Title

Variance Estimation and Confidence Intervals from Genome-wide Association Studies Through High-dimensional Misspecified Mixed Model Analysis.

Authors

Dao, Cecilia
Zhao, Hongyu
Jiang, Jiming
et al.

Publication Date

2022-09-01

DOI

10.1016/j.jspi.2022.01.003

Peer reviewed

HHS Public Access

Author manuscript

J Stat Plan Inference. Author manuscript; available in PMC 2023 September 01.

Published in final edited form as:

J Stat Plan Inference. 2022 September ; 220: 15–23. doi:10.1016/j.jspi.2022.01.003.

Variance Estimation and Confidence Intervals from Genome-wide Association Studies Through High-dimensional Misspecified Mixed Model Analysis

Cecilia Dao[1], **Jiming Jiang**[2], **Debashis Paul**[2], **Hongyu Zhao**[1]

[1]*Yale University,* USA

[2]*University of California, Davis,* USA

Abstract

We study variance estimation and associated confidence intervals for parameters characterizing genetic effects from genome-wide association studies (GWAS) in misspecified mixed model analysis. Previous studies have shown that, in spite of the model misspecification, certain quantities of genetic interests are consistently estimable, and consistent estimators of these quantities can be obtained using the restricted maximum likelihood (REML) method under a misspecified linear mixed model. However, the asymptotic variance of such a REML estimator is complicated and not ready to be implemented for practical use. In this paper, we develop practical and computationally convenient methods for estimating such asymptotic variances and constructing the associated confidence intervals. Performance of the proposed methods is evaluated empirically based on Monte-Carlo simulations and real-data application.

Keywords

asymptotic approximation; confidence intervals; GWAS; heritability; mis-LMM; variance; unbiasedness

1 Introduction

Genome-wide association studies (GWAS) have proved successful by scanning the genome for genetic variations, e.g., single nucleotide polymorphisms (SNPs), that are associated with disease status and traits across study subjects. Tens of thousands of SNPs have been identified to be associated with various diseases and traits. For a review of the remarkable discoveries through GWAS, see Visscher et al. (2017). Researchers can use GWAS results to further medical research, such as to determine a person's risk of developing a disease or treat/prevent the disease. Genetic factors may account substantially for disease risk or various traits, and heritability estimates the proportion of variation in a phenotype due to genetic (and environmental) differences between individuals in a population. Historically,

heritability was inferred from resemblance among different degrees of related individuals (e.g., twin studies) without studying specific genetic variations, but today there is an emerging interest in quantifying how much variation can be accounted for from GWAS data due to the recent development of efficient genotyping and sequencing technology and the success of the GWAS strategy. However, when GWAS significant variants were considered, they only explained a small fraction of the genetic component of the phenotypes. The gap between the phenotypic variance explained by significant GWAS results and that estimated from classical heritability methods is known as the “missing heritability problem.”

More precisely, the problem refers to the concept that SNPs that are significant in GWASs cannot fully account for heritability of many diseases and traits. One explanation for missing heritability is that many SNPs jointly affect the phenotype, and SNPs with smaller effects that have not been identified may contribute to heritability as well. To address this issue, Yang et al. (2010) used an approach involving linear mixed models (LMMs) to show that a large proportion of heritability is not missing but rather captured by SNPs with weak effects that do not reach genome-wide significance level. The general idea is to use an LMM to treat the effects of all SNPs as random effects rather than relying on single-SNP association analysis. This approach has been widely used for heritability estimation in the genetics community via the genome-based restricted maximum likelihood (GREML) method (e.g., a popular implementation of GREML, with assumptions regarding the variance of the effect size prior distributions, is the GCTA software in Yang et al. 2011).

In an attempt to make the modelling more accurate, others have proposed extensions of this LMM approach. For instance, Heckerman et al. (2016) proposed to add an environmental random effect (along with a genomic random effect) in the LMM to reduce heritability inflation, and Zhou et al. (2013) proposed to use a hybrid of LMM and regression models to learn the true genetic architecture from the data to estimate heritability. To improve heritability estimation compared to GCTA, Speed et al. (2017) developed the LDAK model to factor in minor allele frequency (MAF), linkage disequilibrium (LD), and genotype certainty. Speed et al. (2020) extended their LDAK model to handle more complex heritabiltiy models by proposing an approximate model likelihood to be computed by GWAS summary statistics. Comprehensive comparisons of heritability estimation methods [Yang, Manolio, et al. (2011), Yang et al. (2015), Speed et al. (2017), Speed et al. (2012), Zaitlen et al. (2013), Bulik-Sullivan et al. (2015)] can be found in Evans et al. (2018). Zhu and Zhou (2020) also provides a review of statistical methods for heritability estimation.

Consider an LMM which can be expressed as

$$y = X\beta + \widetilde{Z}\alpha + \epsilon, \tag{1}$$

where y is an $n \times 1$ vector of observations; X is an $n \times q$ matrix of known covariates; β is a $q \times 1$ vector of unknown regression coefficients (the fixed effects); and $\widetilde{Z} = p^{-1/2}Z$, where Z is an $n \times p$ matrix whose entries are random variables. Furthermore, α is a $p \times 1$ vector of random effects that are distributed as $N(0, \sigma_a^2 I)$, ϵ is an $n \times 1$ vector of errors that is distributed as $N(0, \sigma_\epsilon^2 I)$, and α, ϵ and Z are independent. The heritability parameter is defined

as $h^2 = \sigma_\alpha^2 / (\sigma_\alpha^2 + \sigma_\epsilon^2)$ when the average trace of the genetic relationship matrix is one (i.e., under the GCTA model); for the general equation, see equation B1 in Speed et al. (2012).

