Bioinformatics, 34, 2018, i187–i194
doi: 10.1093/bioinformatics/bty253
ISMB 2018

OXFORD

A scalable estimator of SNP heritability for biobank-scale data

Yue Wu[1] and Sriram Sankararaman[1,2,*]

[1]Department of Computer Science and [2]Department of Human Genetics, University of California, Los Angeles, CA 90095, USA

*To whom correspondence should be addressed.

Abstract

Motivation: Heritability, the proportion of variation in a trait that can be explained by genetic variation, is an important parameter in efforts to understand the genetic architecture of complex phenotypes as well as in the design and interpretation of genome-wide association studies. Attempts to understand the heritability of complex phenotypes attributable to genome-wide single nucleotide polymorphism (SNP) variation data has motivated the analysis of large datasets as well as the development of sophisticated tools to estimate heritability in these datasets. Linear mixed models (LMMs) have emerged as a key tool for heritability estimation where the parameters of the LMMs, i.e. the variance components, are related to the heritability attributable to the SNPs analyzed. Likelihood-based inference in LMMs, however, poses serious computational burdens.

Results: We propose a scalable randomized algorithm for estimating variance components in LMMs. Our method is based on a method-of-moment estimator that has a runtime complexity $\mathcal{O}(NMB)$ for N individuals and M SNPs (where B is a parameter that controls the number of random matrix-vector multiplications). Further, by leveraging the structure of the genotype matrix, we can reduce the time complexity to $\mathcal{O}\left(\frac{NMB}{\max(\log_3 N, \log_3 M)}\right)$.

We demonstrate the scalability and accuracy of our method on simulated as well as on empirical data. On standard hardware, our method computes heritability on a dataset of 500 000 individuals and 100 000 SNPs in 38 min.

Availability and implementation: The RHE-reg software is made freely available to the research community at: https://github.com/sriramlab/RHE-reg.

Contact: sriram@cs.ucla.edu

1 Introduction

A central question in biology is to understand how much of the variation in a trait (phenotype) can be explained by genetics as opposed to environmental factors. The heritability of a trait is a central notion in quantifying the contribution of genetics to the variation in a trait. The heritability of a trait refers to the proportion of variation in the trait that can be explained by genetic variation (Visscher *et al.*, 2008). The narrow-sense heritability (h^2) refers to the proportion of trait variation that can be explained by a linear function of genetic variation (Almasy and Blangero, 1998). Beyond understanding the genetic basis of a phenotype, heritability determines the power of genetic association studies to detect genetic variants associated with a phenotype, the accuracy of using genetic data to predict phenotypes, as well as the response of a phenotype to natural and artificial selection (Houle, 1992).

While family-based studies enabled the estimation of heritability of a wide variety of traits, the availability of genome-wide genetic variation data has enabled a direct estimation of the heritability associated with genotyped single nucleotide polymorphisms (SNPs), termed *SNP heritability*. Initial attempts to estimate heritability from genomic data focused on the variation in a trait could be explained by SNPs that were discovered to be significantly associated with the trait in a genome-wide association study (GWAS). These estimates were found to severely under-estimate the narrow-sense heritability, a phenomenon known as *missing heritability*. A major insight into the mystery of missing heritability emerged in Yang *et al.* (2010) who showed that using all genotyped SNPs jointly to explain variation in a trait led to a substantially larger estimate of heritability than from SNPs that were found to be associated in GWAS. Subsequent analyses suggest that much of missing heritability could be explained by the presence of a large number of SNPs of weak effects that has, in turn, motivated analyses of larger datasets.

Linear mixed models (LMMs) has emerged as a key analytically technique for estimating the heritability of complex traits using

genome-wide SNP variation data. Beyond their application in estimating SNP heritability, LMMs are widely used in association tests where they are used to control for population stratification (Kang *et al.*, 2008a; Lippert *et al.*, 2011; Loh *et al.*, 2015b; Yu *et al.*, 2006; Zhou and Stephens, 2014), in phenotype and disease risk prediction (Makowsky *et al.*, 2011; Speed *et al.*, 2012; Wray *et al.*, 2013; Yang *et al.*, 2010; Zhou *et al.*, 2013), and in understanding the relative contribution of genomic regions to variation in a trait of interest (Makowsky *et al.*, 2011; Wray *et al.*, 2013; Yang *et al.*, 2010). A key step in the application of LMMs is the estimation of their parameters, i.e. often referred to as variance components. Estimation of variance components is a computationally challenging problem on genomic datasets containing large numbers of individuals and SNPs. The most commonly used method for variance components estimation in LMMs relies on maximizing the likelihood of the parameters. Often, a related estimator, known as the restricted maximum likelihood (REML) estimator, is preferred due to a reduced bias relative to maximum likelihood estimators. Both maximum likelihood as well as REML estimation, however, rely on computationally intensive optimization problems. While a number of methods have been proposed to improve the computational efficiency of REML estimators (Kang *et al.*, 2008b; Lippert *et al.*, 2011; Loh *et al.*, 2015a, b; Pirinen *et al.*, 2013; Yang *et al.*, 2011), all of these methods rely on iterative optimization algorithms that do not scale well to biobank-scale datasets consisting of millions of individuals genotyped at tens of millions of SNPs. Further, REML has been shown to yield biased estimates of heritability in ascertained case-control studies (Chen, 2014; Golan *et al.*, 2014).

1.1 Our contributions

We propose a scalable randomized algorithm to estimate variance components of a LMM. Our method is based on Haseman–Elston (HE) regression (Bulik-Sullivan, 2015; Chen *et al.*, 2004; Elston *et al.*, 2000; Haseman and Elston, 1972), a method-of-moment (MoM) estimator of the heritability of a phenotype. The HE-regression estimator, like other MoM estimators, tends to be statistically less efficient compared to REML. On the other hand, HE-regression is computationally attractive as it leads to a set of linear equations in the variance components that can be solved analytically. While this property of HE-regression is appealing, a key computational bottleneck in the application of HE-regression is the computation of an $N \times N$ matrix that summarizes the relationship between all N pairs of individuals in the dataset. As a result, the computation and memory requirements of HE scale quadratically with the number of individuals.

Our randomized HE-regression (RHE-reg) estimator relies on the observation that the key bottleneck in HE-regression can be replaced by multiplying the $N \times M$ (individuals $\times$ SNPs) matrix of genotypes with a small number, B, of random vectors. This leads to a randomized estimator with runtime $\mathcal{O}(NMB)$ and memory requirements $\mathcal{O}(NM)$. Further, we leverage the observation that the genotype matrix has entries in a finite set, i.e. $\{0,1,2\}$ so that the time complexity of matrix-vector multiplication reduces to $\mathcal{O}\left(\frac{NMB}{\max(\log_3(N),\log_3(M))}\right)$ (Liberty and Zucker, 2009). This additional gain in efficiency can be substantial when the number of SNPs or individuals is large. For example, in the UK Biobank, N is of the order of 10^5 while M is of the order of 10^6. Thus, we propose an estimator of variance components with runtime $\mathcal{O}\left(\frac{NMB}{\max(\log_3(N),\log_3(M))} + NM\right)$ and memory requirement $\mathcal{O}(NM)$.

We apply the RHE-reg estimator to the problem of estimating SNP heritability. We show that our method yields unbiased SNP heritability estimates. While our method is statistically inefficient compared to REML (both because it is moment-based as well as the added randomization), we show in practice that the statistical inefficiency is minimal, particularly for large sample sizes. Further, our method is substantially more computationally efficient so that it can be effectively applied to whole-genome genotype data from hundreds of thousands of individuals. REML has been shown to yield biased estimates of heritability in ascertained case-control studies (Chen *et al.*, 2004; Golan *et al.*, 2014) while the RHE-reg estimator can also be applied in this setting.

Finally, since variance component analysis is of interest beyond heritability estimation, the RHE-reg estimator can enable rapid estimation of variance components in all of the settings in which LMMs are used.

