bioRxiv preprint doi: https://doi.org/10.1101/522003; this version posted January 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Scalable multi-component linear mixed models with application to SNP heritability estimation

Ali Pazokitoroudi[1], Yue Wu[1], Kathryn S. Burch[2], Kangcheng Hou[3], Bogdan Pasaniuc[4,5,6], and Sriram Sankararaman[*1,5,6]

[1]Department of Computer Science, UCLA, Los Angeles, California
[2]Bioinformatics Interdepartmental Program, UCLA, Los Angeles, California
[3]Department of Computer Science, Zhejiang University, Hangzhou, Zhejiang, China
[4]Department of Pathology and Laboratory Medicine, David Geffen School of Medicine, UCLA, Los Angeles, California
[5]Department of Human Genetics, David Geffen School of Medicine, UCLA, Los Angeles, California
[6]Department of Computational Medicine, David Geffen School of Medicine, UCLA, Los Angeles, California

Abstract

A central question in human genetics is to find the proportion of variation in a trait that can be explained by genetic variation. A number of methods have been developed to estimate this quantity, termed narrow-sense heritability, from genome-wide SNP data. Recently, it has become clear that estimates of narrow-sense heritability are sensitive to modeling assumptions that relate the effect sizes of a SNP to its minor allele frequency (MAF)and linkage disequilibrium (LD) patterns [3]. A principled approach to estimate heritability while accounting for variation in SNP effect sizes involves the application of linear Mixed Models (LMMs) with multiple variance components where each variance component represents the fraction of genetic variance explained by SNPs that belong to a given range of MAF and LD values. Beyond their importance in accurately estimating genome-wide SNP heritability, multiple variance component LMMs are useful in partitioning the contribution of genomic annotations to trait heritability which, in turn, can provide insights into biological processes that are associated with the trait.

Existing methods for fitting multi-component LMMs rely on maximizing the likelihood of the variance components. These methods pose major computational bottlenecks that makes it challenging to apply them to large-scale genomic datasets such as the UK Biobank which contains half a million individuals genotyped at tens of millions of SNPs.

We propose a scalable algorithm, RHE-reg-mc, to jointly estimate multiple variance components in LMMs. Our algorithm is a randomized method-of-moments estimator that has a runtime that is observed to scale as $\mathcal{O}(\frac{NMB}{\max(\log_3(N),\log_3(M))}+K^3)$ for N individuals, M SNPs, K variance components, and $B \approx 10$ being a parameter that controls the number of random matrix-vector multiplication. RHE-reg-mc also efficiently computes standard errors. We evaluate the accuracy and scalability of RHE-reg-mc for estimating the total heritability as well as in partitioning heritability. The ability to fit multiple variance components to SNPs partitioned according to their MAF and local LD allows RHE-reg-mc to obtain relatively unbiased estimates of SNP heritability under a wide range of models of genetic architecture. On the UK Biobank dataset consisting of $\approx 300,000$ individuals and $\approx 500,000$ SNPs, RHE-reg-mc can fit 250 variance components, corresponding to genetic variance explained by 1 MB blocks, in ≈ 40 minutes on standard hardware.

1 Introduction

Heritability is a central parameter in understanding the contribution of genetic variation to trait variation [22]. Narrow-sense heritability refers to the maximal proportion of variation in a trait that can be

bioRxiv preprint doi: https://doi.org/10.1101/522003; this version posted January 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

explained by a linear function of genetic variation[22]. Over the last decade, there has been substantial attention focused on estimating narrow-sense heritability from genome-wide SNP genotype data [25]. These SNP heritability estimates are of great interest in understanding the genetic basis of complex traits and can inform strategies for designing future genetic studies. While a number of methods have been developed to estimate SNP heritability [25, 27, 11, 5, 19], recent studies have shown that the estimates of SNP heritability from these methods can be highly sensitive to modeling assumptions such as the joint distribution of the effect sizes at causal variants, their allele frequencies as well as the levels of correlation or linkage disequilibrium (LD) between causal variants and the genotyped SNPs[19, 3]. An observation from these studies is that methods that assume that the effect size of each SNP on a trait comes from the same distribution can yield biased estimates of SNP heritability [3]. One potential strategy to relax the assumption of identically distributed effect sizes consists of parametrizing the distribution of SNP effect sizes in terms of relevant covariates that are expected to influence their distributions. For example, methods that partition SNPs into bins based on minor allele frequency (MAF) and local LD patterns thereby allowing the distribution of effect sizes to vary as a function of MAF and LD tend to be robust to variation in the underlying genetic models [3]. Such partitioning strategies are also motivated by a number of recent studies suggest that the distribution of effect sizes of SNPs on a trait can have a complex relationship with their allele frequencies and LD [3, 4]. Beyond the dependency of SNP effect sizes on MAF and LD, the SNP effect sizes have been observed to vary as a function of a number of genomic annotations such as whether a SNP lies in the protein coding regions or in regions of open chromatin [7, 6, 4]. These observations, together, motivate statistical models that allow for the flexible modeling of genetic effects on phenotype where the genetic effects vary as a function of covariates such as MAF, LD or other genomic annotations.

Linear mixed models (LMMs) are an important class of models that permit the representation of flexible relationships between genetic variation and complex traits, *i.e.*, traits that are modulated by multiple genetic and environmental factors. LMMs have been applied successfully in linkage analysis in family studies, in genomic selection and risk prediction, as well as in association analysis to control for individual relatedness and population stratification.

LMMs [15] have been applied to estimate SNP heritability attributed to genome-wide SNPs (h^2_{SNP}) [25]. In the simplest setting, the LMM is endowed with two parameters, *i.e.*, variance components, corresponding to the phenotypic variance explained by genetic and residual factors respectively. These models, termed *single-component* LMMs, as they employ a single variance component to capture all the genetic effects, assume that the effect of each SNP on the phenotype is drawn independently from the same underlying distribution. However, when this assumption is violated, these single component LMMs are expected to yield biased estimates of SNP heritability. To overcome this limitation, multi-components LMMs have been proposed which assign SNPs into one of several variance components and assume that the effect sizes at SNPs assigned to a given variance component are drawn independently from the same distribution [24]. By binning SNPs according to their MAF and LD and assigning a variance component to each bin, these multi-components LMMs have been shown to yield heritability estimates that are accurate across a range of underlying genetic architectures.

While multi-component LMMs appear to be well-suited to estimate genome-wide SNP heritability, estimating their parameters, *i.e.*, the variance components, is computationally demanding. The most common approach to estimate the variance components is to search for parameter values that maximize the likelihood. Usually a particular form of maximum likelihood, the restricted maximum likelihood (REML) estimator [16], is preferred due to a reduced bias relative to the full maximum likelihood estimator. Computing maximum likelihood or REML estimators, however, can be challenging. Methods for computing maximum likelihood or REML rely on iterative optimization algorithms. These algorithms do not scale well to large data sets like the UK biobank which contains $\approx 100,000$ individuals and over a million SNPs. While a number of algorithmic approaches have been proposed for efficient inference in LMMs, many of these are designed to leverage the specific structure of single-component LMMs and cannot be applied to the multi-component setting [13, 28]. Thus, efficient parameter estimation in multi-component LMMs remains a challenging computational problem.

bioRxiv preprint doi: https://doi.org/10.1101/522003; this version posted January 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

1.1 Our contribution

We propose a fast multi-component variance components estimation algorithm for linear mixed models with many variance components based on a randomized method-of-moments (MoM) estimator. Our estimator can be viewed as a generalization of the classic Haseman-Elston (HE) regression [8] estimator to the multi-component setting as well as of a recently proposed randomized version of HE regression for the single component setting[23].