The LMM (1) is the model used by Yang et al. (2010) where it is assumed that the effects of all the SNPs (random effects) are nonzero. The restricted maximum likelihood (REML) estimator of the heritability is given by $\hat{h}^2 = \hat{\sigma}_\alpha^2 / (\hat{\sigma}_\alpha^2 + \hat{\sigma}_\epsilon^2)$, where the estimates of the variance components $\hat{\sigma}_\alpha^2$ and $\hat{\sigma}_\epsilon^2$ are based on the REML method [e.g., Jiang (2007), Section 1.3.2]. In reality, however, only a subset of the SNPs are potentially nonzero. Let α be the vector of effect sizes across the whole genome, where non-causal SNPs have a zero effect. Without loss of generality, we can assume that $\alpha = \{\alpha_{(1)}', 0'\}$, where $\alpha_{(1)}'$ is the vector of the first m components of α ($1 \le m \le p$), and 0 is the $(p - m) \times 1$ vector of zeros. Correspondingly, we have $\tilde{Z} = [\tilde{Z}_{(1)}; \tilde{Z}_{(2)}]$, where $Z_{(j)} = p^{-1/2} Z_{(j)}$, $j = 1, 2$, $Z_{(1)}$ is $n \times m$, and $Z_{(2)}$ is $n \times (p - m)$. Therefore, the true LMM can be expressed as

$$y = X\beta + \tilde{Z}_{(1)}\alpha_{(1)} + \epsilon, \tag{2}$$

With respect to the true model (2), the assumed model (1) is misspecified. We call the latter a misspecified LMM, or mis-LMM.

Jiang et al. (2016) showed that even under a mis-LMM, $\hat{\sigma}_\epsilon^2$ and $\hat{h}^2$ are consistent by investigating the asymptotic behavior of the estimators as the sample size and the number of SNPs increase to infinity, such that their ratio converges to a finite, nonzero constant. However, the asymptotic variances of the REML estimators have complex forms that are not ready to be implemented for practical use. This issue is important, from a practical point of view, because the asymptotic variance is used to obtain the standard error of the estimator, and confidence interval for the associated parameter, in applications. The main goal of the current paper is to propose accurate estimators of the variance of $\hat{\sigma}_\epsilon^2$ and $\hat{h}^2$ along with confidence intervals that are robust even under the mis-LMM. The proposed variance estimators are derived based on asymptotic approximation; they have analytic expressions and are simple to use. Using the variance estimators and Jiang et al. (2016), we construct approximate $100(1 - \alpha)\%$ confidence intervals for the associated parameters.

In this paper, we first derive the variance estimators and associated confidence intervals, providing technical details in the the Appendix. Then, we compare the performance of our method with that of GREML under the GCTA model through simulation studies and a real data example using the UK Biobank data. We end with a discussion of the results.

2 Derivation of variance estimators

As noted, Jiang et al. (2016) showed that REML estimators of certain variance components of genetic interest are consistently estimable and asymptotically normal; however, the corresponding asymptotic variances do not have expressions suitable for implementation. Thus, our first objective is to derive (simple) estimators of those asymptotic variances. Let us begin with estimation of $\text{var}(\hat{\sigma}_\epsilon)$. By Jiang et al. (2016), we have the expression

$$\hat{\sigma}_\epsilon^2 = \frac{y' P_{\hat{\gamma}}^2 y}{\mathrm{tr}(P_{\hat{\gamma}})}, \tag{3}$$

where $P_\gamma = V_\gamma^{-1} - V_\gamma^{-1} X (X' V_\gamma^{-1} X)^{-1} X' V^{-1}$ with $V_\gamma = I_n + \gamma \tilde{Z}\tilde{Z}'$, $\gamma = \sigma_\alpha^2 / \sigma_\epsilon^2$ and $\hat{\gamma} = \hat{\sigma}_\alpha^2 / \hat{\sigma}_\epsilon^2$. Some technical (see Subsection A.1 of the Appendix) derivations lead to the following approximation:

$$\hat{\sigma}_\epsilon^2 \approx \frac{\mathrm{E}(U_{\gamma,y})}{\mathrm{E}(U_{\gamma,y}) - \mathrm{E}(S_{\gamma,y})} \cdot \frac{y' P_\gamma^2 y}{\mathrm{tr}(P_\gamma)} + \frac{\mathrm{E}(S_{\gamma,y})}{\mathrm{E}(S_{\gamma,y}) - \mathrm{E}(U_{\gamma,y})} \cdot \frac{y' Q_\gamma y}{\mathrm{tr}(P_\gamma \tilde{Z}\tilde{Z}')}, \tag{4}$$

where $\gamma = \gamma_*$, which is the asymptotic limit of $\hat{\gamma}$ according to Jiang et al. (2016). Denote the right side of (4) by $\tilde{\sigma}_\epsilon^2$, then, by the law of total variance, we have

$$\mathrm{var}(\hat{\sigma}_\epsilon^2) \approx \mathrm{var}(\tilde{\sigma}_\epsilon^2) = \mathrm{E}\{\mathrm{var}(\tilde{\sigma}_\epsilon^2 \mid Z)\} + \mathrm{var}\{\mathrm{E}(\tilde{\sigma}_\epsilon^2 \mid Z)\}. \tag{5}$$

It can be shown that the second term on the right side of (5) is of lower order than the first term; therefore, we have

$$\mathrm{var}(\hat{\sigma}_\epsilon^2) \approx \mathrm{E}\{\mathrm{var}(\tilde{\sigma}_\epsilon^2 \mid Z)\}. \tag{6}$$

To obtain a further approximation, define

$$A = \frac{\mathrm{tr}(Q_\gamma \tilde{Z}\tilde{Z}')\{\mathrm{tr}(P_\gamma^2)\mathrm{tr}(Q_\gamma \tilde{Z}\tilde{Z}') - \mathrm{tr}^2(Q_\gamma)\}}{\mathrm{tr}^2(P_\gamma)\mathrm{tr}^2(P_\gamma \tilde{Z}\tilde{Z}')},$$
$$B = \frac{\mathrm{tr}(Q_\gamma \tilde{Z}\tilde{Z}')}{\mathrm{tr}(P_\gamma \tilde{Z}\tilde{Z}')} - \frac{\mathrm{tr}(Q_\gamma)}{\mathrm{tr}(P_\gamma)}.$$

Then, it can be shown (see Subsection A.2 of the Appendix) that the right side of (6) can be approximated by $2\sigma_\epsilon^2 \mathrm{E}(A) / \{\mathrm{E}(B)\}^2$. Thus, in conclusion, we obtain the following estimator of $\mathrm{var}(\hat{\sigma}_\epsilon)$:

$$\widehat{\mathrm{var}(\hat{\sigma}_\epsilon^2)} = 2\hat{\sigma}_\epsilon^2 \frac{\hat{A}}{\hat{B}^2}, \tag{7}$$

where $\hat{A}$, $\hat{B}$ are A, B with γ replaced by $\hat{\gamma}$, respectively.

