Statistica Sinica Preprint No: SS-2025-0022	
Title	Simple Inferential Analyses of Big Gwas Data
Manuscript ID	SS-2025-0022
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202025.0022
Complete List of Authors	Jiming Jiang, Leqi Xu, Jiangshan Zhang and Hongyu Zhao
Corresponding Authors	Jiming Jiang
E-mails	jimjiang@ucdavis.edu
Notice: Accepted author version.	

SIMPLE INFERENTIAL HERITABILITY ANALYSES OF BIG GWAS DATA

Jiming Jiang[1], Leqi Xu[2], Jiangshan Zhang[1] and Hongyu Zhao[2]

University of California, Davis[1] and Yale University[2]

Abstract: **We derive simple, closed-form estimators of variance components of genetic interest, including the heritability and variance of environmental errors, in genome-wide association studies (GWAS) involving Biobank-size number of individuals. Consistency and asymptotic normality of the proposed estimators are established. Inferential analyses, including confidence intervals and hypothesis testing, for variance components of genetic interest are developed based on the asymptotic distribution. The method has significant advantage over the existing BOLT-REML method, designed for such Big GWAS data, in that the latter is not capable of carrying out the inferential analyses. The new method also has potentially computational advantage. Finite-sample performance of the proposed method, both in terms of statistical properties of the proposed estimators and confidence intervals and in terms of computational efficiency, is studied empirically and compared with the BOLT-REML method. Empirical comparison also shows that our method has significant computational advantage over a moment-matching method called mmhe, and similar statistical performance as mmhe. While most of the theoretical results are established under the assumption of independent single nucleotide polymorphisms (SNPs), we demonstrate**

Jiming Jiang's ORCID ID: 0000-0001-6364-4717.

how to extend the results to the case of C-dependent SNPs. A real-data example of the UK Biobank data is discussed.

Key words and phrases: ANOVA, Big GWAS Data, heritability, proportion of causal SNPs, random matrix theory, variance components

1. Introduction

Modern data science is characterized by high-dimensional, massive data. This is no exception in genome-wide association studies (GWAS), which have been successful in scanning the genome for genetic variations, for example, single nucleotide polymorphisms (SNPs), that are associated with disease status and traits. Tens of thousands of SNPs have been identified to be associated with various diseases and traits. See, for example, Visscher *et al.* (2017) for a review. Recent years have seen the rapid accumulation of human genetic data; as a result, massive genetic databases have been built. For example, in 2017, the UK Biobank database already involved approximately half a million individuals genotyped at nearly one million SNPs (Bycroft *et al.* 2018). Such a massive database presents computational challenges to the current GWAS analysis tools and software packages. For example, BOLT-REML (Loh *et al.* 2015) was developed to deal with the computational challenge of variance components estimation in case of Big GWAS data (see below). However, computation is not

the only thing that matters. Measure of uncertainty, and associated inferential tools are also important, which are characteristics of statistical data science.

The BOLT-REML method was developed to overcome the computational challenges in implementing the restricted maximum likelihood (REML) method in the analysis of linear mixed models (LMM; e.g., Jiang and Nguyen 2021, ch. 1). The computational challenge is mainly due to the large number of individuals, n, rather than the number of SNPs, p, even though the latter can be much larger than n. This is because the restricted log-likelihood, whose maximizer is the REML estimator, involves the inverse of an $(n-q)\times(n-q)$ matrix, where q is the rank of the matrix of covariates, X (e.g., Jiang and Nguyen 2021, p. 15). Here, q is a fixed number, which is typically small in GWAS. Importantly, the matrix cannot be simplified (e.g., having a block-diagonal structure) due to the nature of the GWAS LMM, in which the same set of SNP-specific random effects are shared by all the individuals. Thus, one essentially needs to invert an $n\times n$ matrix, which is computationally challenging if n is Biobank size, in the scale of hundreds of thousands.

More importantly, although asymptotic distribution of the REML estimator is well known under the standard LMM (e.g., Jiang and Nguyen 2021, sec. 1.8.3), the LMM in GWAS is different from the standard LMM. Namely, a standard LMM is conditional on X, and the design matrix, Z, associated with the

random effects. In other words, the X and Z matrices are considered fixed—there is no variation associated with these that needs to be considered when assessing the uncertainty. The LMM in GWAS, however, assumes the entries of Z are random variables associated with the genome-wide information. This additional variation needs to be taken into consideration when assessing the uncertainty. Such an uncertainty measure, and the resulting asymptotic distribution, cannot be straightforwardly obtained from the standard LMM inferential tools, and this is why such inferential tools have not been equipped with BOLT-REML. Although there has been recent studies in establishing the asymptotic distribution of the REML estimator under the GWAS LMM (e.g., Jiang *et al.* 2016), the asymptotic variance of the distribution is not in a form ready to be implemented for inference. Furthermore, the asymptotic variance of the REML estimator of the heritability involves an additional unknown parameter, ω, which is the proportion of causal SNPs. Unless ω is known, or can be consistently estimated, inference based on the asymptotic distribution of the REML heritability estimator cannot be executed practically. See Dao *et al.* (2022) for an alternative approach to the inferential problem.