Being a Method-of-Moments estimator, our proposed estimator is statistically less efficient than REML but it is computationally attractive. Recently, a scalable estimator of variance components for a single-component LMM based on a randomized version of HE regression (RHE-reg), was proposed. This method has a runtime complexity $\mathcal{O}(\frac{NMB}{\max(\log_3(N),\log_3(M))})$ for N individuals and M SNPs and a parameter B that controls the number of random matrix-vector multiplications[23]. In this paper, we extend RHE-reg to the multiple component setting where we assume that SNPs have been classified into one of K different non-overlapping functional categories. Our method, RHE-reg-mc, estimates the variance component of each category in time $\mathcal{O}(\frac{NMB}{\max(\log_3(N),\log_3(M))} + K^3)$

The time complexity of RHE-reg-mc scales with the size of the genotype matrix as long as the number of components or partitions K is less than $(MN)^{1/3}$ where M and N are the number of SNPs and individuals respectively. For example, in the full set of UK Biobank genotypes ($M \approx 10^6$, $N \approx 10^5$), this allows us to define 10^4 variance components while maintaining scalability.

We apply RHE-reg-mc to the problem of estimating genome-wide SNP heritability when the underlying genetic architecture assumes that the effect sizes are a function of the MAF and LD patterns at the SNP. To account for variation in the SNP effect sizes, we bin SNPs according to their MAF and LD patterns (following previous approaches [26] and assign a distinct variance component for each bin. In small scale experiments containing around $14,000$ individuals, we show that RHE-reg-mc yields approximately unbiased estimates of SNP heritability albeit with larger standard errors relative to multi-component REML methods. RHE-reg-mc remained relatively unbiased in large-scale experiments on about $337,000$ individuals. Importantly, RHE-reg-mc was the only method that used individual-level genotypes that could run on these datasets. Our bench-marking experiments show that RHE-reg-mc is around 400 times faster than other state-of-the-art methods such as BOLT-REML on a dataset of $100,000$ individuals and $500,000$ SNPs even when the latter were required to estimate only a single component. Therefore, RHE-reg-mc is the only method that used individual-level genotypes that can run efficiently on large datasets such as full UK Biobank which contains half a million individuals and millions of SNPs. Beyond estimating SNP heritability, we also show that RHE-reg-mc accurately partitions heritability across 1 Mb regions which corresponds to jointly estimating 250 variance components.

2 Methods

2.1 Motivation

Various methods have been developed to estimate narrow sense heritability [1]. According to recent studies [18, 3], our assumptions about the levels of LD, MAF and effect sizes of underlying causal variants have significant effects on the accuracy of estimates of narrow sense heritability. However,methods which partition SNPs into bins based on minor allele frequency (MAF) and linkage disequilibrium (LD) [24] obtain reduced bias in their estimates of heritability.

To understand why the estimation of narrow-sense heritability is sensitive to assumptions about the underlying causal variants, and why single component LMMs are expected to yield biased estimates of SNP heritability, consider the following single-component linear mixed model:

$$\begin{aligned} \boldsymbol{y}|\boldsymbol{\epsilon},\boldsymbol{\beta} &= \boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\epsilon} \\ \boldsymbol{\epsilon} &\sim \mathcal{N}(\mathbf{0},\sigma_e^2\boldsymbol{I}_N) \\ \boldsymbol{\beta} &\sim \mathcal{N}(\mathbf{0},\frac{\sigma_g^2}{M}\boldsymbol{I}_M) \end{aligned} \tag{1}$$

bioRxiv preprint doi: https://doi.org/10.1101/522003; this version posted January 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Here $\boldsymbol{y}$ is a N-vector of phenotypes which is centered, and $\boldsymbol{X}$ is a $N \times M$ matrix of standardized genotypes obtained by centering and scaling each column of the genotype matrix $\boldsymbol{G}$ where $g_{i,j} \in \{0, 1, 2\}$ denotes the number of minor alleles carried by individual i at SNP j, N and M are the number of individuals and SNPs respectively. $\boldsymbol{\beta}$ is a M-vector of SNP effect sizes. σ_e^2 is the residual variance while σ_g^2 is the variance component corresponding to the M SNPs. In this model, the SNP heritability is defined as $h^2_{SNP} = \frac{\sigma_g^2}{\sigma_g^2+\sigma_e^2}$. This model assumes that the variance of the effect size is the same across all SNPs. When effect sizes are coupled with MAF or LD, the mismatch between the model assumptions and data can lead to biased heritability estimates.

To tackle this problem, the LMM can be extended to include multiple components as follows:

$$\begin{aligned} \boldsymbol{y}|\boldsymbol{\epsilon}, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K &= \sum_{k=1}^{K} \boldsymbol{X}_k \boldsymbol{\beta}_k + \boldsymbol{\epsilon} \qquad (2) \\ \boldsymbol{\epsilon} &\sim \mathcal{N}(\boldsymbol{0}, \sigma_e^2 \boldsymbol{I}_N) \\ \boldsymbol{\beta}_k &\sim \mathcal{N}(\boldsymbol{0}, \frac{\sigma_k^2}{M_k} \boldsymbol{I}_{M_k}), k \in \{1, \ldots, K\} \qquad (3) \end{aligned}$$

Here each of the M SNPs is assigned to one of K non-overlapping categories. Each category k contains M_k SNPs, $k \in \{1, \ldots, K\}$, $\sum_k M_k = M$. Let $\boldsymbol{G}_k$ denote the $N \times M_k$ genotype matrix for category k such that $g_{k,i,j} \in \{0, 1, 2\}$ denotes the number of minor alleles carried by individual i at SNP j in category k. Let $\boldsymbol{X}_k$ be a $N \times M_k$ matrix of standardized genotypes obtained by centering and scaling each column of $\boldsymbol{G}_k$ so that $\sum_i g_{k,i,m} = 0$ and $\sum_k g^2_{k,i,m} = N$ for $m \in \{1, 2, \ldots, M\}$. Let $\boldsymbol{\beta}_k$ be a M_k-vector of SNP effect sizes for the k-th category. In the above model, σ_e^2 is the residual variance, and σ_k^2 is the variance component of the k-th category. In this model, the total SNP heritability is defined as :

$$h^2_{SNP} = \frac{\sum_{k=1}^{K} \sigma_k^2}{(\sum_{k=1}^{K} \sigma_k^2) + \sigma_e^2} \qquad (4)$$

The SNP heritability of category k is defined as:

$$h_k^2 = \frac{\sigma_k^2}{(\sum_{k=1}^{K} \sigma_k^2) + \sigma_e^2}, k \in \{1, \ldots, K\} \qquad (5)$$

The model in Equation 3 has K partitions such that the variance of effect sizes can differ among the partitions. We can partition SNPs into bins based on minor allele frequency (MAF) and local LD patterns thereby allowing the distribution of effect sizes to vary as a function of MAF and LD tend to be robust to variation in the underlying genetic models and achieve unbiased estimation of SNP heritability. However, in this setting we have several challenges. First, can we estimate the parameters of the model efficiently? Second, for a given genotype, what is the minimum value of K required for an accurate estimation of heritability?

