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

- **Background**: commonly kinship estimators can have severe biases.
- **Results**: kinship matrices of different bias types result in equal association statistics and performance in simulations and 1000 Genomes.
- Intercept and relatedness (PCs in Principal Component Analysis (PCA), random effect in Linear Mixed-effects Models (LMM)) coefficients compensate for the kinship bias.

Model

$x_{ij} \in \{0, 1, 2\}$: genotype of ind. j, biallelic SNP i.

p_i: ancestral allele frequency. φ_{ij}: kinship coefficient

$$\mathrm{E}[\boldsymbol{x_i}] = 2p_i\mathbf{1}, \qquad \mathrm{Cov}(\boldsymbol{x_i}) = 4p_i(1-p_i)\boldsymbol{\Phi}$$

where $\boldsymbol{x}_i = (x_{ij})$ is the length-n column vector of genotypes at locus i, $\boldsymbol{\Phi} = (\varphi_{ij})$ is the $n \times n$ kinship matrix, and $\mathbf{1}$ is a length-n column vector of ones [1].

Kinship estimators

Standard estimator

Ratio-of-means (ROM) [1,2]:

$$\hat{p}_i = \frac{1}{2n}\sum_{j=1}^{n} x_{ij}\,, \qquad \hat{\varphi}_{ij}^{\text{std-ROM}} = \frac{\sum_{i=1}^{m}(x_{ij}-2\hat{p}_i)(x_{ik}-2\hat{p}_i)}{\sum_i^m 4\hat{p}_i(1-\hat{p}_i)} \xrightarrow[m\to\infty]{a.s.} \frac{\varphi_{ij}-\bar{\varphi}_j-\bar{\varphi}_k+\bar{\varphi}}{1-\bar{\varphi}}$$

Mean-of-ratios (MOR)

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

Popkin estimator[1]

$$A_{ij} = \frac{1}{m}\sum_{i=1}^{m} w_i\left((x_{ij}-1)\right)(x_{ik}-1)-1)\,, \qquad \hat{\varphi}_{ij}^{\text{popkin}} = 1 - \frac{A_{ij}}{\min\limits_{j\neq k} A_{jk}},$$

$w_i = 1$ for ROM, $\quad \hat{\varphi}_{ij}^{\text{popkin-ROM}} \xrightarrow[m\to\infty]{a.s.} \varphi_{ij}$,

$w_i = \left(\hat{p}_i(1-\hat{p}_i)\right)^{-1}$ for MOR.

Association models

LMM [3,4]: $\quad \boldsymbol{y} = \mathbf{1}\alpha + \boldsymbol{x}_i\beta_i + \boldsymbol{s} + \boldsymbol{\epsilon}, \qquad \boldsymbol{s}\sim\mathrm{Normal}(\mathbf{0}, 2\sigma^2\boldsymbol{\Phi})$

PCA [3,4]: $\quad \boldsymbol{y} = \mathbf{1}\alpha + \boldsymbol{x}_i\beta_i + \boldsymbol{U}_d\boldsymbol{\gamma}_d + \boldsymbol{\epsilon}, \qquad \boldsymbol{\Phi} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{\mathrm{T}}$