Next, we consider estimation of $\mathrm{var}(\hat{\gamma})$. Using similar arguments (details omitted), it can be shown that

$$\mathrm{var}(\hat{\gamma}) \approx \frac{\mathrm{E}(C)}{\{\mathrm{E}(B)\}^2}, \tag{8}$$

where

$$C = \mathrm{tr}\left[\left\{\frac{P_\gamma}{\mathrm{tr}(P_\gamma)} - \frac{P_\gamma \widetilde{Z}\widetilde{Z}'}{\mathrm{tr}(P_\gamma \widetilde{Z}\widetilde{Z}')}\right\}^2\right].$$

Thus, an estimator of $\mathrm{var}(\hat{\gamma})$ is given by

$$\widehat{\mathrm{var}(\hat{\gamma})} = \frac{\hat{C}}{\hat{B}^2}. \tag{9}$$

The variance estimator (9) is used to obtain a variance estimator for $\hat{h}^2$, the REML estimator of the heritability, h^2. Using the expression $\hat{h}^2 = \hat{\gamma}/(1+\hat{\gamma})$, and the delta-method (e.g., Jiang 2010, sec. 4.2), we obtain

$$\mathrm{var}(\hat{h}^2) \approx \frac{\mathrm{var}(\hat{\gamma})}{(1+\gamma)^4}, \tag{10}$$

where, again, $\gamma = \gamma_*$, the limit of $\hat{\gamma}$. Thus, an estimator of $\mathrm{var}(\hat{h}^2)$ is given by

$$\widehat{\mathrm{var}(\hat{h}^2)} = \frac{\hat{C}}{(1+\hat{\gamma})^4 \hat{B}^2}. \tag{11}$$

Note that all of the variance estimators obtained here are guaranteed to be nonnegative (and positive with probability one), a desirable property for a variance estimator. In particular, one can take square root of the variance estimator, and use it to construct a large-sample confidence interval for the corresponding parameter. Let θ denote a parameter of interest, such as σ_ϵ^2, h^2, and $\hat{\theta}$ be its estimator. Let $\widehat{\mathrm{var}(\hat{\theta})}$ be a variance estimator for $\hat{\theta}$ that is guaranteed nonnegative. Since $\hat{\theta}$ has an asymptotically normal distribution due to Theorem 3.2 in Jiang et al. (2016), given $\alpha \in (0, 1)$, an approximate $100(1-\alpha)\%$ confidence interval for θ is

$$\left[\hat{\theta} - z_{\alpha/2}\sqrt{\widehat{\mathrm{var}(\hat{\theta})}},\quad \hat{\theta} + z_{\alpha/2}\sqrt{\widehat{\mathrm{var}(\hat{\theta})}}\right], \tag{12}$$

where $z_{\alpha/2}$ is the $\alpha/2$ critical value of $N(0, 1)$ [i.e., $\mathrm{P}(Z > z_{\alpha/2}) = \alpha/2$ for $Z \sim N(0, 1)$].

3 Simulation Studies

We simulate scenarios similar to that in Jiang et al. (2016). Specifically, we simulate the allele frequencies for p SNPs from the Uniform[0.05, 0.5] distribution, and denote f_j as the allele frequency of the jth SNP, for $j = 1, 2, \ldots p$. The genotype matrix $U \in \{0, 1, 2\}^{n\times p}$ has rows corresponding to the individuals and the columns corresponding to the SNPs. The genotype value of each individual for the jth SNP is sampled from $\{0, 1, 2\}$ with probabilities $(1-f_j)^2$, $2f_j(1-f_j)$, f_j^2, respectively. Let the standardized genotype matrix Z be such that each column of U is standardized to have zero mean and unit variance, and

then let $\widetilde{Z} = p^{-\frac{1}{2}}Z$. We express the relationship between the phenotypic vector y and the standardized genotype matrix $\widetilde{Z}$ in the LMM in (1).

As previously noted, (1) assumes that $\alpha_j \sim N(0, \sigma_\alpha^2)$ for all $j \in \{1, 2, \ldots, p\}$ when in reality, only a subset m of the SNPs is associated with the phenotype. Thus, a correct model is (2) and the heritability should be

$$h^2_{\text{true}} = \frac{\omega\sigma_\alpha^2}{\omega\sigma_\alpha^2 + \sigma_\epsilon^2}, \tag{13}$$

where $\omega = m/p$. Since it is not possible to identify all of the m SNPs in practice, we follow model (2) to simulate the phenotypes and use all of the SNPs in Z to estimate the variance components, σ_α^2 and σ_ϵ^2, in model (1). We therefore estimate the heritability as

$$\hat{h}^2 = \frac{\hat{\sigma}_\alpha^2}{\hat{\sigma}_\alpha^2 + \hat{\sigma}_\epsilon^2}, \tag{14}$$

where the estimates of the variance components are their REML estimates.

In the simulations, given n, p, and m, we set the true parameters as $\beta = 0$, $\sigma_\epsilon^2 = \sigma_{\epsilon 0}^2$, and $\sigma_\alpha^2 = \sigma_{\alpha 0}^2$ for $(\sigma_{\epsilon 0}^2, \sigma_{\alpha 0}^2) \in \{(0.8, \frac{0.2}{\omega}), (0.6, \frac{0.4}{\omega}), (0.4, \frac{0.6}{\omega}), (0.2, \frac{0.8}{\omega})\}$,, so that the heritability parameter is varied. We perform simulations with those true parameters under misspecifications of $\omega \in \{0.005, 0.01, 0.05, 0.1, 0.5\}$. Note that $\gamma = \frac{\sigma_\alpha^2}{\sigma_\epsilon^2}$. We simulate the data under model (2), but compute REML estimates under model (1). For each scenario, we carry out 300 replications, and report the results (see below).