2.2 Method-of-moments for estimating multiple variance components

To estimate the variance components of the multi-component LMM, we use a Method-of-Moments (MoM) estimator that searches for parameter values so that the population moments are close to the sample moments. Since $\mathbb{E}[\boldsymbol{y}] = 0$, we derived the MoM estimates by equating the population covariance to the empirical covariance. The population covariance is given by:

$$cov(\boldsymbol{y}) = E[\boldsymbol{y}\boldsymbol{y}^T] - E[\boldsymbol{y}]E[\boldsymbol{y}^T] = \sum_k \sigma_k^2 \boldsymbol{K}_k + \sigma_e^2 \boldsymbol{I}_N \qquad (6)$$

Here $\boldsymbol{K}_k = \frac{\boldsymbol{X}_k \boldsymbol{X}_k^T}{M_k}$ is the genetic relatedness matrix (GRM) computed from all SNPs of k-th category. Using $\boldsymbol{y}\boldsymbol{y}^{\mathrm{T}}$ as our estimate of the empirical covariance, we need to solve the following least squares problem to find the variance components.

bioRxiv preprint doi: https://doi.org/10.1101/522003; this version posted January 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

$$(\tilde{\sigma_1^2}, \ldots, \tilde{\sigma_K^2}, \tilde{\sigma_e^2}) = argmin_{(\sigma_1^2, \ldots, \sigma_K^2, \sigma_e^2)} ||\boldsymbol{y}\boldsymbol{y}^T - \sum_k \sigma_k^2 \boldsymbol{K}_k + \sigma_e^2 \boldsymbol{I}||_F^2 \tag{7}$$

It is not hard to see that the MoM estimator satisfies the following normal equations:

$$\begin{bmatrix} \boldsymbol{T} & \boldsymbol{b} \\ \boldsymbol{b}^T & N \end{bmatrix} \begin{bmatrix} \tilde{\sigma_1^2} \\ \vdots \\ \tilde{\sigma_k^2} \\ \tilde{\sigma_e^2} \end{bmatrix} = \begin{bmatrix} \boldsymbol{c} \\ \boldsymbol{y}^T\boldsymbol{y} \end{bmatrix} \tag{8}$$

Here $\boldsymbol{T}$ is a $K \times K$ matrix with entries $T_{k,l} = tr(\boldsymbol{K}_k \boldsymbol{K}_l), k, l \in \{1, \ldots, K\}$, $\boldsymbol{b}$ is a K-vector with entries $b_k = tr(\boldsymbol{K}_k) = N$ (because $\boldsymbol{X}_k$s is standardized), and $\boldsymbol{c}$ is a K-vector with entries $c_k = \boldsymbol{y}^T \boldsymbol{K}_k \boldsymbol{y}$.

Every GRM $\boldsymbol{K}_k$ can be computed in time $\mathcal{O}(N^2 M_k)$ and $\mathcal{O}(N^2)$ memory. Given K GRMs, the quantities $T_{k,l}$, c_k, $k, l \in \{1, \ldots, K\}$, can be computed in $\mathcal{O}(NM)$. Given the quantities $T_{k,l}$, c_k, the normal equation 8 can be solved in $\mathcal{O}(K^3)$. Therefore, the total time complexity for estimating the variance components is $\mathcal{O}(N^2 M + K^3)$.

2.2.1 RHE-reg-mc: Randomized estimator of multiple variance components

The key bottleneck in solving the normal equation 8 is the computation of $T_{k,l}$, $k, l \in \{1, \ldots, K\}$ which takes $\mathcal{O}(N^2 M)$. Instead of computing the exact value of $T_{k,l}$, we use Hutchinson's estimator of the trace [9]. This estimator uses the fact that for a given $N \times N$ matrix $\boldsymbol{C}$, $\boldsymbol{z}^T \boldsymbol{C} \boldsymbol{z}$ is an unbiased estimator of $tr(\boldsymbol{C})$ ($E[\boldsymbol{z}^T \boldsymbol{C} \boldsymbol{z}] = tr[\boldsymbol{C}]$) where $\boldsymbol{z}$ be a random vector with mean zero and covariance $\boldsymbol{I}_N$. Hence, we can estimate the values $T_{k,l}$, $k, l \in \{1, \ldots, K\}$ as follows:

$$T_{k,l} = tr(\boldsymbol{K}_k \boldsymbol{K}_l) \approx \widehat{T_{k,l}} = \frac{1}{B} \frac{1}{M_k M_l} \sum_b \boldsymbol{z}_b^T \boldsymbol{X}_k \boldsymbol{X}_k^T \boldsymbol{X}_l \boldsymbol{X}_l^T \boldsymbol{z}_b \tag{9}$$

Here $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_B$ are B independent random vectors with zero mean and covariance $\boldsymbol{I}_N$. In our method, we draw these random vectors independently from a standard normal distribution. Note that computing $T_{k,l}$ by using the unbiased estimator involves four matrix-vector multiplications which is repeated B times. Therefore, the total running time for estimating the values $T_{k,l}$ is $\mathcal{O}(NMB)$.

Moreover, we can leverage the structure of the genotype matrix which only contains entries in $\{0, 1, 2\}$. For a fixed genotype matrix $\boldsymbol{X}_k$, we can improve the per iteration time complexity of matrix-vector multiplication from $\mathcal{O}(NM)$ to $\mathcal{O}(\frac{NM}{max(\log_3 N, \log_3 M)})$ by using the Mailman algorithm [12]. Solving the normal equations takes $\mathcal{O}(K^3)$ time so that the overall time complexity of our algorithm is $\mathcal{O}(\frac{NMB}{\max(\log_3(N), \log_3(M))} + K^3)$.

2.3 Computing the Standard Errors of the estimates

We obtain standard errors for RHE-reg-mc using a block jackknife [10]. The jackknife is a useful resampling technique for estimating the variance and bias of an estimator [21]. A jackknife subsample is created by leaving out a subset of observations from a dataset. The jackknife estimate of a parameter can be found by estimating the parameter for each subsample omitting the i-th jackknife block. A naive way to compute jackknife estimate requires computing the estimator of the parameters for every subsample. For instance, in our problem, if we define J jackknife blocks, then we need to run RHE-reg-mc for every subsample which takes $\mathcal{O}(J(\frac{NMB}{\max(\log_3(N), \log_3(M))} + K^3))$. We propose an efficient way to compute the jackknife estimate in time $\mathcal{O}(\frac{NMB}{\max(\log_3(N), \log_3(M))} + JK^3)$.

Let $\boldsymbol{X}$ be a $N \times M$ matrix of standardized genotypes where N and M are the number of individuals and SNPs respectively. To generate J jackknife subsamples, we partition X into J non-overlapping blocks $\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(J)}$ such that $\boldsymbol{X} = [\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}, \ldots, \boldsymbol{X}^{(J)}]$. Note that for every j, $\boldsymbol{X}^{(j)}$ is a $N \times M_j$ matrix where M_j is the number of SNPs in the j-th block.

We create the j-th jackknife subsample by removing the j-th block $\boldsymbol{X}^{(j)}$ from $\boldsymbol{X}$. To estimate the variance components of the j-th jackknife subsample, we need to compute the corresponding quantities of

bioRxiv preprint doi: https://doi.org/10.1101/522003; this version posted January 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

the jth subsample in the normal equations 8. Let $\boldsymbol{K}_k^{(-j)}$ be the GRM of the k-th partition which is created by removing the j-th block $\boldsymbol{X}^{(j)}$ from $\boldsymbol{X}$ where $k \in \{1, \ldots, K\}$, $j \in \{1, \ldots, J\}$. In Appendix A we show that we can compute $tr(\widehat{\boldsymbol{K}_k^{(-j)}\boldsymbol{K}_l^{(-j)}})$ and $\boldsymbol{y}^T\boldsymbol{K}_i^{(-j)}\boldsymbol{y}$, for all $k, l \in \{1, \ldots, K\}$, $j \in \{1, \ldots, J\}$, in time $\mathcal{O}(\frac{NMB}{\max(\log_3(N), \log_3(M))})$ which does not effect the total running time of the algorithm.

Therefore, for every jackknife subsample, we can estimate the corresponding variance components in $\mathcal{O}(K^3)$ by solving the corresponding normal equations 8. Given the estimates of variance components for each jackknife subsamples, we can compute jackknife estimate of the variance, estimate the bias, and bias-corrected jackknife estimate of the variance components.