2 Materials and methods

We observe genotypes from N individuals at M SNPs. The genotype vector for individual i is a length M vector denoted by $\boldsymbol{g}_i \in \{0,1,2\}^M$. The jth entry of $\boldsymbol{g}_i$ denotes the number of minor allele carried by individual i at SNP j. Let $\boldsymbol{G}$ be the $N \times M$ genotype matrix where $\boldsymbol{G} = [\boldsymbol{g}_1^T \dots \boldsymbol{g}_N^T]$. $\boldsymbol{X}$ is a $N \times M$ matrix of standardized genotypes obtained by centering and scaling each column of $\boldsymbol{G}$ so that $\sum_n g_{n,m} = 0$ and $\sum_n g_{n,m}^2 = 1$ for all $n \in \{1,\dots,N\}$. Let $\boldsymbol{y}$ is an N-vector of phenotypes and $\boldsymbol{\beta}$ be an M-vector of SNP effect sizes.

2.1 Linear mixed model

We assume the vector of phenotypes $\boldsymbol{y}$ is related to the genotypes by a LMM:

$$\boldsymbol{y}|\epsilon, \boldsymbol{\beta} = \boldsymbol{X}\boldsymbol{\beta} + \epsilon \tag{1}$$

$$\epsilon|\sigma_e^2 \sim \mathcal{N}(0, \sigma_e^2 \boldsymbol{I}_N) \tag{2}$$

$$\boldsymbol{\beta}|\sigma_g^2 \sim \mathcal{N}\left(0, \frac{\sigma_g^2}{M}\boldsymbol{I}_M\right). \tag{3}$$

Here $\boldsymbol{y}$ is centered so that $\sum_n y_n = 0$. σ_e^2 is the residual variance while σ_g^2 is the variance component corresponding to the M SNPs. The SNP heritability is defined as $h^2 = \frac{\sigma_g^2}{\sigma_g^2+\sigma_e^2}$.

In this model, we have $\mathbb{E}[\boldsymbol{y}] = 0$ while the population covariance of the phenotype vector $\boldsymbol{y}$ is:

$$\begin{aligned} cov(\boldsymbol{y}) &= \mathbb{E}[\boldsymbol{y}\boldsymbol{y}^{\mathrm{T}}] - \mathbb{E}[\boldsymbol{y}]\mathbb{E}[\boldsymbol{y}]^{\mathrm{T}} \\ &= \sigma_g^2 \frac{\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}}{M} + \sigma_e^2 \boldsymbol{I}_N \end{aligned} \tag{4}$$

$$= \sigma_g^2 \boldsymbol{K} + \sigma_e^2 \boldsymbol{I}_N. \tag{5}$$

Here $\boldsymbol{K} = \frac{1}{M}\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}$ is the genetic relatedness matrix (GRM) computed from all SNPs. One approach to estimate the SNP heritability is HE-regression (Haseman and Elston, 1972) which is a MoM estimator obtained by equating the population covariance to the empirical covariance [several variants of HE-regression have been proposed; what we consider here is HE-CP (Sham and Purcell, 2001)]. The empirical covariance of the phenotype vector $\boldsymbol{y}$ is estimated by $\boldsymbol{y}\boldsymbol{y}^{\mathrm{T}}$. The MoM estimator is obtained by solving the following ordinary least squares (OLS) problem (see Appendix A1 for details):

$$(\widehat{\sigma_g^2}, \widehat{\sigma_e^2}) = argmin_{\sigma_g^2,\sigma_e^2} ||\boldsymbol{y}\boldsymbol{y}^{\mathrm{T}} - (\sigma_g^2 \boldsymbol{K} + \sigma_e^2 \boldsymbol{I})||F^2. \tag{6}$$

The MoM estimator satisfies the normal equations:

$$\begin{bmatrix} \mathrm{tr}[\boldsymbol{K}^2] & \mathrm{tr}[\boldsymbol{K}] \\ \mathrm{tr}[\boldsymbol{K}] & N \end{bmatrix} \begin{bmatrix} \widehat{\sigma_g^2} \\ \widehat{\sigma_e^2} \end{bmatrix} = \begin{bmatrix} \boldsymbol{y}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{y} \\ \boldsymbol{y}^{\mathrm{T}}\boldsymbol{y} \end{bmatrix}. \tag{7}$$

Solving the normal equations requires computing $\mathrm{tr}[\boldsymbol{K}^2] = \sum_{i,j} K_{i,j}^2$, $\mathrm{tr}[\boldsymbol{K}] = \sum_i K_{i,i}$, $\boldsymbol{y}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{y} = \sum_{i,j} K_{i,j} y_i y_j$ and $\boldsymbol{y}^{\mathrm{T}}\boldsymbol{y} = \sum_{n=1}^{N} y_n^2$. The GRM $\boldsymbol{K}$ can be computed in time $\mathcal{O}(MN^2)$ and requires $\mathcal{O}(N^2)$ memory. Given the GRM, computing each of the coefficients for the normal equation requires $\mathcal{O}(N^2)$ time. Finally, given each of the coefficients, we can solve analytically solve for the $\widehat{\sigma_g^2}$ and $\widehat{\sigma_e^2}$. Indeed, we can write

$$\widehat{\sigma_g^2} = \frac{\boldsymbol{y}^{\mathrm{T}}(\boldsymbol{K} - \boldsymbol{I})\boldsymbol{y}}{\mathrm{tr}[K^2] - N}. \tag{8}$$

Thus, the key bottleneck in solving the HE-regression lies in computing the GRM.

2.2 RHE-reg: a randomized estimator of heritability

Given that $\boldsymbol{K} = \frac{1}{M}\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}$, we can compute the quantities $\mathrm{tr}[\boldsymbol{K}] = \frac{1}{M}\sum_{i,j} X_{i,j}^2$, $\boldsymbol{w} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}$, $\mathrm{tr}[\boldsymbol{y}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{y}] = \frac{1}{m}\sum_{m=1}^{M} w_m^2$. For standardized genotypes, $\mathrm{tr}[\boldsymbol{K}] = N$ while $\mathrm{tr}[\boldsymbol{y}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{y}]$ can be computed in $\mathcal{O}(MN)$ time.

The one remaining quantity that we need to compute efficiently is $\mathrm{tr}[\boldsymbol{K}^2]$. Given a $N \times N$ matrix $\boldsymbol{A}$ and a random vector $\boldsymbol{z}$ with mean zero and covariance $\boldsymbol{I}_N$, we use the following identity to construct a randomized estimator of the trace of matrix $\boldsymbol{A}$ (see Appendix A2 for a proof):

$$\mathbb{E}[\boldsymbol{z}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{z}] = \mathrm{tr}[\boldsymbol{A}]. \tag{9}$$

Equation (9) leads to the following unbiased estimator of the trace of $\boldsymbol{K}^2$ given B random vectors, $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_B$, drawn independently from a distribution with zero mean and identity covariance matrix $\boldsymbol{I}_N$:

$$\begin{aligned} L_B \equiv \widehat{\mathrm{tr}[\boldsymbol{K}^2]} &= \frac{1}{B}\sum_b \boldsymbol{z}_b^{\mathrm{T}}\boldsymbol{K}\boldsymbol{K}\boldsymbol{z}_b \\ &= \frac{1}{B}\frac{1}{M^2}\sum_b \boldsymbol{z}_b^{\mathrm{T}}\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{z}_b\,. \\ &= \frac{1}{B}\frac{1}{M^2}\sum_b ||\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{z}_b||^2 \end{aligned} \tag{10}$$

In practice, we draw each entry of $\boldsymbol{z}$ independently from a standard normal distribution. We note that the estimator L_B involves two matrix-vector multiplications of $N \times M$ matrix repeated B times for a total runtime of $\mathcal{O}(NMB)$.