Empirical analysis using 1000 Genomes

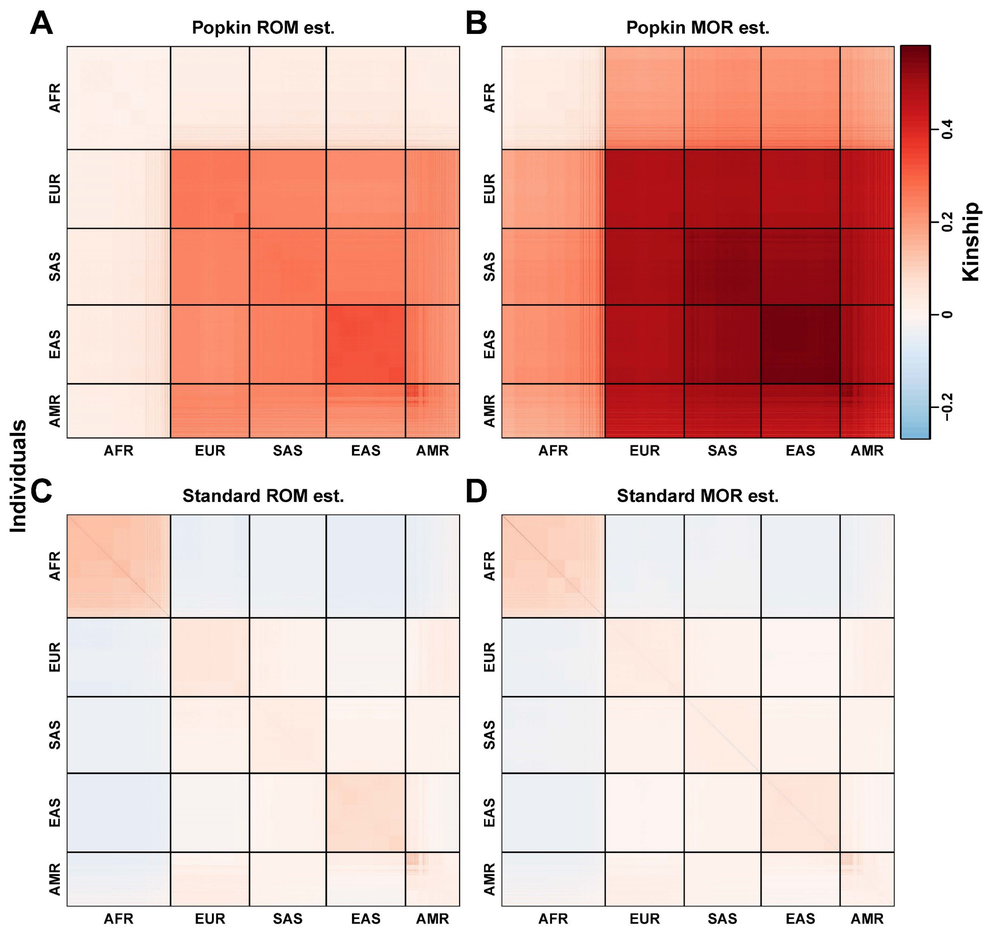

Figure 1: **Kinship estimates on 1000 Genomes**

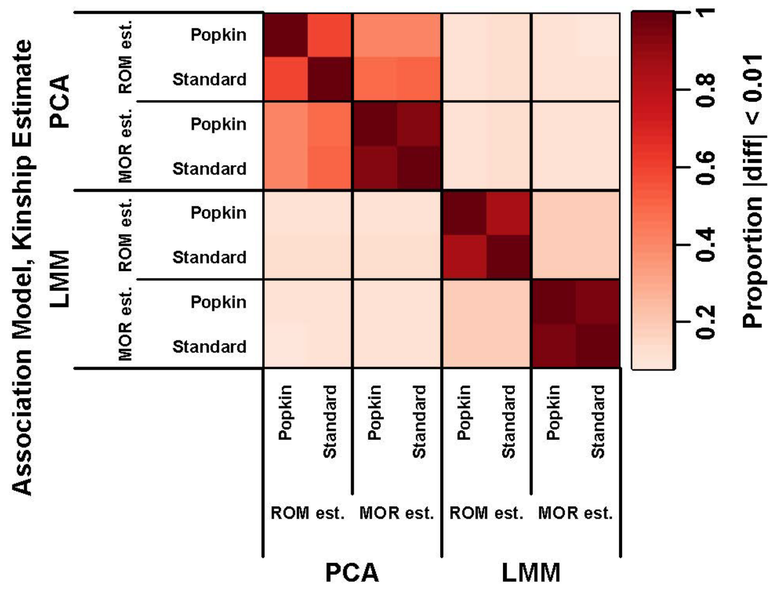

Figure 2: **Approximate agreement between p-values on 1000 Genomes**

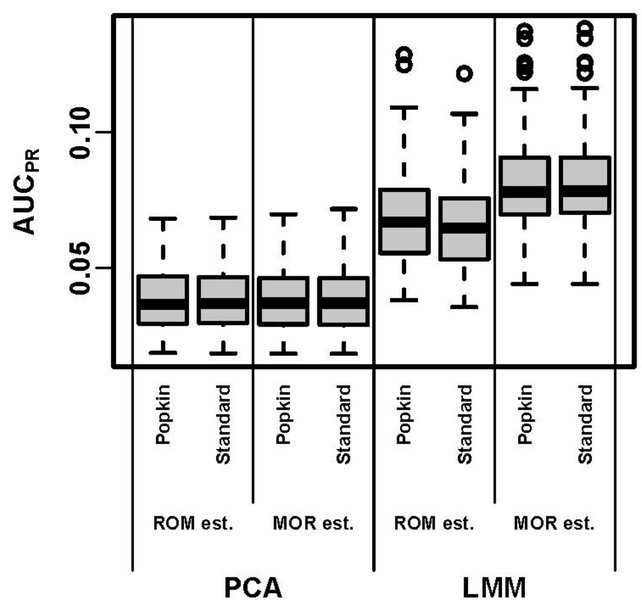

Figure 3: **Distributions of Area Under the Precision-Recall Curve ($\mathbf{AUC_{PR}}$) on 1000 Genomes**