Let $\text{var}(\hat{\theta})$ be the sample variance of all of the simulated $\hat{\theta}$s for $\theta \in \{\sigma_\epsilon^2, h^2\}$, $\hat{v} = \widehat{\text{var}(\hat{\theta})}$, and $\text{E}(\hat{v})$ be the sample mean of all of the simulated $\hat{v}$s. The percentage of relative bias (%RB) is defined as

$$\%\,\text{RB} = 100 \times \left\{ \frac{\text{E}(\hat{v}) - \text{var}(\hat{\theta})}{\text{var}(\hat{\theta})} \right\}.$$

We also look at the sample standard deviation of all of the simulated $\hat{v}$s, denoted as $\text{s}(\hat{v})$. The N_λ for $\lambda \in \{0.01, 0.05, 0.1\}$ is the empirical coverage probability for large sample confidence intervals of θ at level λ. Since θ is bounded, we also consider the large sample truncated confidence intervals of θ and denote the empirical coverage probability by T_λ for $\lambda \in \{0.01, 0.05, 0.1\}$. To find the truncated confidence intervals of $\theta \in \{\sigma_\epsilon^2, h^2\}$, we use the quantiles of the truncated normal distribution, where the mean is the REML estimate of θ and the variance is the variance estimate of REML estimate of θ. In particular, since the lower bound of $\theta = \sigma_\epsilon^2$ is 0, we truncate the lower bound by 0 but not the upper bound. For $\theta = h^2$, we truncate the lower bound by 0 and the upper bound by 1. The quantiles of the truncated normal distribution at levels $\lambda/2$ and $1 - \lambda/2$ are the lower and upper bounds of the confidence interval, respectively.

In the following tables, we showcase some results for $\theta \in \{\sigma_\epsilon^2, h^2\}$ using our method in comparison to the GREML method under GCTA. Other simulations are given in the supplementary materials.

4 Real data example

We apply the proposed method to a real data example using a subset of indviduals with height data from the UK Biobank (UKBB) database. We consider $n = 4{,}986$ Caucasian individuals who are unrelated up to the 3rd degree using KING (Manichaikul et al. 2010) to avoid inflating the heritability estimate. The UKBB performed genotype imputation using IMPUTE4 and the Haplotype Reference Consortium reference panel (Bycroft et al. 2018). We have retained imputed SNPs with high imputation quality (INFO scores greater than 0.8). Then, we have removed imputed SNPs with a missing call rate exceeding 0.05, a Hardy-Weinberg equilibrium exact test p-value below 1×10^{-10}, or a minor allele frequency below 0.05. After quality control, $p = 6{,}133{,}110$ SNPs remained for analysis.

After the preprocessing described above, we apply the LMM approach described in model (1) to obtain REML estimates of the variance components, and then estimate their variances and construct confidence intervals for the parameters of interest. For the matrix X of fixed effects, in addition to the intercept, we account for sex, age, and population stratification using the first twenty principal component scores derived from genotype data provided by the UKBB.

We obtain REML estimates $\hat{\sigma}_\epsilon^2 = 20.150$, $\hat{\gamma} = 1.003$, and $\hat{h}^2 = 0.5009$. Using our approach, we get the following variance estimates for our parameter of interests: $\widehat{\text{var}}(\hat{\sigma}_\epsilon^2) = 7.117$, and $\widehat{\text{var}}(\hat{h}^2) = 0.0045$. The variance estimates from GCTA are comparable: $\widehat{\text{var}}(\hat{\sigma}_\epsilon^2) = 7.242$, $\widehat{\text{var}}(\hat{h}^2) = 0.0046$. The corresponding 95% confidence intervals for σ_ϵ^2 are (14.921, 25.379) and (14.875, 25.424) for our method and GCTA, respectively. The 95% confidence intervals for h^2 are (0.3697, 0.6321) and (0.3685, 0.6332) for our method and GCTA, respectively. The heritability of height estimated by LMM/REML have similar results in other data sets (e.g., Yang et al.(2010), Zhou et al. (2013), Golan et al. (2014)).

5 Discussion

In this paper, we developed variance estimators and their associated confidence intervals that are robust under misspecified mixed model analysis in GWAS studies, supported by the theory established in Jiang et al. (2016). In our simulation studies, the GREML under the GCTA model performed satisfactory in terms of variance estimators and associated confidence intervals despite not taking into account misspecification. In fact, our proposed method and GCTA method, performed similarly in our simulation studies and real data example.

We also considered a nonparametric approach to construct bootstrap confidence intervals. Particularly, let F be the true distribution of $\theta(F) \in \{\sigma_\epsilon^2, h^2\}$, and let $\hat{\theta} \equiv \theta(\hat{F}) \in \{\hat{\sigma}_\epsilon^2, \hat{h}^2\}$ be

the REML estimate of $\theta(F)$. Let $\hat{F}^*$ denote a bootstrap approximation to $\hat{F}$. Since the sampling distribution of $\theta(\hat{F}) / \theta(F) \approx \theta(\hat{F}^*) / \theta(\hat{F})$, we constructed an approximate $100(1 - \alpha)\%$ confidence interval for θ as $\left(\frac{\hat{\theta}}{q^*_{1-\alpha/2}}, \frac{\hat{\theta}}{q^*_{\alpha/2}}\right)$, where q^*_t is the t-th quantile of the bootstrap sampling distribution of $\theta(\hat{F}^*) / \theta(\hat{F})$. However, simulation studies results based on the method of confidence interval construction in this paper performed better based on empirical coverage probabilities so it was presented here.

6 Conclusion

The GREML method under the GCTA model is based on the standard LMM analysis (e.g., Jiang 2007), which does not take into account (i) that the LMM is misspecified (see Section 1); (ii) that the design matrix, Z, for the random effects is random; and (iii) the asymptotic framework is different than the standard assumption that the number of random effects is, at most, of the same order as the sample size. In typical GWAS, the number of random effects, which correspond to the SNPs, is typically of higher order than the sample size. On the other hand, our method is fully supported by the recently established theory on high-dimensional mis-LMM analysis (Jiang et al. 2016), based on which, the variance estimators are derived in the current paper. Thus, in a way, the results of this paper provide additional justification for the use of the GREML method under the GCTA model for inference.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Research of all authors was supported in part by NSF grant DMS-1713120. Research of Debashis Paul was also supported in part by NSF grants DMS-1811405 and DMS-1915894. Research of Hongyu Zhao was also supported in part by NSF grant DMS-1902903 and NIH R01 GM134005. The UK Biobank resource was used under an approved data request (ref: 29900). The authors wish to thank Professor Can Yang for helpful discussions. The authors are grateful to the referees for their suggestions and comments which helped improve the paper in important ways.