The RHE-reg estimator $(\widetilde{\sigma_g^2}, \widetilde{\sigma_e^2})$ is obtained by solving the Normal equations [Equation (7)] by replacing $\mathrm{tr}[K^2]$ with L_B.

$$\begin{bmatrix} L_B & \mathrm{tr}[\boldsymbol{K}] \\ \mathrm{tr}[\boldsymbol{K}] & N \end{bmatrix} \begin{bmatrix} \widetilde{\sigma_g^2} \\ \widetilde{\sigma_e^2} \end{bmatrix} = \begin{bmatrix} \boldsymbol{y}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{y} \\ \boldsymbol{y}^{\mathrm{T}}\boldsymbol{y} \end{bmatrix}. \tag{11}$$

The RHE-reg estimator of the SNP heritability is then obtained by $h_{rhe}^2 = \frac{\widetilde{\sigma_g^2}}{s_y^2}$ where $s_y^2 = \frac{\boldsymbol{y}^{\mathrm{T}}\boldsymbol{y}}{N-1}$ is the unbiased estimator of the phenotypic variance.

2.3 Sub-linear computations

The key bottleneck in the RHE-reg is the computation of L_B which involves repeated multiplication of the normalized genotype matrix $\boldsymbol{X}$ by a real-valued vector. Leveraging the fact that each element of the genotype matrix $\boldsymbol{G}$ takes values in the set {0, 1, 2}, we can improve the complexity of these multiplication operations from $\mathcal{O}(NM)$ to $\mathcal{O}\left(\frac{NM}{\max(\log_3 N, \log_3 M)}\right)$ using the Mailman algorithm (Liberty and Zucker, 2009).

2.3.1 The Mailman algorithm

Consider a $M \times N$ matrix $\boldsymbol{A}^T$ whose entries take values in {0, 1, 2}. Assume that the number of SNPs $M = \log_3(N)$. The naive way to compute the product $\boldsymbol{A}^T\boldsymbol{b}$ for any real-valued vector $\boldsymbol{b}$ takes $O(log_3(N) * N)$ time.

The Mailman algorithm decomposes the matrix $\boldsymbol{A}$ as $\boldsymbol{A}^T = \boldsymbol{U}_n\boldsymbol{P}$. $\boldsymbol{U}_n$ is a $\log_3(N) \times N$ matrix whose column contains all possible vectors over {0, 1, 2} of length $\log_3(N)$. And P is an indicator matrix, where entry $P_{i,j} = 1$ if the ith column is the same as jth column in matrix $\boldsymbol{A} : \boldsymbol{A}^{(j)} = \boldsymbol{U}_n^{(i)}$. The decomposition of matrix $\boldsymbol{A}$ takes $\mathcal{O}(N \log_3(N))$ time. The desired product $\boldsymbol{A}^T\boldsymbol{b}$ is computed in two steps as $\boldsymbol{c} = \boldsymbol{P}\boldsymbol{b}$ followed by $\boldsymbol{U}_n\boldsymbol{c}$, each of which can be computed in only $\mathcal{O}(N)$ operations (Liberty and Zucker, 2009).

For a matrix $\boldsymbol{A}^{\mathrm{T}}$ with $M > \lceil \log_3(N) \rceil$, we partition $\boldsymbol{A}^{\mathrm{T}}$ into $\lceil \frac{M}{\lceil \log_3(N) \rceil} \rceil$ sub-matrices each of size $\lceil log_3(N) \times N \rceil$ each of which can be multiplied in time $\mathcal{O}(N)$ for a total computational cost of $\mathcal{O}\left(\frac{NM}{\log_3(N)}\right)$.

2.3.2 Application of the Mailman algorithm to RHE-reg

Now consider the standardized genotype $\boldsymbol{X}$, which could be written as $\boldsymbol{X} = (\boldsymbol{G} - \boldsymbol{M})\boldsymbol{\Sigma}$, where $\boldsymbol{M}$ is a matrix where the ith column contains the sample mean of the ith SNP ($\boldsymbol{M} = \boldsymbol{1}_N\bar{\boldsymbol{g}}^{\mathrm{T}}$), and $\boldsymbol{\Sigma}$ is an $M \times M$ diagonal matrix, with the inverse of variance of each SNP as the diagonal entries.

Thus, when we compute $\boldsymbol{y}^T\boldsymbol{K}\boldsymbol{y} = \frac{1}{M}\boldsymbol{y}^T\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{y} = \frac{1}{M}||\boldsymbol{\Sigma}(\boldsymbol{G}^{\mathrm{T}}\boldsymbol{y} - \boldsymbol{M}^{\mathrm{T}}\boldsymbol{y})||2^2$ in Equation (11), computing $\boldsymbol{G}^T\boldsymbol{y}$ using the Mailman algorithm takes $\mathcal{O}\left(\frac{NM}{\max(\log_3 M, \log 3}N)\right)$ operations. Similarly, to compute each term in the sum of the randomized estimator of $\mathrm{tr}[K^2]$ [Equation (10)], $\boldsymbol{X}^T\boldsymbol{z}_b$, we can substitute $\boldsymbol{X}^T\boldsymbol{z}_b$ with $\boldsymbol{\Sigma}\boldsymbol{G}^T\boldsymbol{z}_b - \boldsymbol{\Sigma}\boldsymbol{M}^T\boldsymbol{z}_b$. The first term $\boldsymbol{\Sigma}\boldsymbol{G}^T\boldsymbol{z}_b$ can again be computed using $\mathcal{O}\left(\frac{NM}{\max(\log_3 M, \log_3 N)}\right)$ using the Mailman algorithm, and the second term $\boldsymbol{\Sigma}\boldsymbol{M}^T\boldsymbol{z}_b$ is equivalent to scaling the N-vector $\boldsymbol{z}_b$ which can be computed in time $\mathcal{O}(N + M)$.

2.4 Computing the standard error

We show in Appendix A4 that the variance of the RHE-reg estimator of σ_g^2 can be approximated by the variance of the exact HE-regression estimator with an additional contribution due to the randomization:

$$\mathrm{Var}\left[\widetilde{\sigma_g^2}\right] \approx \mathrm{Var}\left[\widehat{\sigma_g^2}\right] + \frac{1}{B(\mathrm{tr}[K^2] - N)^2}\left(\sigma_g^4\right)\mathrm{tr}\left[K^2\right].$$

Here B is the number of samples used and $\boldsymbol{z}$ is a random vector with mean zero and identity covariance matrix. For samples with low-levels of relatedness, we can assume $\boldsymbol{K} \approx \boldsymbol{I}$ and our estimates of σ_g^2 and $\mathrm{tr}[\boldsymbol{K}^2]$ to estimate the variance. Further, we show in Appendix A4 that we can estimate the variance (and hence, the standard error) of the RHE-reg estimator in sub-linear time without assuming that $\boldsymbol{K} \approx \boldsymbol{I}$.

2.5 Some remarks on the RHE-reg estimator

1. The RHE-reg is biased as we show in Appendix A3 with a bias that decreases with B. In practice, the bias appears to be small (see Fig. 1).

