Supplementary Information
Confidence intervals for heritability via Haseman-Elston regression

Tamar Sofer

Contents

1 Mathematical derivation

Suppose that a quantitative trait Y, measured on n individuals, follows the regression model

$$y_i = \boldsymbol{w}_i^T \boldsymbol{\beta} + \epsilon_i, \qquad i = 1, \ldots, n$$

where

$$E[\boldsymbol{\epsilon}] = \mathbf{0} \quad (1)$$

$$\text{var}[\boldsymbol{\epsilon}] = \sigma_e^2 \mathbf{I}_{n\times n} + \sigma_a^2 \mathbf{A}_1 + \ldots + \sigma_k^2 \mathbf{K} = \boldsymbol{\Sigma} \quad (2)$$

and $\mathbf{A}, \ldots, \mathbf{K}$ are $n \times n$ matrices modeling correlations between individuals. Let $a_{i,j}, \ldots, k_{i,j}$ denote the i, j entries of the matrices $\mathbf{A}, \ldots, \mathbf{K}$. Assuming that random effects due to $\mathbf{A}, \ldots, \mathbf{K}$ are independent, we have that:

$$E[\epsilon_i \epsilon_j] = \text{cov}(\epsilon_i, \epsilon_j) = \sigma_e^2 \mathcal{I}_{(i=j)} + \sigma_a^2 a_{i,j} + \ldots + \sigma_k^2 k_{i,j}.$$

Let $\widehat{\boldsymbol{\beta}}$ be an unbiased estimator of $\boldsymbol{\beta}$, and let $\hat{\epsilon}_i = y_i - \boldsymbol{w}_i^T \widehat{\boldsymbol{\beta}}$ be an estimator $\epsilon_i, i = 1, \ldots, n$. We estimate the variance components in a residual regression, i.e. by taking the vector all unique pairs of residuals $\hat{\epsilon}_i \hat{\epsilon}_j, i \leq j$ (we can do it by taking the upper diagonal sub-matrix of $\hat{\boldsymbol{\epsilon}} \hat{\boldsymbol{\epsilon}}^T$ that includes the diagonal), denoted by $\tilde{\boldsymbol{\epsilon}}^d$ and regressing it according to the above model. The regression design matrix is given by:

$$\mathbf{X} = \begin{pmatrix} 1 & a_{1,1} & \ldots & k_{1,1} \\ 0 & a_{1,2} & \ldots & k_{1,2} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & a_{1,n} & \ldots & k_{1,n} \\ 1 & a_{2,2} & \ldots & k_{2,2} \\ 0 & a_{2,3} & \ldots & k_{2,3} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & a_{2,n} & \ldots & k_{2,n} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & a_{n-1,n-1} & \ldots & k_{n-1,n-1} \\ 0 & a_{n-1,n} & \ldots & k_{n-1,n} \\ 1 & a_{n,n} & \ldots & k_{n,n} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & a_{1,2} & \ldots & k_{1,2} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & a_{1,n} & \ldots & k_{1,n} \\ 1 & 1 & 1 & 1 \\ 0 & a_{2,3} & \ldots & k_{2,3} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & a_{2,n} & \ldots & k_{2,n} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 1 & 1 \\ 0 & a_{n-1,n} & \ldots & k_{n-1,n} \\ 1 & 1 & 1 & 1 \end{pmatrix}$$

(because $a_{i,i}, \ldots, k_{i,i} = 1$ for all i). Denote, for simplicity of presentation, the vector of off-diagonal elements of $\mathbf{A}, \ldots, \mathbf{K}$ by $\boldsymbol{l} = (l_{1,2}, l_{1,3}, \ldots, l_{1,n}, l_{2,3}, \ldots, l_{n-1,n})^T, l = 1, \ldots, k$, and the vector of off-diagonal elements of $\hat{\boldsymbol{\epsilon}}\hat{\boldsymbol{\epsilon}}^T$ by $\tilde{\boldsymbol{\epsilon}}$. Then the least squares estimator of $(\sigma_e^2, \sigma_a^2, \ldots, \sigma_k^2)$ is given by $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\tilde{\boldsymbol{\epsilon}}^d$. Clearly, we have that

$$(\mathbf{X}^T\mathbf{X}) = \begin{pmatrix} n & n & n & \ldots & n \\ n & n + \boldsymbol{a}^T\boldsymbol{a} & n + \boldsymbol{a}^T\boldsymbol{b} & \ldots & n + \boldsymbol{a}^T\boldsymbol{k} \\ \vdots & & & & \vdots \\ n & n + \boldsymbol{k}^T\boldsymbol{a} & n + \boldsymbol{k}^T\boldsymbol{b} & \ldots & n + \boldsymbol{k}^T\boldsymbol{k} \end{pmatrix}.$$

This is most likely a positive definite matrix as (we assume that) the matrices $\mathbf{A}, \ldots, \mathbf{K}$ are not highly correlated. In addition, we have that

$$\mathbf{X}^T\tilde{\boldsymbol{\epsilon}} = \begin{pmatrix} \sum_{i=1}^n \hat{\epsilon}_i^2 \\ \sum_{i=1}^n \hat{\epsilon}_i^2 + \boldsymbol{a}^T\tilde{\boldsymbol{\epsilon}} \\ \vdots \\ \sum_{i=1}^n \hat{\epsilon}_i^2 + \boldsymbol{k}^T\tilde{\boldsymbol{\epsilon}} \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n \hat{\epsilon}_i^2 \\ \sum_{i=1}^n \hat{\epsilon}_i^2 \\ \vdots \\ \sum_{i=1}^n \hat{\epsilon}_i^2 \end{pmatrix} + \begin{pmatrix} 0 \\ \boldsymbol{a}^T\tilde{\boldsymbol{\epsilon}} \\ \vdots \\ \boldsymbol{k}^T\tilde{\boldsymbol{\epsilon}} \end{pmatrix}.$$