2.4 Including covariates

We can extend our model in Equation 2 to include covariates as follows:

$$\boldsymbol{y}|\boldsymbol{\epsilon}, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k = \boldsymbol{W}\boldsymbol{\alpha} + \sum_k \boldsymbol{X}_k\boldsymbol{\beta}_k + \boldsymbol{\epsilon} \tag{10}$$

Here $\boldsymbol{W}$ is a $N \times C$ matrix of covariates while $\boldsymbol{\alpha}$ is a C-vector of fixed effects. In Appendix B, we show that in this setting, we need to solve the following normal equations to estimate the variance components.

$$\begin{bmatrix} \boldsymbol{T} & \boldsymbol{b} \\ \boldsymbol{b}^T & N - C \end{bmatrix} \begin{bmatrix} \sigma_1^2 \\ \vdots \\ \sigma_k^2 \\ \sigma_e^2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{c} \\ \boldsymbol{y}^T\boldsymbol{V}\boldsymbol{y} \end{bmatrix} \tag{11}$$

Here $\boldsymbol{V} = \boldsymbol{I}_N - \boldsymbol{W}(\boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{W}^T$ and $\boldsymbol{T}$ is a $K \times K$ matrix where $T_{k,l} = tr(\boldsymbol{K}_k\boldsymbol{V}\boldsymbol{K}_l\boldsymbol{V})$, and b is a $K-$vector where $b_k = tr(\boldsymbol{V}\boldsymbol{K}_k)$, and $\boldsymbol{c}$ is a K- vector where $c_k = \boldsymbol{y}^T\boldsymbol{V}\boldsymbol{K}_k\boldsymbol{V}\boldsymbol{y}$. Commonly, the number of covariates C is small (tens to hundreds) so that including covariates does not significantly affect the computational cost. The cost of computing of the elements of the normal equations 11 includes the cost of inverting $\boldsymbol{W}^T\boldsymbol{W}$ which is a $C \times C$ matrix and multiplying $\boldsymbol{W}$ by a real-valued N-vector which can be done in $\mathcal{O}(C^3 + NC)$.

3 Results

3.1 Estimating total heritability by partitioning based on MAF and LD

3.1.1 Simulations

We performed simulations to compare the performance of RHE-reg-mc with several state-of-the-art methods for heritability estimation that cover the spectrum of methods that have been proposed. These methods include single-component as well as multi-component LMMs which use likelihood maximization for parameter estimation. These methods require access to individual-level genotypes and phenotypes. We also compared to methods that only require access to summary statistics and are typically quite scalable. GCTA and BOLT-REML estimate heritability by maximizing the restricted maximum likelihood (REML). BOLT-REML is a computationally efficient approximate method to compute the REML estimator. GCTA-ldms is the extension of GCTA to a multi-component LMM where the variance components are typically defined by binning SNPs according to their MAF as well as local LD. LDAK is similar to GCTA, except that it assumes allelic effects are a function of LD scores. Among the summary statistic methods, LD score regression (LDSC) uses the slope from the GWAS χ^2 statistics regressed on the LD scores to estimate the h^2_{SNP}. Stratified LD score method (S-LDSC) is an extension of LDSC for partitioning heritability from summary statistics. SumHer is the summary statistic analog of LDAK [17].

We considered two simulation settings. The small-scale setting was designed so that we could compare the accuracies of our method to state-of-the-art methods on the same data. In this setting, we simulated phenotypes from a subsampled set of genotypes from the UK Biobank [20]. Specifically, we chose $M = 14,000$ SNPs from chromosome 1 of the UK Biobank Axiom array and a subset of $N = 9,000$ individuals. In the

bioRxiv preprint doi: https://doi.org/10.1101/522003; this version posted January 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

large-scale simulation setting, we simulated phenotypes for the full set of UK Biobank genotypes consisting of $M = 590,000$ array SNPs and $N = 337,000$ unrelated white British individuals.

We simulated phenotypes from genotypes using the following model for the genetic architecture:

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \\ \boldsymbol{\epsilon}|\sigma_e^2 &\sim \mathcal{N}(\mathbf{0}, \sigma_e^2 \boldsymbol{I}_N) \\ \boldsymbol{\beta}_m &\sim \mathcal{N}(\mathbf{0}, c[f_m(1-f_m)]^a w_m^b), m \in \{1, \ldots, M\} \end{aligned} \tag{12}$$

where $a \in \{0, 0.75\}$,$b \in \{0, 1\}$, c is a constant, β_m, f_m and w_m are the effect size, the minor allele frequency and LDAK score of m^{th} SNP respectively. The LD score of a SNP is defined to be the sum of the squared correlation of the SNP with all other SNPs that lie within a specific distance, and the LDAK score of a SNP is computed based on local levels of LD such that the LDAK score tends to be higher for SNPs in regions of low LD[19]. The above models relating genotype to phenotype are commonly used in methods for estimating SNP heritability: the GCTA Model (when $a = b = 0$ in Equation 12), which is used by the software GCTA [26] and LD Score regression (LDSC) [2], and the LDAK Model (where $a = 0.75, b = 1$ in Equation 12) used by software LDAK [19]. Moreover, under each model, we varied the proportion and minor allele frequency (MAF) of causal variants (CVs). Proportion of causal variants set to be either 100% or 1%, and MAF of causal variants drawn uniformly from $[0, 1]$ or $[0.01, 0.05]$ to consider genetic architectures that are either infinitesimal or sparse as well genetic architectures that include a mixture of common and rare SNPs as well as one that includes only common SNPs. The GCTA Model assumes that heritability is independent of LD, while the LDAK Model assumes that heritability varies according to local levels of LD[18].

We generated 100 sets of simulated phenotypes for each setting of parameters and report accuracies averaged over these 100 sets.

3.1.2 Accuracy

We compared the accuracy of RHE-reg-mc with the most popular single component methods such as BOLT-REML [14], GCTA [26], LDAK [19], LDSC [2], SumHer [17], as well as multi-components methods such as GCTA-ldms and S-LDSC (BOLT-REML can also be estimate multi-component variance components but we did not explore this option). In the small scale setting, we compared RHE-reg-mc with BOLT-REML, GCTA, GCTA-ldms, and LDAK, each of which use individual-level phenotypes and genotypes[19, 2, 26, 17]. In the large scale setting, we only compared RHE-reg-mc with methods that use summary statistics like SumHer,LDSC, and S-LDSC because the existing methods based on individual-level phenotypes and genotype are not scalable to large data sets. For the multi-component methods, we applied GCTA-ldms by binning SNPs into 8 bins based on 2 bins for MAF (MAF$<$ 0.05, MAF$>$ 0.05), MAF refers to the frequency at which the second most common allele occurs in a SNP, and 4 bins based on quartiles of the LD score of a SNP. The LD score of a SNP is defined to be the sum of the squared correlation of the SNP with all other SNPs that lie within a specific distance. We used the default LD score computation that is used by GCTA. For RHE-reg-mc, we used 16 bins formed by the combination of 4 bins based on MAF (MAF$<$ 0.009,0.009 $<$MAF$<$ 0.011,0.011 $<$MAF$<$ 0.05,0.05 $<$MAF$<$ 0.5) as well as 4 bins based on quartiles of the LDAK score of a SNP. The LDAK score of a SNP is computed based on local levels of LD such that the LDAK score tends to be higher for SNPs in regions of low LD[19].