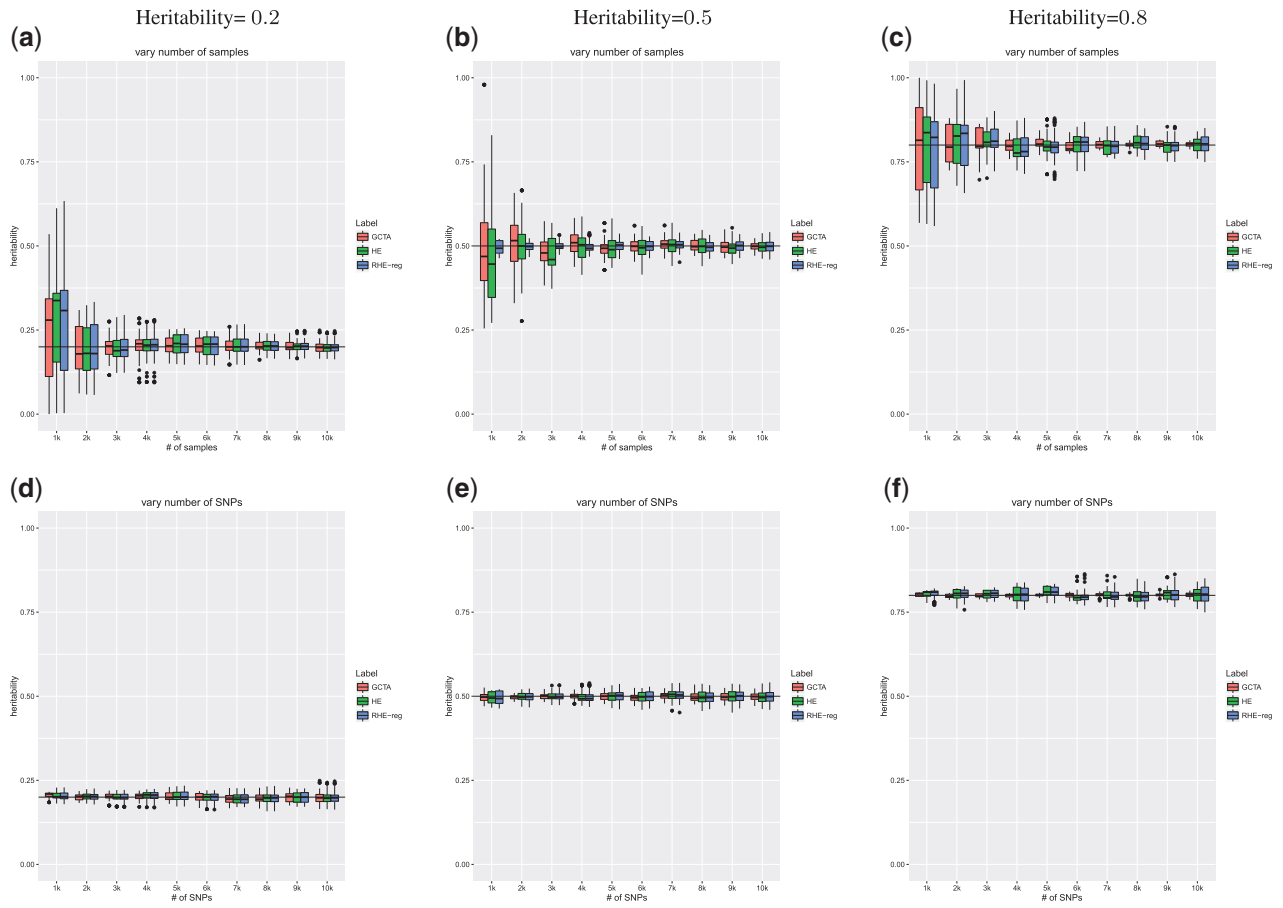

Fig. 1. RHE-reg accurately estimates heritability: in the first series of (**a–c**), we simulated genotypes with varying sample size while fixing the number of SNPs to 10 000. The phenotype in each of the (a), (b) and (c) is simulated with true heritability of 0.2, 0.5 and 0.8, respectively. The second series of (**d–f**) considers genotype data with varying number of SNPs while the number of samples is fixed at 10 000. All three methods that we evaluated (GCTA, HE-reg and RHE-reg) have similar accuracies. GCTA which estimates the REML has smaller standard errors when the heritability is large ($h^2 = 0.80$). For lower values of true heritability ($h^2 = 0.20, h^2 = 0.50$), the estimates from REML, HE-regression and RHE-reg are comparable. HE and RHE-reg have similar variance suggesting that randomization only makes a minor contribution to the statistical accuracy

2. Equation (3) assumes an infinitesimal model for the phenotype. However, all our results only depend on the second moment of the SNP effect sizes. Thus, the RHE-reg estimator can yield valid estimates for non-infinitesimal architectures.
3. In a number of settings, it is desirable to include covariates, such as age or sex, in the analysis. This changes the model in Equation (3) to:

$$\boldsymbol{y}|\epsilon, \boldsymbol{\beta} = \boldsymbol{W}\alpha + \boldsymbol{X}\boldsymbol{\beta} + \epsilon. \tag{12}$$

Here $\boldsymbol{W}$ is a $N \times C$ matrix of covariates while α is a C-vector of coefficients. In this setting, we transform Equation (12) by multiplying by the projection matrix $\boldsymbol{V} = \boldsymbol{I}_N - \boldsymbol{W}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W})^{-1}\boldsymbol{W}^{\mathrm{T}}$:

$$\boldsymbol{V}\boldsymbol{y} = \boldsymbol{V}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{V}\epsilon. \tag{13}$$

The RHE-regression estimator applied to Equation (13) then must satisfy the following moment conditions:

$$\begin{bmatrix} J_B & \mathrm{tr}[\boldsymbol{V}\boldsymbol{K}] \\ \mathrm{tr}[\boldsymbol{V}\boldsymbol{K}] & N - C \end{bmatrix} \begin{bmatrix} \widetilde{\sigma_g^2} \\ \widetilde{\sigma_e^2} \end{bmatrix} = \begin{bmatrix} \boldsymbol{y}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{K}\boldsymbol{V}\boldsymbol{y} \\ \boldsymbol{y}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{y} \end{bmatrix}. \tag{14}$$

Here J_B is a randomized estimator of $\mathrm{tr}[\boldsymbol{V}\boldsymbol{K}\boldsymbol{V}\boldsymbol{K}]$ analogous to Equation (10). The cost of computing the RHE-reg estimator now includes the cost of computing the inverse of $\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}}$ as well as multiplying $\boldsymbol{W}$ by a real-valued vector for an added computational cost of $\mathcal{O}(C^3 + NC)$. Typically, the number of covariates C is small (tens to hundreds) so that the presence of covariance does not greatly increase the computational burden.

4. The variance components model [Equations (3) and (5)] can be extended in a straightforward manner to more than two variance components. A number of recent studies have explored the utility of these models to partition heritability based on functional annotations as well as other categories.
5. The accuracy and the runtime of RHE-reg depends on the choice of the number of random vectors B. In practice, we find that the estimator is highly accurate with a small $B \approx 100$ even for moderate sample sizes $N \approx 5\,000$ as we show empirically (Fig. 2). Further, for larger sample sizes, even smaller values of B should be adequate. It is also possible to choose increasing values of B and to terminate when the estimate of $\mathrm{tr}[\boldsymbol{K}^2]$ does not change considerably. We have not explored this option in detail in this work.

3 Results

3.1 Simulations

We performed simulations to measure the performance of RHE-reg to other methods for heritability estimation in terms of accuracy,

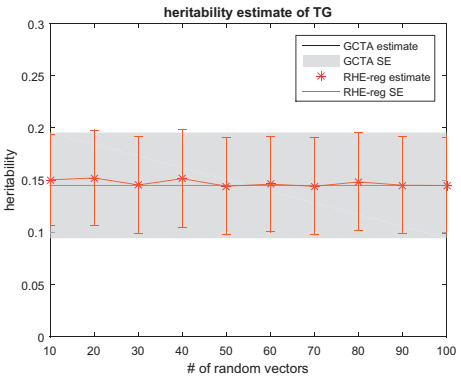

Fig. 2. Impact of the number of random vectors on the accuracy of RHE-reg: we ran RHE-reg with a different number of random vectors *B*, and compared the point estimate and standard error to GCTA. The gray area indicates the standard error computed by GCTA. As RHE-reg use more random vectors, the estimate converges. In fact, even with 10 random vectors, the point estimation is accurate

running time and memory usage. We compared RHE-reg to two methods for computing REML estimates: GCTA (Yang *et al.*, 2011) (which implements an exact numerical optimization algorithm to compute the REML) as well as implementations of HE-regression.

3.2 Accuracy

In our first set of simulations, we compared the accuracy of RHE-reg to our implementation of exact HE-regression as well as GCTA, an implementation that computes the REML. We simulated genotypes assuming each SNP is drawn independently from a Binomial distribution with allele frequency that is sampled uniformly from the interval (0, 1). Given the genotypes, we simulated phenotypes under an infinitesimal model, i.e. with effect size at each SNP drawn independently from a normal distribution with mean zero and variance equal to the heritability divided by the number of SNPs. We considered different values for the true SNP heritability of the phenotype to be 0.2, 0.5 and 0.8.

In our first series of experiments, we fixed the number of SNPs at $M = 10\,000$ and varied the number of individuals $N = 1k, 2k \ldots 10k$. In the second series of experiments, we varied the number of SNPs $M = 1k, 2k \ldots 10k$ while fixing the number of individuals to be $N = 10\,000$. We repeated each experiment 100 times in order to assess the variance of each of the estimators. We estimated heritability using RHE-reg with $B = 100$ random vectors.