Lemma 1:

$$\left(\mathbf{X}^T\mathbf{X}\right)^{-1} \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} = \begin{pmatrix} \frac{1}{n} \\ 0 \\ \vdots \\ 0 \end{pmatrix} \Leftrightarrow \left(\mathbf{X}^T\mathbf{X}\right) \begin{pmatrix} \frac{1}{n} \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$$

Proof: Because $\left(\mathbf{X}^T\mathbf{X}\right)$is non-singular, and from the properties of the inverse matrix, we have that $\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\left(\mathbf{X}^T\mathbf{X}\right)\boldsymbol{v} = \boldsymbol{v}$ for every $\boldsymbol{v}$. ∎

Lemma 2: *Variance component estimators corresponding to the matrices* $\mathbf{A}, \ldots, \mathbf{K}$ *depend only on the between-observation residuals of the form* $\epsilon_i\epsilon_j$ *for* $i \neq j$ *and do not depend on* $\epsilon_i^2, i = 1, \ldots, n$.

Proof: By noting that

$$(\mathbf{X}^T\mathbf{X})\begin{pmatrix} \frac{1}{n} \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix},$$

we get from Lemma 1 that

$$(\mathbf{X}^T\mathbf{X})^{-1}\begin{pmatrix} \sum_{i=1}^n \hat{\epsilon}_i^2 \\ \sum_{i=1}^n \hat{\epsilon}_i^2 \\ \vdots \\ \sum_{i=1}^n \hat{\epsilon}_i^2 \end{pmatrix} = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n \hat{\epsilon}_i^2 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

which proves that the term $\sum_{i=1}^n \hat{\epsilon}_i^2$ contributes only to the estimator $\widehat{\sigma}_e^2$. ∎

Lemma 3: *Denote by* $\sigma_T^2 = \sigma_e^2 + \sigma_a^2 + \ldots + \sigma_k^2$. *Then* $\widehat{\sigma}_T^2 = \frac{1}{n}\sum_{i=1}^n \hat{\epsilon}_i^2$.

Proof: We show that $\widehat{\sigma}_e^2 + \widehat{\sigma}_a^2 + \ldots + \widehat{\sigma}_k^2 = \frac{1}{n}\sum_{i=1}^n \hat{\epsilon}_i^2$. In the proof of Lemma 2 we saw that

$$(\mathbf{X}^T\mathbf{X})^{-1}\begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} = \begin{pmatrix} \frac{1}{n} \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Since this $(\mathbf{X}^T\mathbf{X})^{-1}$ is symmetric, it follows that

$$\begin{pmatrix} 1 & 1 & \ldots & 1 \end{pmatrix}(\mathbf{X}^T\mathbf{X})^{-1} = \begin{pmatrix} \frac{1}{n} & 0 & \ldots & 0 \end{pmatrix}$$

Therefore

$$\begin{aligned} \widehat{\sigma}_e^2 + \widehat{\sigma}_a^2 + \ldots + \widehat{\sigma}_k^2 &\equiv \begin{pmatrix} 1 & 1 & \ldots & 1 \end{pmatrix}(\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\tilde{\boldsymbol{\epsilon}}) \\ &= \begin{pmatrix} \frac{1}{n} & 0 & \ldots & 0 \end{pmatrix}(\mathbf{X}^T\tilde{\boldsymbol{\epsilon}}) = \frac{1}{n}\sum_{i=1}^n \hat{\epsilon}_i^2, \end{aligned}$$

which completes the proof. ∎

Lemma 4: *An estimator of the ratio between any variance component (or sum of variance components) and the total variance is a ratio between two quadratic forms.*

Proof: For $\boldsymbol{L} = \mathbf{A}, \ldots, \mathbf{K}$, a quantity of the form $\boldsymbol{l}^T\tilde{\boldsymbol{\epsilon}} = \hat{\boldsymbol{\epsilon}}^T\boldsymbol{L}^-\hat{\boldsymbol{\epsilon}}/2$, where the matrix $\boldsymbol{L}^{-1}$ is the matrix $\boldsymbol{L}$ with all diagonal values set to 0. An estimator a variance component σ_l^2 is a linear sum of the quadratic forms $\hat{\boldsymbol{\epsilon}}^T\mathbf{A}^-\hat{\boldsymbol{\epsilon}}, \ldots, \hat{\boldsymbol{\epsilon}}^T\mathbf{K}^-\hat{\boldsymbol{\epsilon}}$, with coefficients the entries of the corresponding row of $(\mathbf{X}^T\mathbf{X})^{-1}$. Since a weighted sum of quadratic forms is a quadratic form, any variance component (and a sum of variance components) is also a quadratic form. Similarly, the total variance estimators is the quadratic form $\frac{1}{n}\hat{\boldsymbol{\epsilon}}^T\hat{\boldsymbol{\epsilon}}$. ∎

Theorem: *We say that two matrices $\boldsymbol{C}_1$ and $\boldsymbol{C}_2$ are orthogonal in the trace inner product, or "trace orthogonal" if $tr(\boldsymbol{C}_1\boldsymbol{C}_2) = 0$. If a matrix $\boldsymbol{L}^-$ is trace orthogonal to all other matrices in the set $\{\mathbf{A}^-, \ldots, \mathbf{K}^-\}$, then*

$$\widehat{\sigma}_l^2 = \frac{1}{\sum_{j>i} l_{i,j}^2} \sum_{j>i} l_{i,j}\hat{\epsilon}_i\hat{\epsilon}_j.$$