Figure 1 and Table 1 show that both multi-component methods, *i.e.*, GCTA-ldms and RHE-reg-mc, have the least bias across the settings considered. RHE-reg-mc tends to have larger standard errors relative to GCTA-ldms consistent with the lower statistical efficiency of method-of-moments estimators relative to REML estimators. In the large-scale setting (Figure 2 and Table 2), RHE-reg-mc has reduced bias as well as low standard errors compared to other methods.

bioRxiv preprint doi: https://doi.org/10.1101/522003; this version posted January 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

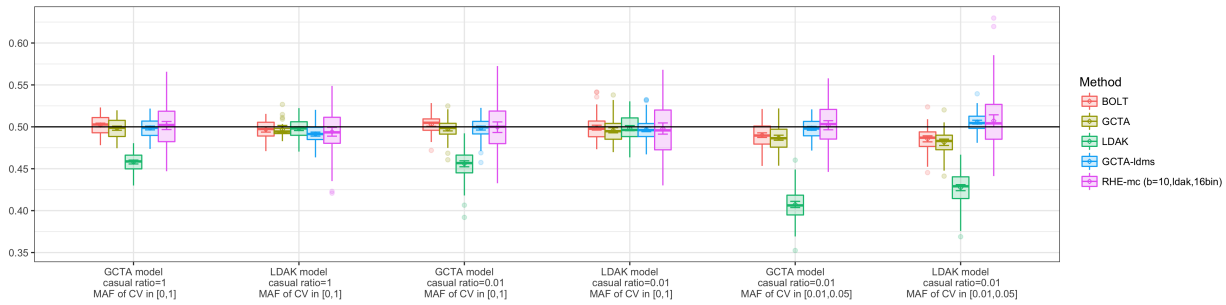

Figure 1: **RHE-reg-mc is relatively unbiased in small scale simulated data** ($M = 14,000$ array SNPs and $N = 9,000$ individuals): We ran 100 replicates where the true heritability of the phenotype is 0.5. We compared methods for heritability estimation under both GCTA and LDAK models with different proportions of causal variants as well as MAF distributions. BOLT-REML (run with single component), GCTA, and LDAK are single component methods. GCTA-ldms is a multi-component method that is applied by binning SNPs to 8 bins based on 2 MAF bins and 4 bins based on quartiles of LD scores. RHE-reg-mc is our proposed multi-component method that is applied by binning SNPs to 16 bins based on 4 MAF bins and 4 bins based on LDAK weights, and $b = 10$ random vector. GCTA-ldms and RHE-reg-mc are relatively unbiased.

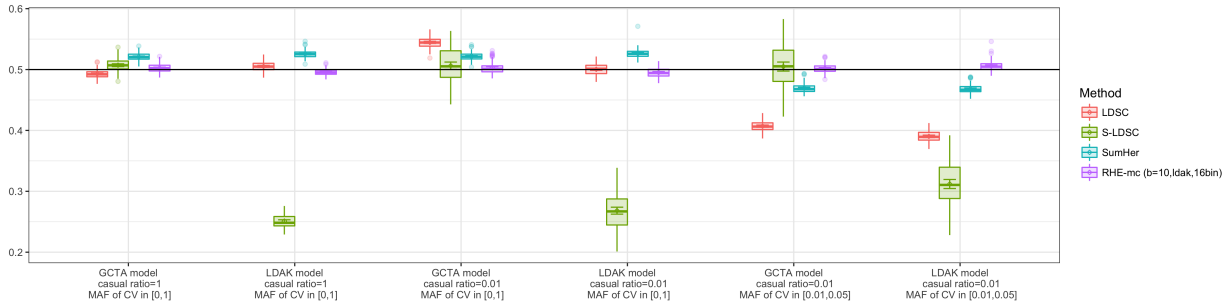

Figure 2: **RHE-reg-mc is relatively unbiased in large scale simulated data** (M=590K array SNPs and N=337K individuals): We ran 100 replicates where the true heritability of the phenotype is 0.5. We compared methods for heritability estimation under both GCTA and LDAK models with different proportions of causal variants and MAF. LDSC, S-LDSC and SumHer methods are based on summary statistics. RHE-reg-mc is an individual-level method that is applied by binning SNPs to 16 bins based on 4 MAF bins and 4 LDAK bins.RHE-reg-mc is relatively unbiased across the settings.

Method	ave(|bias|),GCTA model	ave(sd),GCTA model	ave(|bias|),LDAK model	ave(sd),LDAK model
RHE-reg-mc(b=10,16bin)	**0.001**	0.027	**0.005**	0.031
GCTA-ldms	0.001	**0.011**	0.005	**0.012**
LDAK	0.059	0.015	0.025	0.014
GCTA	0.005	0.012	0.008	0.013
BOLT	0.004	0.011	0.006	0.013

Table 1: The average of standard errors and the average of absolute values of bias across all settings are reported for each method for small scale data.

bioRxiv preprint doi: https://doi.org/10.1101/522003; this version posted January 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Method	ave(|bias|),GCTA model	ave(sd),GCTA model	ave(|bias|),LDAK model	ave(sd),LDAK model
RHE-reg-mc(b=10,16bin)	**0.002**	0.007	**0.004**	**0.007**
SumHer	0.024	**0.006**	0.027	0.007
S-LDSC	0.006	0.026	0.222	0.02
LDSC	0.048	0.008	0.038	0.008

Table 2: The average of standard errors and the average of absolute values of bias across all settings are reported for each method for large scale data.

3.2 Computational Efficiency

We compared the running time and memory usage of RHE-reg-mc method with GCTA [26] and BOLT-REML [14]. In this comparison, we used the UK Biobank genotypes consisting of around 500, 000 SNPs over different sample sizes. For each data set, we ran RHE-reg-mc with $B = 10$ random vectors and 22 bins (SNPs partitioned based on chromosomes). We ran GCTA and BOLT-REML for a single variance component. All computations were restricted to a single core on a standard compute machine.

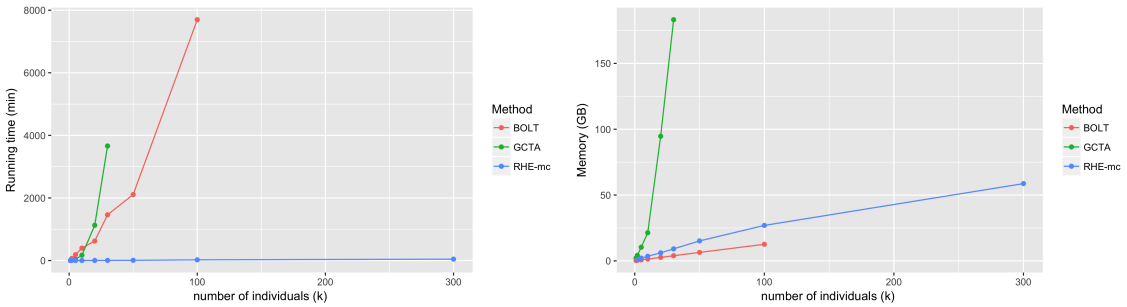

Figure 3: **RHE-reg-mc is efficient**: We compare the runtime and memory usage of methods for heritability estimation. We vary the number of samples while fixing the number of SNPs to 500, 000. We run RHE-reg-mc with 22 bins defined based on chromosomes. We ran GCTA and BOLT-REML assuming a single variance component as multiple variance component versions tend to be more computationally intensive.

Figure 3 shows that we could not run single-component GCTA to sample sizes beyond 50, 000 due to memory constraints. Single-component BOLT-REML took about 5 days to run on 100, 000 individuals while the computation of RHE-reg-mc was about 20 minutes. Our bench-marking experiments show that RHE-reg-mc is around 400 times faster than other state-of-the-art methods such as BOLT-REML on a dataset of $100k$ individuals and $500k$ SNPs. Extrapolating this result, we expect that RHE-reg-mc could run on large datasets such as the full UK Biobank which contains half a million individuals genotyped at tens of millions of SNPs efficiently. The memory usage of RHE-reg-mc is linear with respect to sample size. The running time and accuracy of RHE-reg-mc relies on the choice of the number of random vectors B. In practice, it turns out that the estimator is highly accurate with a small $B \approx 10$ across all datasets analyzed.