Figure 1 compares the estimates of each of the three methods (RHE-reg, HE-regression and GCTA) to the true heritability. First, we observe that all three methods obtain estimates of heritability that are quite close to each other as well as to the true heritability across the range of parameters explored. Second, RHE-reg and HE-regression are virtually indistinguishable in the variance of their estimates in each configuration. This suggests that the randomization makes a negligible contribution to the statistical accuracy of the MoM estimators. In some cases, RHE-reg even has a smaller variance than HE-regression. Third, as expected, REML obtains estimators that are closer to the true heritability compared to either of the MoM estimators for a high value of true heritability. For lower values of true heritability ($h^2 = 0.20, h^2 = 0.50$), the estimates from REML, HE-regression and RHE-reg are comparable. This result is also expected given that REML is asymptotically equivalent to MoM when the phenotypic correlation between individuals is small (Sham *et al.*, 2000; Sham and Purcell, 2001). Finally, the sample size has a bigger effect than the number of SNPs on the accuracy of each of the methods, consistent with theory (Visscher *et al.*, 2014).

3.3 Computational efficiency

In the second set of simulations, we compared the runtime and memory usage of different methods. We compared RHE-reg to two REML methods, GCTA (Yang *et al.*, 2011) and BOLT-REML (Loh *et al.*, 2015a) (a computationally efficient approximate method to compute the REML) as well as an exact MoM method MMHE (Ge *et al.*, 2017). In this experiment, we simulated genotype data consisting of 100 000 SNPs over sample sizes of $N = 10k, 20k, 30k, 50k, 100k$ and $500\,k$ and then simulated phenotypes corresponding to the genotype data. For each dataset, we ran RHE-reg with $B = 100$ random vectors. We performed all comparisons on an Intel(R) Xeon(R) CPU 2.10 GHz server with 128 GB RAM. All computations were restricted to a single core, capped to a maximum runtime of 12 h and a maximum memory of 128 GB.

Figure 3 shows that both GCTA and MMHE do not scale to large sample sizes due to the requirement of computing and operating on a GRM that scales quadratically with *N*. GCTA could not complete its computation when running on $N = 100K$ individuals while MMHE did not complete its computation on $N = 50K$. BOLT-REML and RHE-reg scale linearly with sample size. However, RHE-reg is an order of magnitude faster than BOLT-REML. For example, on a dataset of a size of $500\,K$ individuals, RHE-reg computed the heritability in about 30 min compared to 400 min for BOLT-REML. Figure 3 shows that RHE-reg is memory efficient as well.

3.4 Application to real data

We compared the statistical accuracy and runtime of BOLT-REML, GCTA and RHE-reg on the Northern Finland Birth Cohort (NFBC) dataset. The NFBC dataset contains 315 529 SNPs and 5326 individuals after applying standard filters (minor allele frequency >0.05 and Hardy–Weinberg equilibrium *P*-value <0.01) (Sabatti *et al.*, 2009). We applied these methods to estimate the heritability of three phenotypes that were assayed in this dataset: triglycerides (TGs), high-density lipoprotein (HDL) and body mass index (BMI).

We compared the runtime, point estimates of the heritability as well as standard errors for each of the three methods. We computed RHE-reg with $B = 100$ random vectors. As shown in Table 1, the heritability estimates of RHE-reg are concordant with the other methods while being an order of magnitude faster to compute. We note that the NFBC dataset has a sample size $N \approx 5000$ so that we expect RHE-reg to be more accurate on larger datasets. The standard error estimates can also be computed in sub-linear time (see Appendix A4).

3.5 Understanding the computational efficiency of RHE-reg

Our implementation of RHE-reg relies on two ideas to obtain computational efficiency: (i) the use of a randomized estimator of the trace, and (ii) the Mailman algorithm for fast matrix-vector multiplication. To explore the contribution of each of these ideas, we compared the runtimes of a MoM estimator with no randomization (HE-reg), RHE-reg using standard matrix-vector multiplication and RHE-reg using the Mailman algorithm. Table 2 shows the runtimes of each of these variants on the NFBC data. We see that the biggest runtime gain arises from applying the randomized estimator (faster by a factor of 10–12 relative to HE-reg) while the application of the Mailman algorithm reduces the runtime further by a factor of 2 (Table 1).

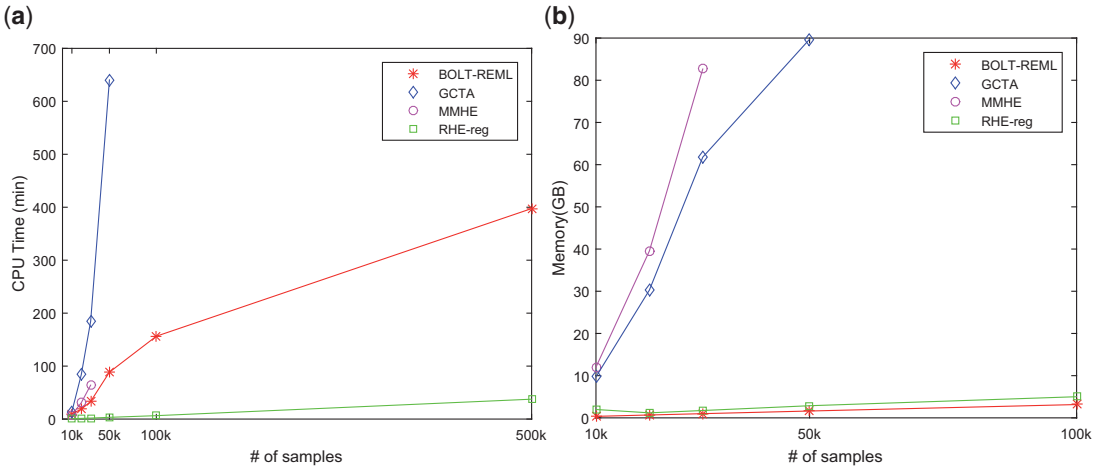

Fig. 3. RHE-reg is efficient: we measured the run time and memory usage of methods for heritability estimation as a function of the number of samples while fixing the number of SNPs to 100 000. We performed all comparisons on an Intel(R) Xeon(R) CPU 2.10 GHz server with 128 GB RAM. All computations were restricted to a single core, capped to a maximum runtime of 12 h and a maximum memory of 128 GB. In (**a**), GCTA could not finish computation on 100 K samples. For MMHE, the computation stopped at sample size of 50 k due to memory constraints. Although BOLT-REML scales linearly, RHE-reg is significantly faster. In (**b**), we observe RHE-reg and BOLT-REML have scalable memory requirements

Table 1. The estimates of heritability from RHE-reg are consistent with those from GCTA and BOLT-REML on the NFBC data while RHE-reg is substantially faster

Method	GCTA		BOLT-REML		RHE-reg	
	Runtime (min)	h_g^2 (SE)	Runtime (min)	h_g^2 (SE)	Runtime (min)	h_g^2 (SE)
TG	11.28	0.145	8.87	0.148	1.61	0.145
	–	(0.051)	–	(0.051)	–	(0.052)
HDL	10.81	0.325	9.72	0.326	1.30	0.349
	–	(0.051)	–	(0.051)	–	(0.052)
BMI	10.85	0.237	9.29	0.235	1.29	0.200
	–	(0.051)	–	(0.051)	–	(0.052)

Note: We estimate the heritability of phenotypes such as triglycerides (TGs), high-density lipoprotein (HDL) and body mass index (BMI) in the NFBC data set.

Table 2. The major gain in computational efficiency arises from the application of the randomized trace estimate

	Runtime (min)	No Mailman (min)	No randomized trace estimate (min)
TG	1.61	3.70	38.5
HDL	1.30	2.60	36.2
BMI	1.29	2.68	36.7

Note: We compare the run time for HE-reg as well as the run time for RHE-reg that does not rely on the Mailman algorithm.