Proof: Without loss of generality, assume that $\mathbf{A}$ is trace orthogonal to $\mathbf{B}, \ldots, \mathbf{K}$. First note that for the symmetric matrices with diagonal values set to zero $\mathbf{A}^-, \ldots, \mathbf{K}^-$, $\text{tr}\left(\mathbf{A}^-\boldsymbol{L}^-\right) = 0$ if and only if $\boldsymbol{a}^T\boldsymbol{l} = 0$. Then

$$(\mathbf{X}^T\mathbf{X}) = \begin{pmatrix} n & n & n & \ldots & n \\ n & n+\boldsymbol{a}^T\boldsymbol{a} & n & \ldots & n \\ n & n & n+\boldsymbol{b}^T\boldsymbol{b} & \ldots & n+\boldsymbol{b}^T\boldsymbol{k} \\ \vdots & \vdots & \vdots & & \vdots \\ n & n & n+\boldsymbol{k}^T\boldsymbol{b} & \ldots & n+\boldsymbol{k}^T\boldsymbol{k}. \end{pmatrix}$$

Denote by $(\mathbf{X}^T\mathbf{X})^{-1}_{[i,j]}$ the i,j element in the matrix $(\mathbf{X}^T\mathbf{X})^{-1}$. First, we notice that the entries $(\mathbf{X}^T\mathbf{X})^{-1}_{[2,j]}, j = 3, \ldots, k+1$ are all 0, because the $(\mathbf{X}^T\mathbf{X})^{-1}_{[i,j]}$ entry is a constant times the i,j minor of $(\mathbf{X}^T\mathbf{X})$, which has two identical columns (corresponding to the $1st$ and $2nd$ columns of $(\mathbf{X}^T\mathbf{X})$ when removing its $2nd$ row). Since the sum of the $2nd$ row of $(\mathbf{X}^T\mathbf{X})^{-1}$ is equal to 0, as we saw before, we get that $(\mathbf{X}^T\mathbf{X})^{-1}_{[2,1]} = -(\mathbf{X}^T\mathbf{X})^{-1}_{[2,2]}$.

We now argue that

$$\begin{aligned}
\widehat{\sigma}_a^2 &\equiv (\mathbf{X}^T\mathbf{X})^{-1}_{[2,1]} \sum_{i=1}^n \epsilon_i^2 + (\mathbf{X}^T\mathbf{X})^{-1}_{[2,2]} \left(\sum_{i=1}^n \epsilon_i^2 + \boldsymbol{a}^T \tilde{\boldsymbol{\epsilon}} \right) \\
&= (\mathbf{X}^T\mathbf{X})^{-1}_{[2,1]} \sum_{i=1}^n \epsilon_i^2 - (\mathbf{X}^T\mathbf{X})^{-1}_{[2,1]} \left(\sum_{i=1}^n \epsilon_i^2 + \boldsymbol{a}^T \tilde{\boldsymbol{\epsilon}} \right) \\
&= -(\mathbf{X}^T\mathbf{X})^{-1}_{[2,1]} \boldsymbol{a}^T \tilde{\boldsymbol{\epsilon}} \stackrel{(4)}{=} \frac{1}{\sum_{j>i} a_{i,j}^2} \sum_{j>i} a_{i,j} \epsilon_i \epsilon_j,
\end{aligned}$$

where we need to show that equality (4) holds to complete the proof. We need to show that $-(\mathbf{X}^T\mathbf{X})^{-1}_{[2,1]} = \frac{1}{\sum_{j>i} a_{i,j}^2}$. Consider now the matrix $\mathbf{X}^T\mathbf{X}$. One can derive its determinant from its second row, as:

$$\begin{aligned}
|\mathbf{X}^T\mathbf{X}| &= (\mathbf{X}^T\mathbf{X})_{[2,1]} M_{2,1} - (\mathbf{X}^T\mathbf{X})_{[2,2]} M_{2,2} + \ldots + (-1)^{k+1} (\mathbf{X}^T\mathbf{X})_{[2,1k]} M_{2,k} \\
&= n M_{2,1} - (n + \boldsymbol{a}^T\boldsymbol{a}) M_{2,2} + 0 + \ldots 0 \\
&= n|\mathbf{X}^T\mathbf{X}|(\mathbf{X}^T\mathbf{X})^{-1}_{[2,1]} - (n + \boldsymbol{a}^T\boldsymbol{a})|\mathbf{X}^T\mathbf{X}|(\mathbf{X}^T\mathbf{X})^{-1}_{[2,2]} \\
&= n|\mathbf{X}^T\mathbf{X}|(\mathbf{X}^T\mathbf{X})^{-1}_{[2,1]} - (n + \boldsymbol{a}^T\boldsymbol{a})|\mathbf{X}^T\mathbf{X}|(\mathbf{X}^T\mathbf{X})^{-1}_{[2,1]} \\
&= -\boldsymbol{a}^T\boldsymbol{a}|\mathbf{X}^T\mathbf{X}|(\mathbf{X}^T\mathbf{X})^{-1}_{[2,1]}
\end{aligned}$$

Therefore, we get that $-(\mathbf{X}^T\mathbf{X})^{-1}_{[2,1]} = \frac{1}{\boldsymbol{a}^T\boldsymbol{a}}$, which completes the proof. ∎

2 Computation

2.1 Variance component estimators

While any unbiased estimator of $\widehat{\boldsymbol{\beta}}$ suffices to generate residuals $\hat{\boldsymbol{\epsilon}}$ and use them to obtain variance component estimators as $\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\tilde{\boldsymbol{\epsilon}}^d$, a more efficient estimator iterates between estimating $\boldsymbol{\beta}$ and the variance component estimator as follows:

1. Initialization step: set $\widehat{\boldsymbol{\beta}}^{(0)} = (\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T\boldsymbol{y}$.

2. Iteration step:

   (a) Given the kth estimator of $\boldsymbol{\beta}$, $\widehat{\boldsymbol{\beta}}^{(k)}$, set $\hat{\boldsymbol{\epsilon}} = \boldsymbol{y} - \mathbf{W}\boldsymbol{\beta}^{(k)}$ and $\tilde{\boldsymbol{\epsilon}}$ is the vector of uppder diagonal matrix (including the diagonal) of $\hat{\boldsymbol{\epsilon}}\hat{\boldsymbol{\epsilon}}^T$. Set $\widehat{\boldsymbol{\sigma}}^{2,(k)} = (\widehat{\sigma}_e^{2,(k)}, \widehat{\sigma}_a^{2,(k)}, \ldots, \widehat{\sigma}_k^{2,(k)}) =$

$\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\tilde{\boldsymbol{\epsilon}}^d$.