Appendix

A.1 Derivation of (4)

Using the identity $B^{-1} = A^{-1} + A^{-1}(A - B)B^{-1}$, the following first-order approximations can be derived: $P_{\hat{\gamma}} \approx P_\gamma - (\hat{\gamma} - \gamma)Q_\gamma$, where $\gamma = \gamma_*$ which is the limit of $\hat{\gamma}$, and $Q_\gamma = P_\gamma \widetilde{Z}\widetilde{Z}' P_\gamma$. With those, and (3), the following approximation can be derived:

$$\hat{\sigma}^2_\epsilon \approx \frac{y' P^2_\gamma y}{\mathrm{tr}(P_\gamma)} - (\hat{\gamma} - \gamma)S_{\gamma, y}, \tag{A.1}$$

where

$$S_{\gamma,y} = \frac{y' R_\gamma y}{\mathrm{tr}(P_\gamma)} - \frac{\mathrm{tr}(Q_\gamma)}{\mathrm{tr}^2(P_\gamma)} y' P_\gamma^2 y \tag{A.2}$$

with $R_\gamma = P_\gamma Q_\gamma + Q_\gamma P_\gamma$.

Next, we obtain an expansion for $\hat{\gamma} - \gamma$. From (3) of Jiang et al. (2016), we have

$$\frac{y' P_\gamma \widetilde{Z}\widetilde{Z}' P_\gamma y}{\mathrm{tr}(P_\gamma \widetilde{Z}\widetilde{Z}')} = \frac{y' P_\gamma^2 y}{\mathrm{tr}(P_\gamma)}. \tag{A.3}$$

The RHS (righthand side) of (A.3) is approximated by (A.1). As for the LHS (lefthand size) of (A.3), one can derive $Q_{\hat{\gamma}} \approx Q_\gamma - (\hat{\gamma} - \gamma) T_\gamma$, where $T_\gamma = P_\gamma \widetilde{Z}\widetilde{Z}' Q_\gamma + Q_\gamma \widetilde{Z}\widetilde{Z}' P_\gamma$. Furthermore, using the elementary expansion of Jiang 2010 (p. 103), we have

$$\frac{1}{\mathrm{tr}(P_{\hat{\gamma}} \widetilde{Z}\widetilde{Z}')} \approx \frac{1}{\mathrm{tr}(P_\gamma \widetilde{Z}\widetilde{Z}')} + \frac{(\hat{\gamma} - \gamma)\mathrm{tr}(Q_\gamma \widetilde{Z}\widetilde{Z}')}{\mathrm{tr}^2(P_\gamma \widetilde{Z}\widetilde{Z}')}.$$

Thus, the LHS of (A.3) can be approximated by

$$\begin{aligned} &\{y' Q_\gamma y - (\hat{\gamma} - \gamma) y' T_\gamma y\} \left\{ \frac{1}{\mathrm{tr}(P_\gamma \widetilde{Z}\widetilde{Z}')} + \frac{(\hat{\gamma} - \gamma)\mathrm{tr}(Q_\gamma \widetilde{Z}\widetilde{Z}')}{\mathrm{tr}^2(P_\gamma \widetilde{Z}\widetilde{Z}')} \right\} \\ &\approx \frac{y' Q_\gamma y}{\mathrm{tr}(P_\gamma \widetilde{Z}\widetilde{Z}')} - (\hat{\gamma} - \gamma) U_{\gamma,y}, \end{aligned} \tag{A.4}$$

where

$$U_{\gamma,y} = \frac{y' T_\gamma y}{\mathrm{tr}(P_\gamma \widetilde{Z}\widetilde{Z}')} - \frac{\mathrm{tr}(Q_\gamma \widetilde{Z}\widetilde{Z}')}{\mathrm{tr}^2(P_\gamma \widetilde{Z}\widetilde{Z}')} y' Q_\gamma y.$$

By equating the LHS to the RHS, i.e., (A.1) to (A.4), we obtain the following:

$$\begin{aligned} \hat{\gamma} - \gamma &\approx \frac{1}{S_{\gamma,y} - U_{\gamma,y}} \left\{ \frac{y' P_\gamma^2 y}{\mathrm{tr}(P_\gamma)} - \frac{y' Q_\gamma y}{\mathrm{tr}(P_\gamma \widetilde{Z}\widetilde{Z}')} \right\} \\ &\approx \frac{1}{\mathrm{E}(S_{\gamma,y}) - \mathrm{E}(U_{\gamma,y})} \left\{ \frac{y' P_\gamma^2 y}{\mathrm{tr}(P_\gamma)} - \frac{y' Q_\gamma y}{\mathrm{tr}(P_\gamma \widetilde{Z}\widetilde{Z}')} \right\}. \end{aligned} \tag{A.5}$$

Combining (A.1) and (A.5), we obtain

$$\begin{aligned} \hat{\sigma}_\epsilon^2 &\approx \frac{y' P_\gamma^2 y}{\mathrm{tr}(P_\gamma)} + \frac{S_{\gamma,y}}{\mathrm{E}(U_{\gamma,y}) - \mathrm{E}(S_{\gamma,y})} \left\{ \frac{y' P_\gamma^2 y}{\mathrm{tr}(P_\gamma)} - \frac{y' Q_\gamma y}{\mathrm{tr}(P_\gamma \widetilde{Z}\widetilde{Z}')} \right\} \\ &\approx \frac{y' P_\gamma^2 y}{\mathrm{tr}(P_\gamma)} + \frac{\mathrm{E}(S_{\gamma,y})}{\mathrm{E}(U_{\gamma,y}) - \mathrm{E}(S_{\gamma,y})} \left\{ \frac{y' P_\gamma^2 y}{\mathrm{tr}(P_\gamma)} - \frac{y' Q_\gamma y}{\mathrm{tr}(P_\gamma \widetilde{Z}\widetilde{Z}')} \right\} \\ &= \frac{\mathrm{E}(U_{\gamma,y})}{\mathrm{E}(U_{\gamma,y}) - \mathrm{E}(S_{\gamma,y})} \cdot \frac{y' P_\gamma^2 y}{\mathrm{tr}(P_\gamma)} \\ &\quad + \frac{\mathrm{E}(S_{\gamma,y})}{\mathrm{E}(S_{\gamma,y}) - \mathrm{E}(U_{\gamma,y})} \cdot \frac{y' Q_\gamma y}{\mathrm{tr}(P_\gamma \widetilde{Z}\widetilde{Z}')} \\ &= \widetilde{\sigma}_\epsilon^2. \end{aligned}$$