3.3 Partitioning heritability

To examine the ability of RHE-reg-mc to partition the heritability explained across genomic regions, we partition SNPs into 1 MB regions so that the first 50 bins each explains 1% of the variance ($h^2_{SNP} = 0.5$) while the rest of the bins have zero heritability. Figure 4 shows that RHE-reg-mc accurately estimates heritability in each partition on 100, 000 individuals and 460, 000 SNPs. Further, RHE-reg-mc computed the partitioned heritability in 20 minutes.

bioRxiv preprint doi: https://doi.org/10.1101/522003; this version posted January 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

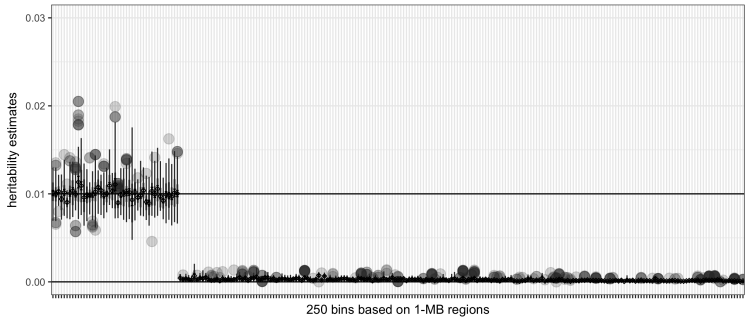

Figure 4: **RHE-reg-mc accurately partitions heritability**: We simulated phenotypes from 460,000 SNPs of UK Biobank genotype over 100,000 individuals. Each of the first 50 bins has a heritability of 1% ($h^2_{SNP} = 0.5$) while the rest of the bins have zero heritability.

3.4 Application to phenotypes in the UK Biobank

Finally, we used RHE-reg-mc to partition the heritability of three phenotypes (trunk fat percentage, diastolic blood pressure, and systolic blood pressure) on the UK Biobank dataset consisting of around 500,000 SNPs and 300,000 individuals . Partitioning SNPs according to 22 chromosomes, we see that longer chromosomes length have higher heritability consistent with studies in other traits such as height [25] (Figure 5).

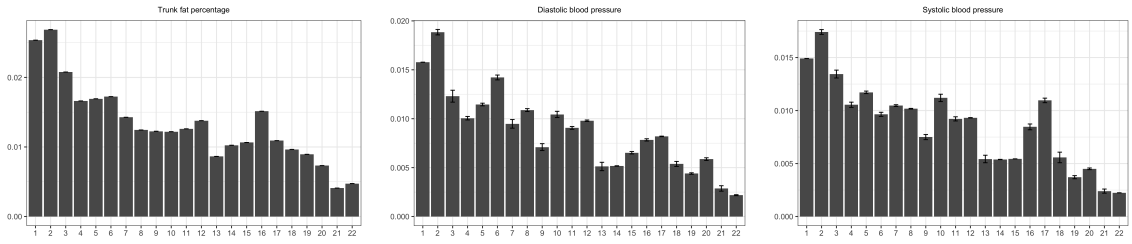

Figure 5: **RHE-reg-mc heritability estimates for trunk fat percentage, diastolic blood pressure, and systolic blood pressure in the UK Biobank**

4 Discussion

We have described RHE-reg-mc, a scalable estimator of multiple variance components in linear mixed models. RHE-reg-mc uses a randomized Method-of-Moments estimator to estimate a large number of variance components on datasets with hundreds of thousands of individuals and SNPs in less than an hour. The ability to estimate multiple variance components efficiently is useful both in obtaining unbiased genome-wide SNP heritability estimates as well as in heritability partitioning analyses.

There are several ways to further improve the runtime and memory usage of RHE-reg-mc method. First, in the context of multiple phenotypes measured for a fixed genotype matrix, the matrix T in normal equation 8 is the same for every given phenotype. Hence, we just need to compute the sufficient statistics of matrix $\boldsymbol{T}$ once in the multiple phenotype setting. Second, RHE-reg-mc can perform its computation in a streaming version which can lead to a highly memory efficient implementation.

5 Acknowledgments

This research was conducted using the UK Biobank Resource under applications 33127 and 33297. We thank the participants of UK Biobank for making this work possible. We thank Rob Brown for feedback on this manuscript. This work was funded by NIH grants R01HG009120 (B.P. and K.S.B.), R35GM125055 (S.S.), an Alfred P. Sloan Research Fellowship (S.S.), and a NSF grant III-1705121 (Y.W. and S.S.).

bioRxiv preprint doi: https://doi.org/10.1101/522003; this version posted January 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

References

[1] Laura Almasy and John Blangero. Multipoint quantitative-trait linkage analysis in general pedigrees. *The American Journal of Human Genetics*, 62(5):1198–1211, 1998.

[2] Brendan K Bulik-Sullivan, Po-Ru Loh, Hilary K Finucane, Stephan Ripke, Jian Yang, Nick Patterson, Mark J Daly, Alkes L Price, Benjamin M Neale, Schizophrenia Working Group of the Psychiatric Genomics Consortium, et al. Ld score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature genetics*, 47(3):291, 2015.

[3] Luke M Evans, Rasool Tahmasbi, Scott I Vrieze, Gonçalo R Abecasis, Sayantan Das, Steven Gazal, Douglas W Bjelland, Teresa R Candia, Michael E Goddard, Benjamin M Neale, et al. Comparison of methods that use whole genome data to estimate the heritability and genetic architecture of complex traits. *Nature genetics*, 50(5):737, 2018.

[4] Hilary K Finucane, Brendan Bulik-Sullivan, Alexander Gusev, Gosia Trynka, Yakir Reshef, Po-Ru Loh, Verneri Anttila, Han Xu, Chongzhi Zang, Kyle Farh, et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature genetics*, 47(11):1228, 2015.

[5] David Golan, Eric S Lander, and Saharon Rosset. Measuring missing heritability: inferring the contribution of common variants. *Proceedings of the National Academy of Sciences*, 111(49):E5272–E5281, 2014.

[6] Alexander Gusev, Gaurav Bhatia, Noah Zaitlen, Bjarni J Vilhjalmsson, Dorothée Diogo, Eli A Stahl, Peter K Gregersen, Jane Worthington, Lars Klareskog, Soumya Raychaudhuri, et al. Quantifying missing heritability at known gwas loci. *PLoS genetics*, 9(12):e1003993, 2013.

[7] Alexander Gusev, S Hong Lee, Gosia Trynka, Hilary Finucane, Bjarni J Vilhjálmsson, Han Xu, Chongzhi Zang, Stephan Ripke, Brendan Bulik-Sullivan, Eli Stahl, et al. Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *The American Journal of Human Genetics*, 95(5):535–552, 2014.

[8] JK Haseman and RC Elston. The investigation of linkage between a quantitative trait and a marker locus. *Behavior genetics*, 2(1):3–19, 1972.

[9] MF Hutchinson. A stochastic estimator of the trace of the influence matrix for laplacian smoothing splines. *Communications in Statistics-Simulation and Computation*, 18(3):1059–1076, 1989.

[10] Hans R Kunsch. The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics*, 17(3):1217–1241, 1989.

[11] Sang Hong Lee, Naomi R Wray, Michael E Goddard, and Peter M Visscher. Estimating missing heritability for disease from genome-wide association studies. *The American Journal of Human Genetics*, 88(3):294–305, 2011.

[12] Edo Liberty and Steven W Zucker. The mailman algorithm: A note on matrix–vector multiplication. *Information Processing Letters*, 109(3):179–182, 2009.