(b) Given the kth estimator of $\boldsymbol{\sigma}^2$, $\widehat{\boldsymbol{\sigma}}^{2,(k)}$, let $\widehat{\boldsymbol{\Sigma}}^{(k)} = \widehat{\sigma}_e^{2,(k)}\mathbf{I}_{n\times n} + \widehat{\sigma}_a^{2,(k)}\mathbf{A} + \ldots + \widehat{\sigma}_k^{2,(k)}\mathbf{K}$ with inverse $\widehat{\boldsymbol{\Sigma}}^{-1,(k)}$. Set $\widehat{\boldsymbol{\beta}}^{(k+1)} = (\mathbf{W}^T\widehat{\boldsymbol{\Sigma}}^{-1,(k)}\mathbf{W})^{-1}\mathbf{W}^T\widehat{\boldsymbol{\Sigma}}^{-1,(k)}\boldsymbol{y}$

The iteration step repeats until convergence.

2.2 Confidence intervals for the variance components

From Lemma 4, any variance components (or sum of variance components) is given as a quadratic form. Let $\boldsymbol{Q}$ be the quadratic form corresponding to a variance component estimate $\widehat{\sigma}_l^2$, such that $\widehat{\sigma}_l^2 = \hat{\boldsymbol{\epsilon}}^T\boldsymbol{Q}\hat{\boldsymbol{\epsilon}}$. Then this $\widehat{\sigma}_l^2$ is distributed as the sum of independent $\chi^2_{(1)}$ variables in $\sum_{i=1}^n \lambda_i\chi^2_{(1)}$, where $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of $\boldsymbol{Q}\text{cov}(\hat{\boldsymbol{\epsilon}})$. In practice, for $\text{cov}(\hat{\boldsymbol{\epsilon}})$ we use used the estimated $\widehat{\Sigma}(\widehat{\sigma}_e^2, \ldots, \widehat{\sigma}_k^2)$. Functions in the package CompQuadFrom calculate the probability function (or survival function) of this quadratic form based on $\lambda_1, \ldots, \lambda_n$. While it takes times to compute the eigenvalues, once they are computed, a calculating the probabilities associated with the quadratic form over a grid is simple and quick. We can test the hypothesis $H_0 : \sigma_l^2 = 0$ by calculating the probability

$$\Pr\left(\hat{\boldsymbol{\epsilon}}^T\boldsymbol{Q}\hat{\boldsymbol{\epsilon}} = 0\right) = 1 - \Pr\left(\hat{\boldsymbol{\epsilon}}^T\boldsymbol{Q}\hat{\boldsymbol{\epsilon}} > 0\right),$$

and calculate two-sided confidence intervals for $\widehat{\sigma}_l^2$ by calculating the survival probabilities over a grid, and taking the appropriate quantiles. For example, for a 95% confidence interval we take the values (c_1, c_2) for which

$$\begin{aligned} c_1 &= u : \Pr\left(\boldsymbol{\epsilon}^T\boldsymbol{Q}\boldsymbol{\epsilon} > u\right) = 0.025 \\ c_2 &= u : \Pr\left(\boldsymbol{\epsilon}^T\boldsymbol{Q}\boldsymbol{\epsilon} > u\right) = 0.975. \end{aligned}$$

We find these values using a binary search on the interval $[0, \widehat{\sigma}_T^2]$.

2.3 Computing heritability estimates and their confidence intervals

Suppose that the variance component corresponding to the kinship matrix is σ_k^2, which quadratic form denoted by $\boldsymbol{Q}_k$. We estimate heritability as $\widehat{h}_k = \widehat{\sigma}_k^2/\widehat{\sigma}_T^2$. However, we cannot use the confidence intervals for σ_k^2 to construct confidence intervals for h_k. Instead, we note that the point

estimate is $\widehat{h}_k$ is given by:

$$\widehat{h}_k = \frac{\hat{\boldsymbol{\epsilon}}^T \boldsymbol{Q}_k \hat{\boldsymbol{\epsilon}}}{\frac{1}{n}\hat{\boldsymbol{\epsilon}}^T \mathbf{I} \hat{\boldsymbol{\epsilon}}} = \frac{\boldsymbol{x}^T \widehat{\boldsymbol{\Sigma}}^{1/2} \boldsymbol{Q}_k \widehat{\boldsymbol{\Sigma}}^{1/2} \boldsymbol{x}}{\frac{1}{n}\boldsymbol{x}^T \widehat{\boldsymbol{\Sigma}} \boldsymbol{x}} = \frac{\boldsymbol{x}^T \boldsymbol{F} \boldsymbol{x}}{\boldsymbol{x}^T \mathbf{G} \boldsymbol{x}}$$

where $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, for $\boldsymbol{F} = \widehat{\boldsymbol{\Sigma}}^{1/2} \boldsymbol{Q}_k \widehat{\boldsymbol{\Sigma}}^{1/2}$ and $\mathbf{G} = \widehat{\boldsymbol{\Sigma}}/n$. Thus, it is a ratio between two quadratic forms in (what we assume are) normal variables. For the squared root $\widehat{\boldsymbol{\Sigma}}^{1/2}$, we use the Cholesky decomposition of $\widehat{\boldsymbol{\Sigma}}$.

