf{T}} \boldsymbol{\Phi} \mathbf{1}_n$ 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

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 to be

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

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

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

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

$$\frac{1}{m_1}\sum_{i=1}^{m_1}\mathbf{x}_i\beta_i \approx \left(\frac{1}{m_1}\sum_{i=1}^{m_1}\mathbf{x}_i\right)\left(\frac{1}{m_1}\sum_{i=1}^{m_1}\beta_i\right)$$

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 $\alpha = \mu - 2\hat{\mathbf{p}}^\mathsf{T}\beta$, 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):

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

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

5

3 Results

...

4 Discussion

...

References

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

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