Kinship estimation bias carries over to heritability estimation bias using variance components

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: Standard kinship estimators can have severe biases.
Heritability estimation requires unbiased estimates of the random effect coefficient, which is biased when the Standard kinship estimator is used.
Results: Using Standard kinship estimators result in a downwardly biased heritability estimation when there is population structure. Using an unbiased kinship estimator addresses this source of bias.

Model

$x_{ij} \in \{0,1,2\}$: genotype of ind. j, biallelic SNP i. The number of loci is m, the number of individuals is n. p_i: ancestral allele frequency. φ_{ij}: kinship coefficient

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

where $\boldsymbol{X} = (x_{ij})$ is the complete m × n genotype matrix, $\boldsymbol{\Phi} = (\varphi_{ij})$ is the $n \times n$ kinship matrix, and $\mathbf{1}_n$ is a length-n column vector of ones [1].

Kinship estimators

Standard estimator

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

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

where $\hat{p}_i = \frac{1}{2n}\sum_{j=1}^{n} x_{ij}$.

Mean-of-ratios (MOR, most commonly used one):

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

Popkin estimator (ROM is unbiased even with population structure) **[1]**

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

$w_i = 1$ for ROM, $\quad \hat{\varphi}_{ij}^{\mathrm{popkin-ROM}} \xrightarrow[m\to\infty]{a.s.} \varphi_{ij}, \qquad w_i = \left(\hat{p}_i(1-\hat{p}_i)\right)^{-1}$ for MOR.

Genetic model

The quantitative trait vector y for all individuals is assumed to follow a linear polygenic model

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

where α is the intercept, $\boldsymbol{\beta} = (\beta_i)$ is a vector of genetic effect coefficients for each locus i, and ϵ is a vector of non-genetic effects. Let us shift the mean of genotypes to the intercept and denote $\boldsymbol{s} = \boldsymbol{X}^T\boldsymbol{\beta}$, then:

$$\boldsymbol{y} = \mathbf{1}_n\alpha + \boldsymbol{s} + \boldsymbol{\epsilon},$$

$$\boldsymbol{s} \sim N\left(0, 2\sigma_g^2\boldsymbol{\Phi}\right), \qquad \boldsymbol{\epsilon} \sim N\left(0, \sigma_e^2\boldsymbol{I}_n\right), \qquad \boldsymbol{s}+\boldsymbol{\epsilon} \sim N\left(0, 2\sigma_g^2\boldsymbol{\Phi} + \sigma_e^2\boldsymbol{I}_n\right),$$

and the narrow-sense heritability h^2 is defined as:

$$h^2 = \frac{\sigma_g^2}{\sigma_g^2+\sigma_e^2}.$$

Heritability estimation bias due to kinship bias

For standard ROM estimator: $\boldsymbol{\Phi}' = \frac{1}{1-\bar{\varphi}}\boldsymbol{C\Phi C}$, where $\boldsymbol{C} = \boldsymbol{I} - \frac{1}{\mathrm{n}}\mathbf{1}\mathbf{1}^{\mathrm{T}}$ is the centering matrix. Using standard ROM results in biased genetic variance component [3]:

$$\boldsymbol{y} = \mathbf{1}_n\alpha' + \boldsymbol{s}' + \boldsymbol{\epsilon}, \qquad \boldsymbol{s}' = \boldsymbol{Cs} \sim N\left(\mathbf{0}, 2\sigma_g^{2\prime}\boldsymbol{\Phi}'\right),$$
$$\boldsymbol{s}' = \boldsymbol{s} - \mathbf{1}_n\bar{\boldsymbol{s}}, \qquad \alpha' = \alpha + \bar{\boldsymbol{s}}, \qquad \bar{\boldsymbol{s}} \sim N\left(0, \sigma_g^2\bar{\varphi}\right),$$
$$\sigma_g^{2\prime} = (1-\bar{\varphi})\sigma_g^2, \qquad \sigma_e^{2\prime} = \sigma_e^2,$$

where $\bar{\varphi}$ is the mean value of the unbiased kinship matrix. Then, the heritability is biased [4]:

$$h_{std}^2 = \frac{\sigma_g^{2\prime}}{\sigma_g^{2\prime}+\sigma_e^{2\prime}} = h_{true}^2\frac{1-\bar{\varphi}}{1-\bar{\varphi}h_{true}^2}.$$