Now, we use the saddlepoint approximation for the distribution of a ratio of quadratic forms in normal variables, proposed by Lieberman (1994). For a given potential value of h_k, say h_k^*, we can calculate the survival probability

$$Pr\left(h_k \geq h_k^*\right) \cong 1 - \Phi(\hat{\xi}) + \phi(\hat{\xi}) \left[\frac{1}{\hat{z}} - \frac{1}{\hat{\xi}}\right]$$

where Φ and ϕ are the standard normal cdf and pdf, and

$$\begin{aligned} \hat{z} &= \hat{\omega} \left\{ 2 \sum_{i=1}^{n} \frac{d_i^{*2}}{(1 - 2\hat{\omega} d_i^*)^2} \right\}^{1/2} \\ \hat{\xi} &= \left\{ \sum_{i=1}^{n} \ln(1 - 2\hat{\omega} d_i^*) \right\}^{1/2} \operatorname{sgn}(\hat{\omega}) \end{aligned}$$

and $d_1^*, \ldots, d_n^*$ are the eigenvalues of the matrix $\mathbf{D}^* = \boldsymbol{F} - h_k^* \mathbf{G}$, and $\hat{\omega}$ is the corresponding saddlepoint satisfying

$$\sum_{i=1}^{n} \frac{d_i^*}{1 - 2\hat{\omega} d_i^*} = 0.$$

Confidence intervals are then built, as before, using a binary search to find the values satisfying the required probabilities at the tails.

2.3.1 A faster algorithm when the kinship matrix is the only source of correlation in the model

Computing the eigenvalues $d_1^*(h_k^*), \ldots, d_n^*(h_k^*)$ takes time. However, in the case where we only have a single kinship matrix, denoted by $\mathbf{K}$ we can compute the eigen decomposition of the matrix $\mathbf{K}^-$ once to obtain eigenvalues $\lambda_1, \ldots, \lambda_n$, and then transform these eigenvalues to obtain the

eigenvalues $d_1^*(h_k^*), \ldots, d_n^*(h_k^*)$ for each value h_k^*. To see this, suppose that $\boldsymbol{u}$ is an eigenvector of $\mathbf{K}^-$ with eigenvalues λ. Then, by definition:

$$\mathbf{K}^- \boldsymbol{u} = \lambda \boldsymbol{u}.$$

Since $\boldsymbol{\Sigma} = \sigma_k^2(\mathbf{K}^- + \mathbf{I}) + \sigma_e^2\mathbf{I}$, it is straightforward to see that $\boldsymbol{u}$ is also an eigenvector of $\boldsymbol{\Sigma}$:

$$\boldsymbol{\Sigma}\boldsymbol{u} = \left[\sigma_k^2(\mathbf{K}^- + \mathbf{I}) + \sigma_e^2\mathbf{I}\right]\boldsymbol{u} = \left(\sigma_k^2\lambda + \sigma_k^2 + \sigma_e^2\right)\boldsymbol{u}.$$

Similarly, $\boldsymbol{u}$ is an eigenvector of $\boldsymbol{\Sigma}^{1/2}$ with eigenvalue $\sqrt{\sigma_k^2\lambda + \sigma_k^2 + \sigma_e^2}$, which finally leads us to the transformation between an eigenvalue λ of $\boldsymbol{\Lambda}$ to an eigenvalue of $\mathbf{D}^* = \boldsymbol{F} - h_k^*\mathbf{G}$ given by:

$$d_i^*(h_k^*, \lambda_i) = \frac{1}{2\sum_{i<j} v_{ij}^2}\lambda_i(\lambda_i\sigma_k^2 + \sigma_k^2 + \sigma_e^2) - h_k^*(\lambda_i\sigma_k^2 + \sigma_k^2 + \sigma_e^2)/n.$$

As before, we use the estimated $\widehat{\sigma}_k^2, \widehat{\sigma}_e^2$ instead of the true unknown quantities.

2.3.2 Meta-analysis of across studies when kinship is the only source of correlation

A meta-analytic estimator. Suppose that there are S studies that we wanted to combined in meta-analysis. We assume that kinship is the only source of correlation. Each study has a vector of residuals $\hat{\boldsymbol{\epsilon}}_s = (\hat{\epsilon}_{s,1}, \ldots, \hat{\epsilon}_{s,n_s})^T, s = 1, \ldots, S$. Consider the Haseman-Elston regression, but incomplete, so that only the pairs of multiplied residuals within study are used (i.e. only $\hat{\epsilon}_{s,i}\hat{\epsilon}_{s,j}$ are regressed against entries of the kinship covariance matrix, but not $\hat{\epsilon}_{s,i}\hat{\epsilon}_{t,j}$). Therefore, cross-study kinship estimates are not used in the regression, however no assumption is made on them. In other words, we do not need to assume that participant in one study is genetically independent (no alleles shared IBD) of a participant in another study. It is straightforward to show that the meta-analytic estimator of σ_2^e is given by $\widehat{\sigma}_e^2 = \sum_{s=1}^{S}\sum_{i=1}^{n_s}\hat{\epsilon}_{s,i}^2$. Let $\hat{\boldsymbol{\epsilon}} = (\hat{\boldsymbol{\epsilon}}_1^T, \ldots, \hat{\boldsymbol{\epsilon}}_S)^T$. Then the meta-analysis kinship variance component estimator is given by

$$\widehat{\sigma}_k^2 = \frac{1}{\operatorname{tr}\left(\mathbf{K}_s^-\mathbf{K}_s^-\right)}\hat{\boldsymbol{\epsilon}}^T\mathbf{K}_s^-\hat{\boldsymbol{\epsilon}}$$

where $\mathbf{K}_s^-$ is the block diagonal matrix that have all the study-specific kinship matrix (without their diagonal values) arranged diagonally, as

$$\mathbf{K}^s = \begin{pmatrix} \mathbf{K}_1^- & \mathbf{0} & \dots & \dots & \\ \mathbf{0} & \mathbf{K}_2^- & & & \mathbf{0} \\ \vdots & & \ddots & & \\ \vdots & \mathbf{0} & \mathbf{0} & & \mathbf{K}_S^- \end{pmatrix}$$