A.2 Further approximation to the right side of (6)

Note that $\tilde{\sigma}_\epsilon^2 = y' D_\gamma y$ for some matri D_γ. Thus, by the normal theory (e.g., Jiang 2007, p. 238), we have

$$\mathrm{var}(\tilde{\sigma}_\epsilon^2 \mid Z) = 2\mathrm{tr}(D_\gamma \Sigma D_\gamma \Sigma) = 2\mathrm{tr}\left\{\left(\sigma_\epsilon^2 D_\gamma + \sigma_\alpha^2 \sum_{i=1}^{m} D_\gamma \tilde{Z}_i \tilde{Z}_i'\right)^2\right\},$$

where $\Sigma = \sigma_\epsilon^2 I_n + \sigma_\alpha^2 \sum_{i=1}^m \tilde{Z}_i \tilde{Z}_i'$ is the true covariance matrix of y. Therefore, we have $\mathrm{E}\{\mathrm{var}(\tilde{\sigma}_\epsilon^2 \mid Z)\} = 2\{\sigma_\epsilon^4 \mathrm{E}\{\mathrm{tr}(D_\gamma^2)\} + 2\sigma_\epsilon^2\sigma_\alpha^2 \mathrm{E}(I_1) + \sigma_\alpha^4 \mathrm{E}(I_2)\}$ with $I_1 = \sum_{i=1}^m \mathrm{tr}(D_\gamma \tilde{Z}_i \tilde{Z}_i' D_\gamma)$ and $I_2 = \sum_{i,j=1}^m \mathrm{tr}(D_\gamma \tilde{Z}_i \tilde{Z}_i' D_\gamma \tilde{Z}_j \tilde{Z}_j')$. By the fact that Z_i, $1 \le i \le p$ are i.i.d., it can be shown that $\mathrm{E}(I_1) = \omega \mathrm{E}\{\mathrm{tr}(D_\gamma \tilde{Z}\tilde{Z}' D_\gamma)\}$, and $\mathrm{E}(I_2) \approx \omega^2 \mathrm{E}\{\mathrm{tr}(D_\gamma \tilde{Z}\tilde{Z}' D_\gamma \tilde{Z}\tilde{Z}')\}$, where $\omega = m/p$. Thus,

$$\begin{aligned}
\mathrm{E}\{\mathrm{var}(\tilde{\sigma}_\epsilon^2 \mid Z)\} &= 2[\sigma_\epsilon^4 \mathrm{E}\{\mathrm{tr}(D_\gamma^2)\} + 2\omega\sigma_\epsilon^2\sigma_\alpha^2 \mathrm{E}\{\mathrm{tr}(D_\gamma \tilde{Z}\tilde{Z}' D_\gamma)\} \\
&\quad + \omega^2 \sigma_\alpha^4 \mathrm{E}\{\mathrm{tr}(D_\gamma \tilde{Z}\tilde{Z}' D_\gamma \tilde{Z}\tilde{Z}')\}] \\
&= 2\mathrm{E}\{\mathrm{tr}(\sigma_\epsilon^4 D_\gamma^2 + 2\sigma_\epsilon^2 \omega \sigma_\alpha^2 D_\gamma \tilde{Z}\tilde{Z}' D_\gamma + \omega^2 \sigma_\alpha^4 D_\gamma \tilde{Z}\tilde{Z}' D_\gamma \tilde{Z}\tilde{Z}')\} \\
&= 2\mathrm{E}[\mathrm{tr}\{(\sigma_\epsilon^2 D_\gamma + \omega\sigma_\alpha^2 D_\gamma \tilde{Z}\tilde{Z}')^2\}].
\end{aligned}$$

It follows that the RHS of (6) is approximately equal to

$$2\mathrm{E}\left[\mathrm{tr}\left\{(\sigma_\epsilon^2 D_\gamma + \omega\sigma_\alpha^2 D_\gamma \tilde{Z}\tilde{Z}')^2\right\}\right] = 2\sigma_\epsilon^4 \mathrm{E}\left[\mathrm{tr}\{(aP_\gamma + bP_\gamma \tilde{Z}\tilde{Z}')^2\}\right], \tag{A.6}$$

where

$$a = \frac{\mathrm{E}(U_{\gamma,y})}{\mathrm{tr}(P_\gamma)\{\mathrm{E}(U_{\gamma,y}) - \mathrm{E}(S_{\gamma,y})\}}, \quad b = \frac{\mathrm{E}(S_{\gamma,y})}{\mathrm{tr}(P_\gamma \tilde{Z}\tilde{Z}')\{\mathrm{E}(S_{\gamma,y}) - \mathrm{E}(U_{\gamma,y})\}}$$

since $\omega\sigma_\alpha^2$ is estimated by $\hat{\sigma}_\alpha^2$ and $P_\gamma V_\gamma P_\gamma = P_\gamma$. So when σ_ϵ^2 is pulled out, $\omega\sigma_\alpha^2 / \sigma_\epsilon^2$ is replaced by γ, which is estimated by $\hat{\gamma}$.

Thus, we have

$$aP_\gamma + bP_\gamma \tilde{Z}\tilde{Z}' = \frac{1}{\mathrm{E}(U_{\gamma,y}) - \mathrm{E}(S_{\gamma,y})}\left\{\frac{\mathrm{E}(U_{\gamma,y})}{\mathrm{tr}(P_\gamma)} P_\gamma - \frac{\mathrm{E}(S_{\gamma,y})}{\mathrm{tr}(P_\gamma \tilde{Z}\tilde{Z}')} P_\gamma \tilde{Z}\tilde{Z}'\right\}. \tag{A.7}$$

Furthermore, using, once again, the i.i.d. property, it can be shown that

$$\mathrm{E}(S_{\gamma,y}) = \sigma_\epsilon^2 \mathrm{E}\left\{\frac{\mathrm{tr}(Q_\gamma)}{\mathrm{tr}(P_\gamma)}\right\}, \quad \mathrm{E}(U_{\gamma,y}) = \sigma_\epsilon^2 \mathrm{E}\left\{\frac{\mathrm{tr}(Q_\gamma \tilde{Z}\tilde{Z}')}{\mathrm{tr}(P_\gamma \tilde{Z}\tilde{Z}')}\right\}. \tag{A.8}$$

Combining (6), (A.6)-(A.8), it can be shown $\mathrm{var}(\hat{\sigma}_\epsilon^2) \approx 2\sigma_\epsilon^2 \mathrm{E}(A) / \{\mathrm{E}(B)\}^2$, where A, B are given below (6).