3.6 Accuracy of RHE-reg as a function of the number of random vectors *B*

To explore the impact of the choice of the number of random vectors B on the accuracy of RHE-reg, we compared the heritability estimates of RHE-reg to those obtained from GCTA for the TG phenotype as a function of B. We find good concordance between the estimates from RHE-reg and GCTA even for values of B as low as 10 suggesting that RHE-reg could be even faster in practice with little loss in accuracy (see Fig. 2).

4 Discussion

We proposed a scalable estimator of heritability which is a randomized version of the Haseman–Elston regression (RHE-reg) estimator. The RHE-reg estimator is based on performing a small number of multiplications of the genotype matrix with random vectors with mean zero and identity covariance. Using the properties of the genotype matrix, we can compute this estimator using the Mailman algorithm in $\mathcal{O}\left(\frac{NMB}{\max(\log_3 N, \log_3 M)}\right)$ time on a dataset containing N individuals, M SNPs and with a small number of B random vectors. We show that this estimator achieves similar accuracy as REML-based methods on both simulated and real data. RHE-reg can be effectively applied to whole-genome genotype data of hundreds of thousands of individuals for rapid variance components estimation. Furthermore, RHE-reg is an unbiased estimator and thus can also be applied to ascertained case-control studies.

Acknowledgements

We thank Xiang Zhou and the reviewers for their valuable feedback.

Funding

This work was supported in part by NIH grants R00GM111744, R35GM125055, NSF Grant III-1705121, an Alfred P. Sloan Research Fellowship, and a gift from the Okawa Foundation.

Conflict of Interest: none declared.

References

Almasy,L. and Blangero,J. (1998) Multipoint quantitative-trait linkage analysis in general pedigrees. *Am. J. Hum. Genet.*, **62**, 1198–1211.

Bulik-Sullivan,B. (2015) Relationship between ld score and Haseman-Elston regression. *bioRxiv*, 018283.

Chen,G.-B. (2014) Estimating heritability of complex traits from genome-wide association studies using ibs-based Haseman–Elston regression. *Front. Genet.*, **5**, 107.

Chen,W.-M. *et al.* (2004) Quantitative trait linkage analysis by generalized estimating equations: unification of variance components and Haseman-Elston regression. *Genet. Epidemiol.*, **26**, 265–272.
Elston,R.C. *et al.* (2000) Haseman and Elston revisited. *Genet. Epidemiol.*, **19**, 1–17.
Ge,T. *et al.* (2017) Phenome-wide heritability analysis of the UK Biobank. *PLoS Genet.*, **13**, e1006711.
Golan,D. *et al.* (2014) Measuring missing heritability: inferring the contribution of common variants. *Proc. Natl. Acad. Sci.*, **111**, E5272–E5281.
Haseman,J. and Elston,R. (1972) The investigation of linkage between a quantitative trait and a marker locus. *Behav. Genet.*, **2**, 3–19.
Houle,D. (1992) Comparing evolvability and variability of quantitative traits. *Genetics*, **130**, 195–204.
Kang,H.M. *et al.* (2008a) Accurate discovery of expression quantitative trait loci under confounding from spurious and genuine regulatory hotspots. *Genetics*, **180**, 1909–1925.
Kang,H.M. *et al.* (2008b) Efficient control of population structure in model organism association mapping. *Genetics*, **178**, 1709–1723.
Liberty,E. and Zucker,S.W. (2009) The Mailman algorithm: a note on matrix–vector multiplication. *Inf. Process. Lett.*, **109**, 179–182.
Lippert,C. *et al.* (2011) FaST linear mixed models for genome-wide association studies. *Nat. Methods*, **8**, 833–835.
Loh,P.-R. *et al.* (2015a) Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. *Nat. Genet.*, **47**, 1385.
Loh,P.-R. *et al.* (2015b) Efficient bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.*, **47**, 284.
Makowsky,R. *et al.* (2011) Beyond missing heritability: prediction of complex traits. *PLoS Genet.*, **7**, e1002051.
Pirinen,M. *et al.* (2013) Efficient computation with a linear mixed model on large-scale data sets with applications to genetic studies. *Annal. Appl. Stat.*, **7**, 369–390.
Sabatti,C. *et al.* (2009) Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nat. Genet.*, **41**, 35–46.
Sham,P. and Purcell,S. (2001) Equivalence between Haseman-Elston and variance-components linkage analyses for sib pairs. *Am. J. Hum. Genet.*, **68**, 1527–1532.
Sham,P. *et al.* (2000) Power of linkage versus association analysis of quantitative traits, by use of variance-components models, for sibship data. *Am. J. Hum. Genet.*, **66**, 1616–1630.
Speed,D. *et al.* (2012) Improved heritability estimation from genome-wide snps. *Am. J. Hum. Genet.*, **91**, 1011–1021.
Visscher,P.M. *et al.* (2008) Heritability in the genomics era? Concepts and misconceptions. *Nat. Rev. Genet.*, **9**, 255.
Visscher,P.M. *et al.* (2014) Statistical power to detect genetic (co) variance of complex traits using snp data in unrelated samples. *PLoS Genet.*, **10**, e1004269.
Wray,N.R. *et al.* (2013) Pitfalls of predicting complex traits from snps. *Nat. Rev. Genet.*, **14**, 507.
Yang,J. *et al.* (2010) Common snps explain a large proportion of the heritability for human height. *Nat. Genet.*, **42**, 565.
Yang,J. *et al.* (2011) GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.*, **88**, 76–82.
Yu,J. *et al.* (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.*, **38**, 203.
Zhou,X. and Stephens,M. (2014) Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nat. Methods*, **11**, 407.
Zhou,X. *et al.* (2013) Polygenic modeling with bayesian sparse linear mixed models. *PLoS Genet.*, **9**, e1003264.

Appendix

A1. Method-of-Moments

The MoM principle obtains estimates of the model parameters such that the theoretical moments match the sample moments. In our model, the first theoretical moment, $\mathbb{E}[\boldsymbol{y}]$, is **0** by definition while the corresponding sample moment is also zero since we standardized the phenotypes. The second sample moment is $\boldsymbol{y}\boldsymbol{y}^T$ and the second theoretical moment is $cov(\boldsymbol{y}) = \sigma_g^2\boldsymbol{K} + \sigma_e^2\boldsymbol{I}_N$. Thus, the MoM estimator of (σ_g^2, σ_e^2) is obtained by searching for values of σ_g^2, σ_e^2 such that the sample and theoretical moments are close, i.e. by solving an ordinary least squares (OLS) problem:

$$(\widehat{\sigma_g^2}, \widehat{\sigma_e^2}) = argmin_{\sigma_g^2,\sigma_e^2} ||\boldsymbol{y}\boldsymbol{y}^T - (\sigma_g^2\boldsymbol{K} + \sigma_e^2\boldsymbol{I})||_F^2.$$

Since the Frobenius norm of a matrix $\boldsymbol{A}$, $||\boldsymbol{A}||_F = \sqrt{\text{tr}[\boldsymbol{A}\boldsymbol{A}^T]}$, the OLS problem can be re-written as:

$$(\widehat{\sigma_g^2}, \widehat{\sigma_e^2}) = argmin_{\sigma_g^2,\sigma_e^2} \text{tr}[(\boldsymbol{y}\boldsymbol{y}^T - (\sigma_g^2\boldsymbol{K} + \sigma_e^2\boldsymbol{I}))(\boldsymbol{y}\boldsymbol{y}^T - (\sigma_g^2\boldsymbol{K} + \sigma_e^2\boldsymbol{I}))^T]$$

which leads to Equation (6).