To see that this meta-analytic estimator of σ_k^2 is unbiased, note first that $\text{cov}(\hat{\boldsymbol{\epsilon}}) = (\sigma_e^2+\sigma_k^2)\mathbf{I}+\sigma_k^2\mathbf{K}^-$, where now $\mathbf{K}^-$ is the kinship matrix with kinship coefficients between the individuals across studies. Now, from characteristics of quadratic forms, we have that

$$\begin{aligned} E\left[\widehat{\sigma}_k^2\right] &= E\left[\frac{1}{\text{tr}\left(\mathbf{K}_s^-\mathbf{K}_s^-\right)}\hat{\boldsymbol{\epsilon}}^T\mathbf{K}_s^-\hat{\boldsymbol{\epsilon}}\right] = \frac{1}{\text{tr}\left(\mathbf{K}_s^-\mathbf{K}_s^-\right)}\text{tr}\left(\mathbf{K}_s^-\text{cov}(\hat{\boldsymbol{\epsilon}})\right) \\ &= \frac{1}{\text{tr}\left(\mathbf{K}_s^-\mathbf{K}_s^-\right)}\text{tr}\left(\mathbf{K}_s^-(\sigma_e^2+\sigma_k^2)\mathbf{I}+\sigma_k^2\mathbf{K}^-\right) \\ &= \frac{1}{\text{tr}\left(\mathbf{K}_s^-\mathbf{K}_s^-\right)}\text{tr}\left(\mathbf{K}_s^-\sigma_k^2\mathbf{K}_s^-\right) = \sigma_k^2. \end{aligned}$$

Computing the meta-analytics heritability estimator and confidence intervals. The eigenvalues result shows that all we need to calculate heritability estimates and confidence intervals are eigenvalues of the matrix $\mathbf{K}^-$ (the kinship matrix without the diagonal), estimated σ_e^2, σ_k^2, and the sum of the entries of $\mathbf{K}^-$ ($2\sum_{i<j}k_{ij}^2$). This result could be used to extend our methods to meta-analysis of information from multiple studies. Suppose that each of m independent studies calculated the residuals from a "null model" (i.e. a regression model without genetic fixed effects other than PCs). Then, each study s reports:

1. $\mathcal{K}^s = 2\sum_{i<j}k_{ij}^2$,
2. $\widehat{\sigma}_{k,s}^2$,
3. $\widehat{\sigma}_{e,s}^2$,
4. The number of participants in the study n_s,
5. The eigenvalues $\lambda_1^s, \dots, \lambda_{n_s}^s$ of the matrix $\mathbf{K}_s^-$.

Then, the meta-analysis estimates of the kinship and error variance components, and $\mathcal{K}^S$ are given by:

$$\begin{aligned}
\widehat{\sigma}_k^2 &= \frac{\sum_{s=1}^{S} \mathcal{K}^s \widehat{\sigma}_{k,s}^2}{\sum_{s=1}^{S} \mathcal{K}^s} \\
\widehat{\sigma}_e^2 &= \frac{\sum_{s=1}^{S} n_s \widehat{\sigma}_{e,s}^2}{\sum_{s=1}^{S} n_s}, \\
\mathcal{K}^S &= \sum_{s=1}^{S} \mathcal{K}^s,
\end{aligned}$$

and the eigenvalues of the cross-study $\mathbf{K}^-$ matrix are taken to be $\lambda_1^1, \ldots, \lambda_{n_1}^1, \ldots, \lambda_1^S, \ldots, \lambda_{n_S}^S$. Using these, the central location that can calculate heritability estimates and confidence intervals.

3 The Hispanic Community Health Study/Study of Latinos

The HCHS/SOL, (LaVange et al., 2010; Sorlie et al., 2010)) is a community based cohort study, following self-identified Hispanic individuals from four field centers (Chicago, IL; Miami, FL; Bronx, NY; and San Diego, CA). Individuals were sampled via a two-stage sampling scheme, in which households were randomly sampled from sampled community block units. Almost 13,000 study participants consented for genotyping. HCHS/SOL individuals are classified into 'genetic analysis groups', classes that are based on self reported ethnicities and genetic similarity (Conomos et al., 2016). The genetic analysis groups are Central American, Cuban, Dominican, Mexican, Puerto Rican, and South American. This study was approved by the institutional review boards at each field center, where all subjects gave written informed consent.

3.1 Genotyping, imputation and quality control

Blood samples from HCHS/SOL individuals were genotyped on a custom array consisting of Illumina Omni 2.5M content plus ∼150,000 custom markers selected to include ancestry-informative markers, variants characteristic of Amerindian populations, known GWAS hits and other candidate gene polymorphisms. Quality control was similar to the procedure described in Laurie et al. (2010) and included checks for sample identity, batch effects, missing call rate, chromosomal anomalies (Laurie et al., 2012), deviation from Hardy-Weinberg equilibrium, Mendelian errors, and duplicate

sample discordance. 12,803 samples passed quality control, and 2,232,944 SNPs passed quality filters. Pairwise kinship coefficients and principal components reflecting ancestry were estimated in an iterative procedure which accounts for admixture (Conomos et al., 2016). All common variants were used to estimate kinship coefficients.

3.2 Heritability estimation in the HCHS/SOL

In each group of interest, including all individuals ('pooled' analysis), or specific genetic analysis groups, we randomly removed related individuals, to generate a set of individuals without any pair having kinship coefficient higher than 2^{-11}. Due to the sampling structure of the HCHS/SOL, the correlation between individuals is modeled in a kinship matrix, and also via matrices corresponding to community block units, and households. We estimated variance components via the procedure described here, with the three correlation matrices. We utilized the availability of environmental correlation to also estimate the contribution of modeled environmental factors (block unit and household) to the phenotypic variance. Finally, we also demonstrate the use of our method for meta-analysis by removing individuals from shared household to generate a restricted set in which none of the individuals live in the same house, and used the proposed procedure for calculating heritability in meta-analysis. Note that for this purpose we neglected block unit correlation and assume that there is no correlation due to block unit sharing.