[13] Christoph Lippert, Gerald Quon, Eun Yong Kang, Carl M Kadie, Jennifer Listgarten, and David Heckerman. The benefits of selecting phenotype-specific variants for applications of mixed models in genomics. *Scientific reports*, 3:1815, 2013.

[14] Po-Ru Loh, Gaurav Bhatia, Alexander Gusev, Hilary K Finucane, Brendan K Bulik-Sullivan, Samuela J Pollack, Teresa R de Candia, Sang Hong Lee, Naomi R Wray, Kenneth S Kendler, et al. Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. *Nature genetics*, 47(12):1385, 2015.

[15] Charles E McCulloch and Shayle R Searle. *Generalized, linear, and mixed models*. John Wiley & Sons, 2004.

bioRxiv preprint doi: https://doi.org/10.1101/522003; this version posted January 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

[16] H Desmond Patterson and Robin Thompson. Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58(3):545–554, 1971.

[17] Doug Speed and David Balding. Better estimation of snp heritability from summary statistics provides a new understanding of the genetic architecture of complex traits. *bioRxiv*, page 284976, 2018.

[18] Doug Speed, Na Cai, Michael R Johnson, Sergey Nejentsev, David J Balding, UCLEB Consortium, et al. Reevaluation of snp heritability in complex human traits. *Nature genetics*, 49(7):986, 2017.

[19] Doug Speed, Gibran Hemani, Michael R Johnson, and David J Balding. Improved heritability estimation from genome-wide snps. *The American Journal of Human Genetics*, 91(6):1011–1021, 2012.

[20] Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, et al. Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine*, 12(3):e1001779, 2015.

[21] John Tukey. Bias and confidence in not quite large samples. *Ann. Math. Statist.*, 29:614, 1958.

[22] Peter M Visscher, William G Hill, and Naomi R Wray. Heritability in the genomics era: concepts and misconceptions. *Nature reviews genetics*, 9(4):255, 2008.

[23] Yue Wu and Sriram Sankararaman. A scalable estimator of snp heritability for biobank-scale data. *bioRxiv*, page 294470, 2018.

[24] Jian Yang, Andrew Bakshi, Zhihong Zhu, Gibran Hemani, Anna AE Vinkhuyzen, Sang Hong Lee, Matthew R Robinson, John RB Perry, Ilja M Nolte, Jana V van Vliet-Ostaptchouk, et al. Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nature genetics*, 47(10):1114, 2015.

[25] Jian Yang, Beben Benyamin, Brian P McEvoy, Scott Gordon, Anjali K Henders, Dale R Nyholt, Pamela A Madden, Andrew C Heath, Nicholas G Martin, Grant W Montgomery, et al. Common snps explain a large proportion of the heritability for human height. *Nature genetics*, 42(7):565, 2010.

[26] Jian Yang, S Hong Lee, Michael E Goddard, and Peter M Visscher. Gcta: a tool for genome-wide complex trait analysis. *The American Journal of Human Genetics*, 88(1):76–82, 2011.

[27] Xiang Zhou. A unified framework for variance component estimation with summary statistics in genome-wide association studies. *The annals of applied statistics*, 11(4):2027, 2017.

[28] Xiang Zhou and Matthew Stephens. Genome-wide efficient mixed-model analysis for association studies. *Nature genetics*, 44(7):821, 2012.

bioRxiv preprint doi: https://doi.org/10.1101/522003; this version posted January 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

A Computing jackknife standard errors

For simplicity, here we explain how we can estimate the corresponding parameters of jackknife subsamples when we have only one variance component (one partition). It is easy to see how this approach can be generalized to the multiple variance components setting.

Let $\boldsymbol{X}$ be a $N \times M$ matrix of standardized genotypes where N and M are the number of individuals and SNPs respectively. We partition $\boldsymbol{X}$ to J non-overlapping blocks $\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(J)}$ such that $\boldsymbol{X} = [\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(J)}]$. Note that for every j, $\boldsymbol{X}_j$ is a $N \times M_j$ matrix where M_j is the number of SNPs in the j-th block.

Let $\boldsymbol{K} = \frac{1}{M}\boldsymbol{X}\boldsymbol{X}^T$ be the GRM. Moreover, Let $\boldsymbol{K}^{(-j)}$ be the GRM of the subsample created by removing the j-th block from $\boldsymbol{X}$. We will show how the values $y^T\boldsymbol{K}^{(-j)}y$ and $tr({\boldsymbol{K}^{(-j)}}^2)$ can be computed for every $j \in \{1, \ldots, J\}$ efficiently.

$$\boldsymbol{X}^T\boldsymbol{y} = [{\boldsymbol{X}^{(1)}}^T\boldsymbol{y}, \ldots, {\boldsymbol{X}^{(J)}}^T\boldsymbol{y}]^T \tag{13}$$

$$(X^T\boldsymbol{y})^T(X^T\boldsymbol{y}) = ({\boldsymbol{X}^{(1)}}^T\boldsymbol{y})^2 + \ldots + ({\boldsymbol{X}^{(J)}}^T\boldsymbol{y})^2 \tag{14}$$

Therefore, for every j we have :

$$\boldsymbol{y}^T\boldsymbol{K}^{(-j)}\boldsymbol{y} = \frac{1}{M_j}[(\boldsymbol{X}^T\boldsymbol{y})^T(\boldsymbol{X}^T\boldsymbol{y}) - ({\boldsymbol{X}^{(j)}}^T\boldsymbol{y})^2] \tag{15}$$

$$\boldsymbol{y}^T\boldsymbol{K}\boldsymbol{y} = \frac{1}{M}\boldsymbol{y}^T\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{y} = \frac{1}{M}(X^T\boldsymbol{y})^T(X^T\boldsymbol{y}) \tag{16}$$

Therefore, we can compute $\boldsymbol{y}^T\boldsymbol{K}^{(j)}\boldsymbol{y}$, for every j, and $\boldsymbol{y}^T\boldsymbol{K}\boldsymbol{y}$ in $\mathcal{O}(\frac{BNM}{max(\log_3 N, \log_3 M)})$.

We rewrite $\widehat{tr(\boldsymbol{K}^2)}$ as:

$$\widehat{tr(\boldsymbol{K}^2)} = \frac{1}{BM^2}\sum_b (\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{z}_b)^T(\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{z}_b) \tag{17}$$

Assume that $\boldsymbol{v}_b = \boldsymbol{X}^T\boldsymbol{z}_b$, then we have:

$$\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{z}_b = \boldsymbol{X}\boldsymbol{v}_b = (\boldsymbol{v}_b^T\boldsymbol{X}^T)^T \tag{18}$$

Recall that $\boldsymbol{v}_b$ is a vector of length M. We partition v_b to J sub-vectors corresponding to the partitions of $\boldsymbol{X}$: $\boldsymbol{v}_{b1}, \ldots, \boldsymbol{v}_{bJ}$ such that $\boldsymbol{v}_b = [\boldsymbol{v}_{b1}, \ldots, \boldsymbol{v}_{bJ}]^T$ and $\boldsymbol{v}_{bj}$ is a sub-vector of length M_j, for every $j \in \{1, \ldots, J\}$.

$$\boldsymbol{X}\boldsymbol{v}_b = \boldsymbol{X}^{(1)}\boldsymbol{v}_{b1} + \ldots + \boldsymbol{X}^{(J)}\boldsymbol{v}_{bJ} \tag{19}$$