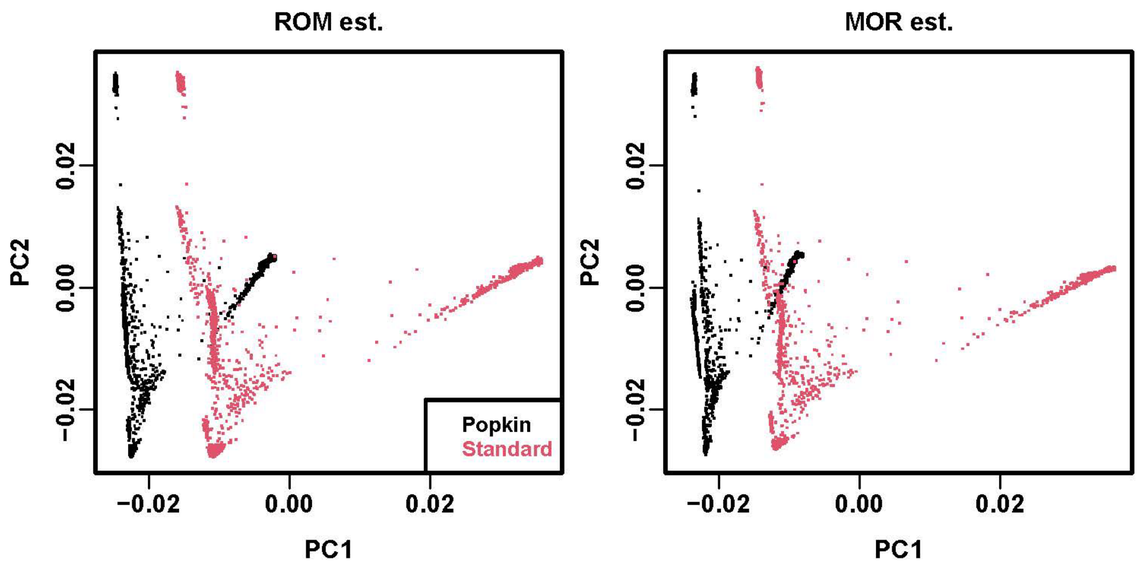

Figure 4: **Visualization of PC shift due to kinship biases**

Proof of association invariability to common kinship biases

For standard kinship estimator:

$$\boldsymbol{\Phi}' = \frac{1}{1-\bar{\varphi}}\mathbf{C}\boldsymbol{\Phi}\mathbf{C}, \qquad \text{where } \mathbf{C} = \mathbf{I} - \frac{1}{\mathrm{n}}\mathbf{1}\mathbf{1}^{\mathrm{T}} \text{ is center matrix.}$$

In LMM model, kinship bias compensated by intercept:

$$\boldsymbol{y} = \mathbf{1}\alpha' + \boldsymbol{x}_i\beta_i + \boldsymbol{s}' + \boldsymbol{\epsilon}, \qquad \boldsymbol{s}' = \mathbf{C}\boldsymbol{s}\sim\mathrm{Normal}\left(\mathbf{0}, 2\sigma'^2\boldsymbol{\Phi}'\right),$$

$$\sigma'^2 = (1-\bar{\varphi})\sigma^2, \qquad \boldsymbol{s}' = \boldsymbol{s} - 1\bar{\boldsymbol{s}}, \qquad \alpha' = \alpha + \bar{\boldsymbol{s}}, \qquad \bar{\boldsymbol{s}}\sim\mathrm{Normal}(0, \sigma^2\bar{\varphi})$$

Similarly, in PCA model:

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

$$\boldsymbol{y} = \mathbf{1}\alpha' + \boldsymbol{x}_i\beta_i + \boldsymbol{U'}_d\boldsymbol{\gamma'}_d + \boldsymbol{\epsilon},$$

$$\boldsymbol{\gamma'}_d = \boldsymbol{\gamma}_d, \qquad \boldsymbol{U'}_d\boldsymbol{\gamma'}_d = \boldsymbol{U}_d\boldsymbol{\gamma}_d - \mathbf{1}\overline{\boldsymbol{U}_d\boldsymbol{\gamma}_d}, \qquad \alpha' = \alpha + \overline{\boldsymbol{U}_d\boldsymbol{\gamma}_d}$$

Reference

1.Ochoa, Alejandro and John D. Storey (2021). “Estimating FST and kinship for arbitrary population structures”. PLoS Genet 17(1), e1009241.

2.Bhatia, Gaurav et al. (2013). “Estimating and interpreting FST: the impact of rare variants”. Genome Res. 23(9), pp. 1514–1521.

3.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.

4. Yao, Yiqi and Alejandro Ochoa (2022). Limitations of principal components in quantitative genetic association models for human studies. bioRxiv, p. 2022.03.25.485885.