We estimated heritability for the FEV1 (a measure of lung function), standing height, depression score (CESD10, a sum of ten questionnaire items related to depression in the past few weeks of filling the form), SBP (systolic blood pressure), and dental caries, a count of tooth decays and cavities across all teeth of a participant. Finally, all regression models were adjusted (via the design matrix $\mathbf{W}$) to the 5 first principal components, study center, age, sex, and genetic analysis group (in the pooled models). For some traits we used additional covariates. Table 1 provides the various estimates and confidence intervals.

Analysis	n	Height	Depression score	SBP	Dental caries	FEV1
Full: Including environmentally correlated individuals						
Heritability	10,255	0.58 (0.47,0.69)	0.04 (0.00,0.10)	0.20 (0.12,0.27)	0.17 (0.09,0.25)	0.25 (0.17,0.33)
Environmental variance	10,255	0.10 (0.06,0.14)	0.06 (0.02,0.10)	0.02 (0.00,0.06)	0.06 (0.02,0.10)	0.04 (0.00,0.09)
Restricted: without environmentally correlated individuals (heritability only)						
Pooled	7,848	0.57 (0.45,0.69)	0.03 (0.00,0.11)	0.19 (0.11,0.28)	0.19 (0.09,0.28)	0.25 (0.16,0.34)
Central American	867	0.26 (0.00,0.88)	0.00 (0.00,0.62)	0.00 (0.00,0.56)	0.17 (0.00,0.75)	0.16 (0.00,0.75)
South American	544	0.48 (0.00,1.00)	0.00 (0.00,0.88)	0.93 (0.06,1.00)	0.00 (0.00,0.88)	0.57 (0.00,1.00)
Mexican	2,862	0.48 (0.28,0.72)	0.04 (0.00,0.22)	0.20 (0.02,0.41)	0.24 (0.06,0.44)	0.27 (0.06,0.50)
Puerto Rican	1,479	0.47 (0.12,0.88)	0.00 (0.00,0.38)	0.27 (0.00,0.62)	0.25 (0.00,0.62)	0.49 (0.12,0.88)
Cuban	1,370	0.83 (0.31,1.00)	0.00 (0.00,0.50)	0.23 (0.00,0.75)	0.06 (0.00,0.62)	0.09 (0.00,0.62)
Dominican	726	0.84 (0.06,1.00)	0.02 (0.00,0.75)	0.71 (0.00,1.00)	0.00 (0.00,0.75)	0.61 (0.00,1.00)
Meta-analysis	7,848	0.53 (0.38,0.72)	0.02 (0.00,0.16)	0.24 (0.09,0.41)	0.19 (0.03,0.34)	0.30 (0.16,0.47)

Table 1: Comparison of heritability estimates obtained in analysis of various subgroups of the HCHS/SOL. The top and bottom parts consider data sets with and without environmentally correlated individuals. When environmental correlation was present, the analysis included all participants. The heritability is $\widehat{\sigma}_k^2/\widehat{\sigma}_T^2$, and the environmental variance proportion is $(\widehat{\sigma}_h^2+\widehat{\sigma}_c^2)/\widehat{\sigma}_T^2$, for σ_h^2, σ_c^2 variance component corresponding to household and community sharing matrices. The models without environmentally correlated individuals had smaller sample sizes (because individuals who shared household were removed), and we compared heritability estimates from the pooled analysis of all individuals, the various ethnic subgroups, and their meta-analysis. SBP is systolic blood pressure. Dental caries is a measure of teeth damage. FEV1 is a measure of lung function. The distribution of the ratio between the appropriate quadratic forms was approximation using the saddlepoint approximation of Lieberman (1994).

4 Additional simulations

We performed additional simulations in simple settings in which we expected that confidence intervals that are based on asymptotic normality will be limited: when the kinship values are small, and when the kinship and household matrices are somewhat "correlated". For the following two settings, we ran 1,000 simulations for each sample size of $n = 1,500,\ \ 3,000,\ \ 5,000$, and $12,784$, with the largest value selected to match the largest HCHS/SOL sample size used. In both settings, the kinship and household matrices were block diagonal with 3×3 matrices, representing sets of three individuals who live in the same house. Therefore, the block corresponding to three people living in the same house in the household matrix was a 3×3 with all values equal to 1. The matrices describing the correlations for the entire simulated sample were thus:

$$\mathbf{H} = \begin{pmatrix} \mathbf{H}_h & \mathbf{0} & \mathbf{0} & \cdots \\ \mathbf{0} & \mathbf{H}_h & \mathbf{0} & \cdots \\ \mathbf{0} & \mathbf{0} & \mathbf{H}_h & \\ \vdots & & & \ddots \end{pmatrix}, \ \mathbf{K} = \begin{pmatrix} \mathbf{K}_h & \mathbf{0} & \mathbf{0} & \cdots \\ \mathbf{0} & \mathbf{K}_h & \mathbf{0} & \cdots \\ \mathbf{0} & \mathbf{0} & \mathbf{K}_h & \\ \vdots & & & \ddots \end{pmatrix},$$

with

$$\mathbf{H}_h = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}$$

in the two settings, and

$$\mathbf{K}_h^1 = \begin{pmatrix} 1 & 0.4 & 0.5 \\ 0.4 & 1 & 0.6 \\ 0.5 & 0.6 & 1 \end{pmatrix}$$

in the "correlated kinship and household matrices" settings 1, and

$$\mathbf{K}_h^2 = \begin{pmatrix} 1 & 0.05 & 0.05 \\ 0.05 & 1 & 0.1 \\ 0.05 & 0.1 & 1 \end{pmatrix}$$

in the "small kinship values" settings 2. In both settings the residuals were sample to have $\boldsymbol{\epsilon} \sim \mathcal{N}\left(\mathbf{0}, \sigma_e^2\mathbf{I} + \sigma_k^2\mathbf{K} + \sigma_h^2\mathbf{H}\right)$, with $\sigma_e^2 = 100, \sigma_k^2 = 40$, and $\sigma_h^2 = 15$. To simulate $\boldsymbol{\epsilon}$ with this distribution we computed the squared root of the matrix $\boldsymbol{\Sigma} = \sigma_e^2\mathbf{I} + \sigma_k^2\mathbf{K} + \sigma_h^2\mathbf{H}$. Then, we used this $\boldsymbol{\Sigma}^{1/2}$ by having $\boldsymbol{\epsilon} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\epsilon}_{iid}$, where $\boldsymbol{\epsilon}_{iid} \sim \mathcal{N}(0, \mathbf{I})$, which is straight forward to sample.