A2. Randomized estimator of trace of a matrix

For a $N \times N$ matrix, $\boldsymbol{A}$, a randomized estimator of $\text{tr}[\boldsymbol{A}]$ is $\widehat{\text{tr}[\boldsymbol{A}]} \equiv \frac{1}{B}\sum_b \boldsymbol{z}_b T\boldsymbol{A}\boldsymbol{z}_b$, where $\boldsymbol{z}_b$ are i.i.d. random vectors with each entry drawn from a standard normal distribution. To see this:

$$\begin{aligned}
\mathbb{E}[\boldsymbol{z}^T\boldsymbol{A}\boldsymbol{z}] &= \mathbb{E}[tr(\boldsymbol{z}^T\boldsymbol{A}\boldsymbol{z})] \quad \boldsymbol{z}^T\boldsymbol{A}\boldsymbol{z} \text{ is a scalar} \\
&= \mathbb{E}[\text{tr}[\boldsymbol{z}\boldsymbol{z}^T\boldsymbol{A}]] \quad \text{cyclic property of the trace} \\
&= \text{tr}[\mathbb{E}[\boldsymbol{z}\boldsymbol{z}^T\boldsymbol{A}]] \quad \text{trace and expectation are linear} \\
&= \text{tr}[\mathbb{E}[\boldsymbol{z}\boldsymbol{z}^T]\boldsymbol{A}] \boldsymbol{A} \text{ is fixed} \\
&= \text{tr}[\boldsymbol{A}] \text{ using the distributional assumptions on } \boldsymbol{z}.
\end{aligned}$$

A3. Bias of the RHE-reg estimator

Our estimator of $\text{tr}[\boldsymbol{K}^2]$ is $L_B \equiv \widehat{tr[K^2]} = \frac{1}{B}\sum_B \boldsymbol{z}_b^T\boldsymbol{K}\boldsymbol{K}\boldsymbol{z}_b$. The RHE-reg estimators for $\left(\sigma_g^2, \sigma_e^2\right)$ are given by: $\begin{bmatrix}\widetilde{\sigma_g^2} \\ \widetilde{\sigma_e^2}\end{bmatrix} = \boldsymbol{A}^{-1}\begin{bmatrix}\boldsymbol{y}^T\boldsymbol{K}\boldsymbol{y} \\ \boldsymbol{y}^T\boldsymbol{y}\end{bmatrix}$ where $\boldsymbol{A} = \begin{bmatrix}L_B & N \\ N & N\end{bmatrix}$.

We first compute the expectation of this estimator:

$$\begin{aligned}
\mathbb{E}\begin{bmatrix}\widetilde{\sigma_g^2} \\ \widetilde{\sigma_e^2}\end{bmatrix} &= \mathbb{E}\left[\boldsymbol{A}^{-1}\begin{bmatrix}\boldsymbol{y}^T\boldsymbol{K}\boldsymbol{y} \\ \boldsymbol{y}^T\boldsymbol{y}\end{bmatrix}\right] \\
&= \mathbb{E}[\boldsymbol{A}^{-1}]\mathbb{E}\begin{bmatrix}\boldsymbol{y}^T\boldsymbol{K}\boldsymbol{y} \\ \boldsymbol{y}^T\boldsymbol{y}\end{bmatrix} \quad \begin{array}{l}\text{since random vectors } \boldsymbol{z}_b \text{ and} \\ \boldsymbol{y} \text{ are independent}\end{array}
\end{aligned}$$

We know that $\mathbb{E}[\boldsymbol{y}\boldsymbol{y}^T] = cov(\boldsymbol{y}) = \sigma_g^2\boldsymbol{K} + \sigma_e^2\boldsymbol{I}$. We can compute $\mathbb{E}[\boldsymbol{y}^T\boldsymbol{K}\boldsymbol{y}]$:

$$\begin{aligned}
\mathbb{E}[\boldsymbol{y}^T\boldsymbol{K}\boldsymbol{y}] &= \mathbb{E}[\text{tr}[\boldsymbol{y}^T\boldsymbol{K}\boldsymbol{y}]] \quad \boldsymbol{y}^T\boldsymbol{K}\boldsymbol{y} \text{ is a scalar} \\
&= \mathbb{E}[\text{tr}[\boldsymbol{y}\boldsymbol{y}^T\boldsymbol{K}]] \quad \text{cyclic property of the trace} \\
&= \text{tr}[\mathbb{E}[\boldsymbol{y}\boldsymbol{y}^T\boldsymbol{K}]] \quad \begin{array}{l}\text{expectation and trace are} \\ \text{linear}\end{array} \\
&= \text{tr}[\mathbb{E}[\boldsymbol{y}\boldsymbol{y}^T]\boldsymbol{K}] \quad \boldsymbol{K} \text{ as constant} \\
&= \text{tr}\left[\sigma_g^2\boldsymbol{K}^2 + \sigma_e^2\boldsymbol{K}\right] \\
&= \sigma_g^2\text{tr}[\boldsymbol{K}^2] + N\sigma_e^2 \quad \text{using } \text{tr}[K] = N.
\end{aligned}$$

And for $\mathbb{E}[\boldsymbol{y}^T\boldsymbol{y}]$, we have;

$$\mathbb{E}[\boldsymbol{y}^T\boldsymbol{y}] = \mathbb{E}[\text{tr}[\boldsymbol{y}^T\boldsymbol{y}]] \quad \boldsymbol{y}^T\boldsymbol{y} \text{ is a scalar}$$

$$
\begin{aligned}
&= \mathbb{E}(\mathrm{tr}[\boldsymbol{y}\boldsymbol{y}^T]] \quad \texttt{cyclic property of the trace} \\
&= \mathrm{tr}[\mathbb{E}[\boldsymbol{y}\boldsymbol{y}^T]] \quad \texttt{expectation and trace are linear} \\
&= \mathrm{tr}[\boldsymbol{K}]\sigma_g^2 + N\sigma_e^2 \\
&= N\sigma_g^2 + N\sigma_e^2.
\end{aligned}
$$

Defining $b \equiv \mathbb{E}\left[\frac{1}{L_B - N}\right]$ and computing $\boldsymbol{A}^{-1} = \begin{bmatrix} \frac{1}{L_B - N} & \frac{-1}{L_B - N} \\ \frac{-1}{L_B - N} & \frac{L_B}{N(L_B - N)} \end{bmatrix}$, we have

$$
\begin{aligned}
\mathbb{E}\begin{bmatrix} \widetilde{\sigma_g^2} \\ \widetilde{\sigma_e^2} \end{bmatrix} &= \mathbb{E}[\boldsymbol{A}^{-1}]\mathbb{E}\left[\begin{bmatrix} \boldsymbol{y}^T\boldsymbol{K}\boldsymbol{y} \\ \boldsymbol{y}^T\boldsymbol{y} \end{bmatrix}\right] \\
&= \begin{bmatrix} b & -b \\ -b & \frac{1}{N} + b \end{bmatrix}\begin{bmatrix} \mathrm{tr}[\boldsymbol{K}^2] + N\sigma_e^2 \\ N\sigma_g^2 + N\sigma_e^2 \end{bmatrix} \\
&= \begin{bmatrix} b(\mathrm{tr}[\boldsymbol{K}^2] - N)\sigma_g^2 \\ b(N - \mathrm{tr}[\boldsymbol{K}^2])\sigma_g^2 + \sigma_g^2 + \sigma_e^2 \end{bmatrix}.
\end{aligned}
$$

We approximate $b = \mathbb{E}\left[\frac{1}{L_B - N}\right]$ using Taylor expansion. As we have: $f(y) \approx f(x) + f'(x)(y-x) + \frac{1}{2}f''(x)(y-x)^2$. Let $X \equiv L_B - N$, and thus $\mu_x = \mathbb{E}[L_B - N] = \mathrm{tr}[\boldsymbol{K}^2] - N$. We have $f(x) = \frac{1}{x}, f'(x) = -\frac{1}{x^2}, f''(x) = \frac{2}{x^3}$.

Thus:

$$
\begin{aligned}
b = \mathbb{E}[f(X)] &\approx \mathbb{E}\left[f(\mu_x) + f'(\mu_x)(X - \mu_x) + \frac{1}{2}f''(\mu_x)(X - \mu_x)^2\right] \\
&= f(\mu_x) + \frac{1}{-\mu_x^2}\mathbb{E}[X - \mu_x] + \frac{1}{2}\frac{2}{\mu_x^3}\mathbb{E}\left[(X - \mu_x)^2\right. \\
&\qquad \left. = \frac{1}{\mu_x} + \frac{1}{\mu_x}\frac{\sigma_x^2}{\mu_x^2}\right]
\end{aligned}
$$

where $\sigma_x^2 = var(X)$.