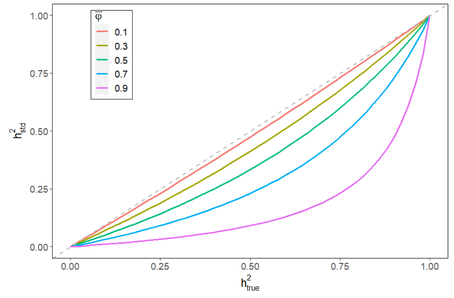

Figure 1: **Relationship between true heritability and biased estimates.**

Results

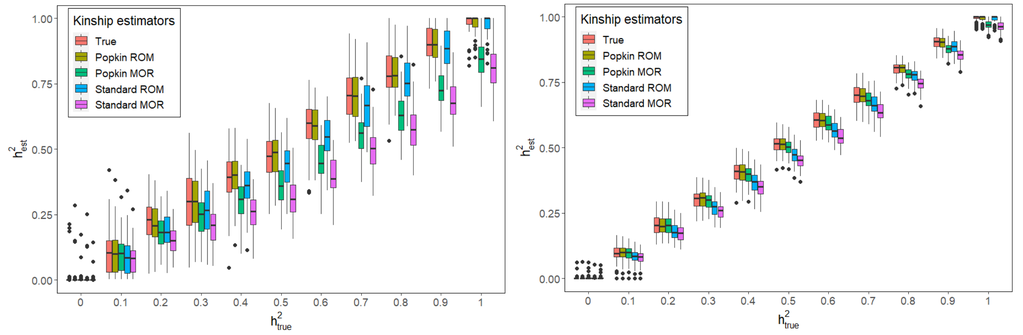

Figure 2: **Simulation evaluation shows that only Popkin ROM results in unbiased heritability estimates.** The left and right panels show simulation results for admixture structure only and admixture plus family structure, respectively.

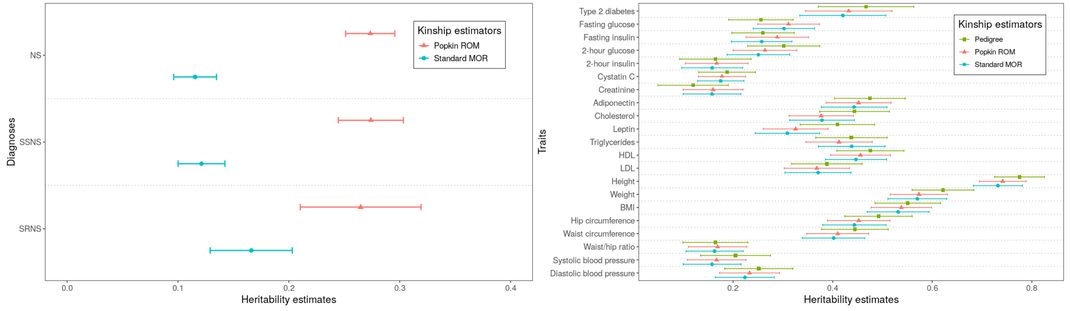

Figure 3: **Magnitude of heritability bias in real datasets depends on the amount of population structure, as quantified by the mean kinship coefficient.** The left panel shows results from Nephrotic Syndrome multiethnic cohort (high mean kinship: 0.13) while the right panel shows results from the San Antonio Family Study (low mean kinship: 0.07).

References

1. Ochoa, Alejandro, and John D. Storey. "Estimating F ST and kinship for arbitrary population structures." PLoS genetics 17.1 (2021): e1009241.
2. Bhatia, Gaurav, et al. "Estimating and interpreting FST: the impact of rare variants." Genome research 23.9 (2013): 1514-1521.
3. Hou, Zhuoran, and Alejandro Ochoa. "Genetic association models are robust to common population kinship estimation biases." Genetics 224.1 (2023): iyad030.
4. Chen, Danfeng, and John D. Storey. "How Kinship estimation Bias Propagates to Heritability. " The 72nd Annual Meeting of The American Society of Human Genetics (2022).