4.1 Results

Tables 2 and 3 provide the results comparing the proposed HE-based method for calculating confidence intervals with the AI-REML method implemented in the GENESIS R package that uses normal approximation for the asymptotic distribution of the variance components for settings 1 and 2 respectively. The performance measures are the same as in the main manuscript: coverage (the proportion of simulations the true value was contained in the confidence interval), width (average width of the confidence interval), and root-mean-square-error comparing the estimated variance component to its true value across simulations.

One can see that the performance of both methods improve with the sample size, but HE confidence intervals always have coverage at least 0.95, and tending to be larger when the sample sizes are smaller, while the REML (GENESIS) confidence intervals often have poor coverage for small sample size and/or small values of the correlation matrices. The household matrix is the same in the two settings. However, in the first settings, the household and kinship matrices are more similar, and this leads to higher uncertainly - wider confidence intervals and larger RMSEs - of the corresponding variance component. Interestingly, the coverage of the household variance component of GENESIS remains poor in setting 2 even in large sample sizes, despite similar width and RMSE to the HE. This is because the estimated kinship variance component was in fact 0 in many of the simulations. The mean width of the confidence intervals tend to be larger for HE (which also have more coverage) and becomes almost the same as that of GENESIS (REML) as the sample size becomes large.

		Kinship			Household		
n	method	coverage	width	RMSE	coverage	width	RMSE
1500	HE	1.00	0.79	0.21	1.00	0.39	0.10
1500	REML - GENESIS	0.61	0.73	0.18	0.59	0.31	0.09
3000	HE	1.00	0.69	0.19	1.00	0.33	0.09
3000	GENESIS	0.70	0.64	0.16	0.69	0.28	0.08
5000	HE	1.00	0.66	0.18	1.00	0.32	0.09
5000	GENESIS	0.73	0.61	0.16	0.72	0.27	0.08
12784	HE	0.97	0.37	0.09	0.96	0.17	0.05
12784	GENESIS	0.95	0.36	0.09	0.95	0.17	0.05

Table 2: Results from simulation setting 1, with $\mathbf{K}_h^1$.

		Kinship			Household		
n	method	coverage	width	RMSE	coverage	width	RMSE
1500	HE	1.00	1.00	0.44	1.00	0.21	0.05
1500	GENESIS	0.38	0.80	0.43	0.89	0.15	0.04
3000	HE	1.00	1.00	0.42	1.00	0.17	0.04
3000	GENESIS	0.46	0.83	0.41	0.83	0.13	0.04
5000	HE	1.00	1.00	0.37	0.96	0.14	0.03
5000	GENESIS	0.56	0.89	0.36	0.84	0.11	0.03
12784	HE	1.00	0.84	0.27	0.97	0.09	0.02
12784	GENESIS	0.76	0.84	0.27	0.88	0.08	0.02

Table 3: Results from simulation setting 2, with $\mathbf{K}_h^2$.

References

Conomos, M. P., Laurie, C. A., Stilp, A. M., Gogarten, S. M., McHugh, C. P., Nelson, S. C., Sofer, T., Fernández-Rhodes, L., Justice, A. E., Graff, M., Young, K. L., Seyerle, A., Avery, C., Taylor, K., Rotter, J., Talavera, G., Daviglus, M., Wassertheil-Smoller, S., Schneiderman, N., Heiss, G., Kaplan, R., Franceschini, N., Reiner, A., Shaffer, G., JohnR and Barr, Kerr, K., Browning, S., Browning, B., Weir, B., Avilés-Santa, L., Papanicolaou, G., Lumley, T., Szpiro, A., North, K., Rice, K., Thornton, T. and Laurie, C. (2016). Genetic Diversity and Association Studies in US Hispanic/Latino Populations: Applications in the Hispanic Community Health Study/Study of Latinos. *The American Journal of Human Genetics*, **98** 165–184.

Laurie, C. et al. (2010). Quality control and quality assurance in genotypic data for genome-wide association studies. *Genetic Epidemiology*, **34** 591–602.

Laurie, C. C., Laurie, C. A., Rice, K., Doheny, K. F., Zelnick, L. R., McHugh, C. P., Ling, H., Hetrick, K. N., Pugh, E. W., Amos, C. et al. (2012). Detectable clonal mosaicism from birth to old age and its relationship to cancer. *Nature Genetics*, **44** 642–650.

LaVange, L. M., Kalsbeek, W. D., Sorlie, P. D., Avilés-Santa, L. M., Kaplan, R. C., Barnhart, J., Liu, K., Giachello, A., Lee, D. J., Ryan, J. et al. (2010). Sample design and cohort selection in the hispanic community health study/study of latinos. *Annals of epidemiology*, **20** 642–649.

Lieberman, O. (1994). Saddlepoint approximation for the distribution of a ratio of quadratic forms in normal variables. *Journal of the American Statistical Association*, **89** 924–928.

Sorlie, P. D., Avilés-Santa, L. M., Wassertheil-Smoller, S., Kaplan, R. C., Daviglus, M. L., Giachello, A. L., Schneiderman, N., Raij, L., Talavera, G., Allison, M., LaVange, L., Chambless, L. E. and Heiss, G. (2010). Design and implementation of the hispanic community health study/study of latinos. *Annals of epidemiology*, **20** 629–641.