Thus $\mathbb{E}\left[\frac{\mu_x}{x}\right] = 1 + \frac{\sigma_x^2}{\mu_x^2}$. Thus $\mathbb{E}\left[\widetilde{\sigma_g^2}\right] = \sigma_g^2 + \frac{\sigma_x^2}{\mu_x^2}\sigma_g^2$, $\mathbb{E}\left[\sigma_e^2\right] = \sigma_e^2 - \frac{\sigma_x^2}{\mu_x^2}\sigma_g^2$, $\mathbb{E}\left[\widetilde{\sigma_g^2} + \widetilde{\sigma_e^2}\right] = \sigma_g^2 + \sigma_e^2$.

For σ_x^2, we have:

$$
\begin{aligned}
\sigma_x^2 &= \mathbb{E}\left[\left(L_B - \mathrm{tr}[\boldsymbol{K}^2]\right)^2\right] \\
&= var(L_B) \\
&= var\left(\frac{1}{B}\sum_B \boldsymbol{z}_b^T\boldsymbol{K}^2\boldsymbol{z}_b\right) \quad \boldsymbol{z}_b \texttt{ are independent} \\
&= \frac{1}{B^2}\sum_B var(\boldsymbol{z}_b^T\boldsymbol{K}^2\boldsymbol{z}_b) \quad \boldsymbol{z}_b \texttt{ are identically distributed} \\
&= \frac{1}{B}\sum_{i,j} \boldsymbol{K}_i^T\boldsymbol{K}_j z_i z_j \quad \texttt{elements of } \boldsymbol{z} \texttt{ are independent} \\
&= \frac{1}{B}\sum_i \boldsymbol{K}_i^2 = \frac{1}{B}\mathrm{tr}[\boldsymbol{K}^2].
\end{aligned}
$$

Here $\boldsymbol{K}_i$ is the ith column of K.

Thus, substituting μ_x and σ_x^2, we get $\mathbb{E}\left[\widetilde{\sigma_g^2}\right] = \sigma_g^2 + \frac{1}{B}\frac{\mathrm{tr}[\boldsymbol{K}^2]}{(\mathrm{tr}[\boldsymbol{K}^2] - N)^2}\sigma_g^2 = \sigma_g^2 + \frac{1}{B}\frac{1}{\mathrm{tr}[\boldsymbol{K}^2] - 2N + \frac{N^2}{\mathrm{tr}[\boldsymbol{K}^2]}}\sigma_g^2$. The bias of the estimator decreases with larger number of random vectors B.

A4. Standard error estimate for the RHE-reg estimator

We define $var(\boldsymbol{y}) \equiv \boldsymbol{\Sigma} = \sigma_g^2\boldsymbol{K} + \sigma_e^2\boldsymbol{I}$. As we know $\widetilde{\sigma_g^2} = \frac{\boldsymbol{y}^T(\boldsymbol{K} - \boldsymbol{I})\boldsymbol{y}}{L_B - N}$. Let $\widetilde{\sigma_g^2} \equiv \frac{A}{B}$ where $A \equiv \boldsymbol{y}^T(K - I)y$ and $B \equiv L_B - N$. Define $\mu_A \equiv \mathbb{E}[A]$, $\mu_B \equiv \mathbb{E}[B]$, $\sigma_A^2 \equiv var(A)$ and $\sigma_B^2 \equiv var(B)$. From Lemma 2 (Appendix A5), we have

$$
\begin{aligned}
var\left(\widetilde{\sigma_g^2}\right) &= var\left(\frac{A}{B}\right) \\
&= \frac{1}{(\mu_B)^2}\sigma_A^2 - 2\frac{\mu_A}{(\mu_B)^3}cov(A,B) + \frac{(\mu_A)^2}{(\mu_B)^4}\sigma_B^2 \\
&= \frac{1}{(\mu_B)^2}\sigma_A^2 + \frac{(\mu_A)^2}{(\mu_B)^4}\sigma_B^2
\end{aligned}
$$

as A, B are independent. By using Lemma 1 (Appendix A5), we have:

$$
\begin{aligned}
\mu_A &= \mathbb{E}\left[\boldsymbol{y}^T(\boldsymbol{K} - \boldsymbol{I})\boldsymbol{y}\right] = \left(\mathrm{tr}[\boldsymbol{K}^2] - N\right)\sigma_g^2 \\
\sigma_A^2 &= var\left(\boldsymbol{y}^T(\boldsymbol{K} - \boldsymbol{I})\boldsymbol{y}\right) = 2\mathrm{tr}[\boldsymbol{\Sigma}(\boldsymbol{K} - \boldsymbol{I})\boldsymbol{\Sigma}(\boldsymbol{K} - \boldsymbol{I})] \\
\mu_B &= \mathrm{tr}[\boldsymbol{K}^2] - N \\
\sigma_B^2 &= \frac{\mathrm{tr}[\boldsymbol{K}^2]}{B}.
\end{aligned}
$$

Thus we have:

$$
SE\left(\widetilde{\sigma_g^2}\right) = \frac{1}{\mathrm{tr}[\boldsymbol{K}^2] - N}\sqrt{2\mathrm{tr}[\boldsymbol{\Sigma}(\boldsymbol{K} - \boldsymbol{I})\boldsymbol{\Sigma}(\boldsymbol{K} - \boldsymbol{I})] + \frac{1}{B}\left(\sigma_g^2\right)^2\mathrm{tr}[\boldsymbol{K}^2]}.
$$

In order to estimate the standard error of $\widetilde{\sigma_g^2}$, we use the plug-in estimator:

$$
\widehat{SE\left(\widetilde{\sigma_g^2}\right)} = \frac{1}{L_B - N}\sqrt{2\mathrm{tr}[\boldsymbol{y}\boldsymbol{y}^T(\boldsymbol{K} - \boldsymbol{I})\boldsymbol{\Sigma}(\boldsymbol{K} - \boldsymbol{I})] + \frac{1}{B}\left(\widetilde{\sigma_g^2}\right)^2 L_B}. \quad (15)
$$

Each term in this estimator could be efficiently computed in $O\left(\frac{NMB}{\max(\log_3 N, \log_3 M)}\right)$.

A5. Useful identities

Lemma 1. For a random vector z that is distributed according to a multivariate normal distribution: $z \sim \mathcal{N}(0, C)$ and for symmetric matrices A and B.

$$
cov\left(\boldsymbol{z}^T\boldsymbol{A}\boldsymbol{z}, \boldsymbol{z}^T\boldsymbol{B}\boldsymbol{z}\right) = 2\mathrm{tr}[\boldsymbol{C}\boldsymbol{A}\boldsymbol{C}\boldsymbol{B}].
$$

Thus

$$
\begin{aligned}
\mathbb{E}\left[\left(\boldsymbol{z}^T\boldsymbol{A}\boldsymbol{z}\right)\left(\boldsymbol{z}^T\boldsymbol{B}\boldsymbol{z}\right)\right] &= 2\mathrm{tr}[\boldsymbol{C}\boldsymbol{A}\boldsymbol{C}\boldsymbol{B}] + \mathbb{E}\left[\left(\boldsymbol{z}^T\boldsymbol{A}\boldsymbol{z}\right)\right]\mathbb{E}\left[\left(\boldsymbol{z}^T\boldsymbol{B}\boldsymbol{z}\right)\right] \\
&= 2\mathrm{tr}[\boldsymbol{C}\boldsymbol{A}\boldsymbol{C}\boldsymbol{B}] + \mathrm{tr}[\boldsymbol{A}\boldsymbol{C}]\mathrm{tr}[\boldsymbol{B}\boldsymbol{C}].
\end{aligned}
$$

Lemma 2. For two random variables, A and B, where B is either discrete or has support $(0,\infty)$, and $\mathbb{E}[A] = \mu_A$, $\mathbb{E}[B] = \mu_B$.

$$
var\left(\frac{A}{B}\right) \approx \frac{1}{(\mu_B)^2}var(A) + 2\frac{-\mu_A}{(\mu_B)^3}cov(A,B) + \frac{(\mu_A)^2}{(\mu_B)^4}var(B).
$$