Assume that $\boldsymbol{c}_{bj} = \boldsymbol{X}^{(j)}\boldsymbol{v}_{bj}$ for every $j \in \{1, \ldots, J\}$ and $b \in \{1, \ldots, B\}$. Let $C_{total} = \sum_b (c_{b1}^{\mathrm{T}} + \ldots + c_{bJ}^{\mathrm{T}})(c_{b1} + \ldots + c_{bJ})$. We then have:

$$\widehat{tr(\boldsymbol{K}^2)} = \frac{1}{BM^2}C_{total} \tag{20}$$

Let $\boldsymbol{d}_b = \boldsymbol{c}_{b1} + \ldots + \boldsymbol{c}_{bJ}$ for every $b \in B$. Now we can compute $\widehat{tr({\boldsymbol{K}^{(-j)}}^2)}$ for every $j \in \{1, \ldots, J\}$ based on the values of $C_{total}, \boldsymbol{d}_b, \boldsymbol{c}_{bj}$ in a way that does not effect the total running time :

$$\begin{aligned}
\widehat{tr({\boldsymbol{K}^{(-j)}}^2)} &= \frac{1}{BM_j^2}[C_{total} - \sum_b (\boldsymbol{c}_{bj}^{\mathrm{T}}\boldsymbol{d}_b + (\boldsymbol{d}_b - \boldsymbol{c}_{bj})^T\boldsymbol{c}_{bj})] \\
&= \frac{1}{BM_j^2}[C_{total} - \sum_b (\boldsymbol{c}_{bj}^{\mathrm{T}}\boldsymbol{d}_b + \boldsymbol{d}_b^T\boldsymbol{c}_{bj} - \boldsymbol{c}_{bj}^{\mathrm{T}}\boldsymbol{c}_{bj})] \\
&= \frac{1}{BM_j^2}[C_{total} - \sum_b (2\boldsymbol{c}_{bj}^{\mathrm{T}}\boldsymbol{d}_b - \boldsymbol{c}_{bj}^{\mathrm{T}}\boldsymbol{c}_{bj})] \\
&= \frac{1}{BM_j^2}[C_{total} - 2\sum_b \boldsymbol{c}_{bj}^{\mathrm{T}}\boldsymbol{d}_b + \sum_b \boldsymbol{c}_{bj}^{\mathrm{T}}\boldsymbol{c}_{bj}]
\end{aligned} \tag{21}$$

bioRxiv preprint doi: https://doi.org/10.1101/522003; this version posted January 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

It is not hard to see that we can compute C_{total} in $\mathcal{O}(BNM)$. Moreover, we can compute c_{bj} for every $b = 1, \ldots, B$ and $j = 1, \ldots, J$ in $\mathcal{O}(NM_j)$, so all c_{bj} can be computed in $\mathcal{O}(BNM)$. We can again use the Mailman algorithm to speed up the vector-matrix multiplications. Therefore, the total running time for computing the variance component of all jackknife subsamples is $\mathcal{O}(\frac{BNM}{max(\log_3 N, \log_3 M)})$.

B Including Covariates

$$\boldsymbol{y}|\boldsymbol{\epsilon}, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k = \boldsymbol{W}\boldsymbol{\alpha} + \sum_k \boldsymbol{X}_k\boldsymbol{\beta}_k + \boldsymbol{\epsilon} \tag{22}$$

Here $\boldsymbol{W}$ is a $N \times C$ matrix of covariates while $\boldsymbol{\alpha}$ is a C-vector of coefficients. It is easy to see that the matrix $\boldsymbol{V} = \boldsymbol{I}_N - \boldsymbol{W}(\boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{W}^T$ is symmetric and idempotent ($\boldsymbol{V}^2 = \boldsymbol{V}$) of rank $N - C$. Therefore, we consider the eigendecomposition of $\boldsymbol{V} = \boldsymbol{E}\boldsymbol{D}\boldsymbol{E}^T$, where $\boldsymbol{D}$ is a diagonal matrix with $N - C$ ones and C zeros on the diagonal (we can assume that first $N - C$ elements are one). Now let the matrix $\boldsymbol{U}_{N\times(N-C)}$ represent the first $N - C$ columns of $\boldsymbol{E}$. It is not hard to see that $\boldsymbol{U}$ satisfies $\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}_{N-C}$, $\boldsymbol{U}\boldsymbol{U}^T = \boldsymbol{V}$, $\boldsymbol{U}^T\boldsymbol{W} = 0$. Now we multiplying by $\boldsymbol{U}^T$ on both sides of the above equation:

$$\boldsymbol{U}^T\boldsymbol{y} = \boldsymbol{U}^T\sum_k \boldsymbol{X}_k\boldsymbol{\beta}_k + \boldsymbol{U}^T\boldsymbol{\epsilon} \tag{23}$$

$$cov(\boldsymbol{U}^T\boldsymbol{y}) = E[\boldsymbol{U}^T\boldsymbol{y}(\boldsymbol{U}^T\boldsymbol{y})^T] - E[\boldsymbol{U}^T\boldsymbol{y}]E[\boldsymbol{U}^T\boldsymbol{y}] \tag{24}$$

The matrix $\boldsymbol{U}^T$ is constant and the vector $\boldsymbol{y}$ is random. Therefore, we have $E[\boldsymbol{U}^T\boldsymbol{y}] = \boldsymbol{U}^T E[\boldsymbol{y}]$.

$$\boldsymbol{U}^T\boldsymbol{y}(\boldsymbol{U}^T\boldsymbol{y})^T = (\boldsymbol{U}^T\sum_k \boldsymbol{X}_k\boldsymbol{\beta}_k + \boldsymbol{U}^T\boldsymbol{\epsilon})(\boldsymbol{U}^T\sum_k \boldsymbol{X}_k\boldsymbol{\beta}_k + \boldsymbol{U}^T\boldsymbol{\epsilon})^T = \tag{25}$$

$$\sum_i\sum_j \boldsymbol{U}^T\boldsymbol{X}_i\boldsymbol{\beta}_i(\boldsymbol{U}^T\boldsymbol{X}_j\boldsymbol{\beta}_j)^T + (\boldsymbol{U}^T\boldsymbol{\epsilon})\sum_i(\boldsymbol{U}^T\boldsymbol{X}_i\boldsymbol{\beta}_i)^T + \sum_i \boldsymbol{U}^T\boldsymbol{X}_i\boldsymbol{\beta}_i(\boldsymbol{U}^T\boldsymbol{\epsilon})^T + \boldsymbol{U}^T\boldsymbol{\epsilon}(\boldsymbol{U}^T\boldsymbol{\epsilon})^T$$

Hence

$$E[\boldsymbol{U}^T\boldsymbol{y}(\boldsymbol{U}^T\boldsymbol{y})^T] = \sum_k \frac{\sigma^2_{g_k}}{M_k}(\boldsymbol{U}^T\boldsymbol{X}_k)(\boldsymbol{U}^T\boldsymbol{X}_k)^T + \sigma^2_\epsilon\boldsymbol{U}^T\boldsymbol{U} \tag{26}$$

Using $\boldsymbol{K}_k = \frac{\boldsymbol{X}_k\boldsymbol{X}_k^T}{M_k}$, we have:

$$cov(\boldsymbol{U}^T\boldsymbol{y}) = \boldsymbol{U}^T(\sum_k \sigma^2_{g_k}\boldsymbol{K}_k)\boldsymbol{U} + \sigma^2_\epsilon\boldsymbol{I}_{N-C} \tag{27}$$

The MoM estimator is obtained by solving the following ordinary least squares problem:

$$(\tilde{\sigma^2_1}, \ldots, \tilde{\sigma^2_K}, \tilde{\sigma^2_e}) = argmin_{(\sigma^2_1,\ldots,\sigma^2_K,\sigma^2_e)}||\boldsymbol{U}^T\boldsymbol{y}(\boldsymbol{U}^T\boldsymbol{y})^T - \boldsymbol{U}^T(\sum_k \sigma^2_k\boldsymbol{K}_k)\boldsymbol{U} - \sigma^2_\epsilon\boldsymbol{I}_{N-C}||^2_F \tag{28}$$