References

[1]. Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, Yang J. (2017), 10 Years of GWAS Discovery: Biology, Function, and Translation, Am J Hum Genet. Jul 6;101(1):5–22. [PubMed: 28686856]

[2]. Yang J, Benyamin B, McEvoy B et al. (2010), Common SNPs explain a large proportion of the heritability for human height. Nat Genet. 42, 565–569. 10.1038/ng.608 [PubMed: 20562875]

[3]. Yang J, Lee SH, Goddard ME, Visscher PM. (2011), GCTA: a tool for genome-wide complex trait analysis, Am J Hum Genet. 2011;88(1):76–82. [PubMed: 21167468]

[4]. Heckerman D, Gurdasani D, Kadie C, Pomilla C, Carstensen T, Martin H, Ekoru K, Nsubuga RN, Ssenyomo G, Kamali A, Kaleebu P, Widmer C, Sandhu MS. (2016), Linear mixed model for heritability estimation that explicitly addresses environmental variation. Proc Natl Acad Sci U S A. 113(27):7377–82. [PubMed: 27382152]

[5]. Zhou X, Carbonetto P, Stephens M. (2013), Polygenic modeling with bayesian sparse linear mixed models, PLoS Genet. 9(2):e1003264. [PubMed: 23408905]

[6]. Speed D, Cai N, Johnson MR, Nejentsev S, & Balding DJ (2017). Reevaluation of SNP heritability in complex human traits. Nature genetics, 49(7), 986–992. [PubMed: 28530675]

[7]. Speed D, Holmes J, & Balding DJ (2020). Evaluating and improving heritability models using summary statistics. Nature Genetics, 52(4), 458–462. [PubMed: 32203469]

[8]. Yang J, Manolio TA, Pasquale LR, Boerwinkle E, Caporaso N, Cunningham JM, … & Hill WG (2011). Genome partitioning of genetic variation for complex traits using common SNPs. Nature genetics, 43(6), 519. [PubMed: 21552263]

[9]. Yang J, Bakshi A, Zhu Z, Hemani G, Vinkhuyzen AA, Lee SH, … & Snieder H (2015). Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. Nature genetics, 47(10), 1114. [PubMed: 26323059]

[10]. Speed D, Hemani G, Johnson MR & Balding DJ Improved heritability estimation from genome-wide SNPs. Am. J. Hum. Genet 91, 1011–1021 (2012). [PubMed: 23217325]

[11]. Zaitlen N et al. Using extended genealogy to estimate components of heritability for 23 quantitative and dichotomous traits. PLoS Genet. 9, e1003520 (2013). [PubMed: 23737753]

[12]. Bulik-Sullivan BK et al. LD score regression distinguishes confounding from poly-genicity in genome-wide association studies. Nat. Genet 47, 291–295 (2015). [PubMed: 25642630]

[13]. Evans LM, Tahmasbi R, Vrieze SI, Abecasis GR, Das S, Gazal S, … & Yang J (2018). Comparison of methods that use whole genome data to estimate the heritability and genetic architecture of complex traits. Nature genetics, 50(5), 737–745. [PubMed: 29700474]

[14]. Zhu H, & Zhou X (2020). Statistical methods for SNP heritability estimation and partition: A review. Computational and Structural Biotechnology Journal.

[15]. Jiang J (2007), Linear and Generalized Linear Mixed Models and Their Applications, Springer, New York.

[16]. Jiang J, Li C, Paul D, Yang C, and Zhao H (2016), On high-dimensional misspecified mixed model analysis in genome-wide association study, Ann. Statist 44, 2127–2160.

[17]. Lohr SL, Divan M (1997), Comparison of confidence intervals for variance components with unbalanced data, Journal of statistical computation and simulation, 58(1), 83–97.

[18]. Burch BD (2007), Comparing pivotal and REML-based confidence intervals for heritability, Journal of agricultural, biological, and environmental statistics, 12(4), 470.

[19]. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM. (2010) Robust relationship inference in genome-wide association studies, Bioinformatics, 26(22):2867–73. [PubMed: 20926424]

[20]. Bycroft C, Freeman C, Petkova D et al. (2018), The UK Biobank resource with deep phenotyping and genomic data, Nature, 562(7726):203–9. [PubMed: 30305743]

[21]. Golan D, Lander ES, Rosset S (2014) Measuring missing heritability: inferring the contribution of common variants. Proc Natl Acad Sci., 111(49):E5272–E5281. [PubMed: 25422463]

[22]. Jiang J (2010), Large Sample Techniques for Statistics, Springer, New York.

Highlights for Review

- Justification of GREML GCTA under misspecification
- Variance estimations and confidence intervals for GWAS mixed model
- Misspecified mixed model GWAS

Table 1:

$\sigma^2_{\epsilon_0} = 0.4$, $\sigma^2_{\alpha_0} = 0.6\frac{p}{m}$ for $\theta = \sigma^2_\epsilon$, $(n, p) = (2000, 20000)$

m	% RB	var($\hat{\theta}$)	E($\hat{v}$)	s($\hat{v}$)	$N_{0.01}$	$N_{0.05}$	$N_{0.1}$	$T_{0.01}$	$T_{0.05}$	$T_{0.1}$
20	3.208	0.009	0.010	0.003	0.983	0.947	0.900	0.983	0.947	0.900
200	−10.250	0.010	0.009	0.001	0.983	0.930	0.860	0.983	0.930	0.860
2, 000	7.608	0.008	0.009	0.001	0.990	0.953	0.907	0.990	0.957	0.907
20, 000	−3.961	0.010	0.009	0.001	0.990	0.947	0.890	0.990	0.947	0.890

Table 2:

GCTA CIs: $\sigma^2_{\epsilon_0} = 0.4$, $\sigma^2_{\alpha_0} = 0.6\frac{p}{m}$ for $\theta = \sigma^2_\epsilon$, $(n, p) = (2000, 20000)$

m	% RB	var($\hat{\theta}$)	E($\hat{v}$)	s($\hat{v}$)	$\text{GCTA}_{0.01}$	$\text{GCTA}_{0.05}$	$\text{GCTA}_{0.1}$
20	3.306	0.009	0.010	0.004	0.987	0.947	0.907
200	−9.926	0.010	0.009	0.001	0.983	0.930	0.863
2, 000	7.653	0.009	0.009	0.001	0.990	0.950	0.910
20, 000	−3.772	0.010	0.009	0.001	0.990	0.943	0.893

Table 3:

$\sigma^2_{\epsilon_0} = 0.4$, $\sigma^2_{\alpha_0} = 0.6\frac{p}{m}$ for $\theta = h^2$, $(n, p) = (2000, 20000)$

m	% RB	var($\hat{\theta}$)	E($\hat{v}$)	s($\hat{v}$)	$N_{0.01}$	$N_{0.05}$	$N_{0.1}$	$T_{0.01}$	$T_{0.05}$	$T_{0.1}$
20	−33.676	0.014	0.009	0.000	0.953	0.880	0.810	0.953	0.880	0.810
200	−15.510	0.011	0.009	0.000	0.980	0.927	0.850	0.980	0.927	0.850
2, 000	4.060	0.009	0.0003	0.000	0.990	0.953	0.897	0.990	0.953	0.900
20, 000	−2.021	0.010	0.009	0.000	0.983	0.953	0.903	0.983	0.953	0.903

Table 4:

GCTA CIs: $\sigma^2_{\varepsilon_0} = 0.4$, $\sigma^2_{\alpha_0} = 0.6\frac{p}{m}$ for $\theta = h^2$, $(n, p) = (2000, 20000)$

m	% RB	var($\hat{\theta}$)	E($\hat{v}$)	s($\hat{v}$)	$\text{GCTA}_{0.01}$	$\text{GCTA}_{0.05}$	$\text{GCTA}_{0.1}$
20	−33.624	0.014	0.009	0.001	0.950	0.883	0.813
200	−15.199	0.011	0.009	0.001	0.977	0.927	0.847
2, 000	4.071	0.009	0.009	0.001	0.987	0.957	0.900
20, 000	−1.804	0.010	0.009	0.001	0.990	0.947	0.903

Table 5:

$\sigma^2_{\epsilon_0} = 0.4$, $\sigma^2_{\alpha_0} = 0.6\frac{p}{m}$ for $\theta = \sigma^2_\epsilon$, $\omega = m/p = 0.01$

(*n*, *p*)	% RB	var($\hat{\theta}$)	E($\hat{v}$)	s($\hat{v}$)	$N_{0.01}$	$N_{0.05}$	$N_{0.1}$	$T_{0.01}$	$T_{0.05}$	$T_{0.1}$
(1000, 10000)	−8.854	0.020	0.019	0.003	0.990	0.927	0.877	0.990	0.940	0.883
(2000, 20000)	−10.250	0.010	0.009	0.001	0.983	0.930	0.860	0.983	0.930	0.860
(3000, 30000)	5.450	0.006	0.006	0.001	0.997	0.957	0.907	0.997	0.957	0.907
(4000, 40000)	−16.138	0.005	0.005	0.000	0.983	0.917	0.870	0.983	0.917	0.870
(5000, 50000)	10.734	0.003	0.004	0.000	0.987	0.967	0.923	0.987	0.967	0.923

Table 6:

GCTA CIs: $\sigma^2_{\epsilon_0} = 0.4$, $\sigma^2_{\alpha_0} = 0.6\frac{p}{m}$ for $\theta = \sigma^2_\epsilon$, $\omega = m/p = 0.01$

(*n*, *p*)	**% RB**	var($\hat{\theta}$)	E($\hat{v}$)	s($\hat{v}$)	**GCTA$_{0.01}$**	**GCTA$_{0.05}$**	**GCTA$_{0.1}$**
(1000, 10000)	−8.477	0.020	0.019	0.004	0.993	0.930	0.877
(2000, 20000)	−9.926	0.010	0.009	0.001	0.983	0.93	0.863
(3000, 30000)	5.408	0.006	0.006	0.001	1.00	0.957	0.907
(4000, 40000)	−16.172	0.006	0.005	0.001	0.987	0.920	0.867
(5000, 50000)	10.868	0.003	0.004	0.001	0.987	0.963	0.923

Table 7:

$\sigma^2_{\epsilon_0} = 0.4$, $\sigma^2_{\alpha_0} = 0.6\frac{p}{m}$ for $\theta = h^2$, $\omega = m/p = 0.01$

(*n*, *p*)	% RB	var($\hat{\theta}$)	E($\hat{v}$)	s($\hat{v}$)	$N_{0.01}$	$N_{0.05}$	$N_{0.1}$	$T_{0.01}$	$T_{0.05}$	$T_{0.1}$
(1000, 10000)	−12.968	0.022	0.019	0.001	0.987	0.913	0.873	0.990	0.917	0.877
(2000, 20000)	−15.510	0.011	0.009	0.000	0.980	0.927	0.850	0.980	0.927	0.850
(3000, 30000)	−2.363	0.006	0.006	0.000	0.993	0.957	0.893	0.993	0.957	0.893
(4000, 40000)	−22.084	0.006	0.005	0.000	0.983	0.910	0.860	0.983	0.910	0.860
(5000, 50000)	5.989	0.004	0.004	0.000	0.993	0.950	0.920	0.993	0.950	0.920

Table 8:

GCTA CIs: $\sigma^2_{\varepsilon_0} = 0.4$, $\sigma^2_{\alpha_0} = 0.6\frac{p}{m}$ for $\theta = h^2$, $\omega = m/p = 0.01$

(*n*, *p*)	**% RB**	var($\hat{\theta}$)	E($\hat{v}$)	s($\hat{v}$)	**GCTA$_{0.01}$**	**GCTA$_{0.05}$**	**GCTA$_{0.1}$**
(1000, 10000)	−12.611	0.022	0.019	0.001	0.990	0.910	0.870
(2000, 20000)	−15.199	0.011	0.009	0.001	0.977	0.927	0.847
(3000, 30000)	−2.393	0.007	0.006	0.000	0.993	0.957	0.893
(4000, 40000)	−22.109	0.006	0.005	0.000	0.983	0.910	0.857
(5000, 50000)	6.140	0.004	0.004	0.000	0.993	0.950	0.923