In this paper, we address the two issues, namely, the computational challenge of Big GWAS data and the inferential problem noted above, with a simple approach. Simplicity is a stone that kills two birds. On the one hand, a simpler

method is computationally more attractive, and therefore more suitable for Big data. On the other hand, because our estimators are (much) simpler, we are able to obtain simple, explicit expressions of the asymptotic distribution, including the asymptotic covariance matrix of the vector of the proposed variance component estimators. From the expressions of the asymptotic covariance matrix, one can clearly see how the asymptotic variances and covariances depend on ω, or $\psi = \omega^{-1}$. Furthermore, a consistent estimator of ψ is proposed, based on which, and the estimators of the variance components, one can obtain asymptotically correct confidence intervals and critical values for hypothesis testing for the variance components of interest.

Before maximum likelihood (ML) and REML methods became popularize in the analysis of LMM, a method of fitting the LMM was well known as the analysis of variance (ANOVA; see Searle, Casella and McCulloch 1992). The basis for ANOVA is the method of moments (MoM), focusing on quadratic forms of the data that are "orthogonal" to the fixed effects. The MoM method is known to be inferior to the ML method in terms of the efficiency, and this is a main reason that the ANOVA method, once dominated the field of mixed model analysis, eventually gave way to ML and REML. However, despite the admiration of the ML in the field of statistics, MoM and its generalized version known as the generalized method of moments (GMM) are widely used, for ex-

ample, in econometrics (e.g., Hansen 1982, McFadden 1989). A main reason is that, in large sample, the difference between ML and GMM becomes practically insignificant. In economic studies, data are often collected in large sample. Furthermore, the GMM estimator is often easier, sometimes much easier, to compute than the ML estimator (e.g., Jiang and Nguyen 2021, ch. 4). This feature is attractive when dealing with large sample, or Big data, of the modern era. For similar reasons, there are renewed interests in the ANOVA-type methods in the analysis of Big GWAS data.

For example, Pazokitoroudi *et al.* (2020) proposed an extended version of the Haseman-Elston regression (Haseman and Elston 1972) combined with a randomized algorithm, called RHE-mc, to estimate a large number of variance components with Big GWAS data. The coefficients of the resulting estimating equations are quadratic forms of the phenotype data and traces of products of genotype matrices; in other words, the estimating equations are ANOVA type. However, Pazokitoroudi *et al.* did not study theoretical properties of their estimators. In terms of inference, the latter authors used a block jackknife method, which is not theoretically justified in the situation of Big GWAS data, neither does it take into account the additional variation of the randomized algorithm that was used. The jackknife method is also likely to increase the computational burden (due to the repeated computation), which is an important issue in dealing

with Big data. Also, the RHE-mc method cannot be directly applied when fixed effects are present in the model. Also see Wu and Sankararaman (2018) and Hou *et al.* (2019) for some related earlier work.

Similarly, ANOVA is the basis for deriving our simple, closed-form estimators of the variance components in GWAS. However, unlike RHE-mc, our methods, which can also incorporate fixed effects, are theoretically well justified. Furthermore, along with our simple estimators, we provide an equally simple, and theoretically justified inference procedure, including the standard errors and confidence intervals. It should be noted that, although our estimators appear in simple forms (see below), justification of their asymptotic properties is not straightforward under the GWAS LMM setting. In fact, random matrix theory (RMT) and martingale limit theory are used to establish consistency and asymptotic normality of the estimators of variance components of genetic interest. It is also worthy of noting that finding a consistent estimator of ω, or equivalently ψ, was a previously unsolved challenging problem (e.g., Jiang *et al.* 2016).

The ANOVA estimators are derived in Section 2 and their asymptotic behaviors are studied. Estimation of ψ is considered in Section 3 and consistency of $\hat{\psi}$ is established. The results lead to inference about τ^2 and h^2, such as confidence intervals with asymptotically correct coverage probability. Extensions of the results are discussed in Section 4. Some results of simulation studies are

presented in Section 5, including comparison with BOLT-REML in terms of statistical performance as well as performance of the ANOVA confidence intervals. A real-data application to Big GWAS data is discussed in Section 6, which also provides comparison of ANOVA and BOLT-REML in terms of the computational efficiency. Some discussion and remarks are offered in Section 7. Proofs of the theoretical results are deterred to a supplement.

2. Some highlights of results

Before engaging in the detailed developments, we would like to present some highlights of the results obtained in this paper. For simplicity, here we consider a simple form of the GWAS model; extensions will be considered in the sequel. Following Jiang *et al.* (2016), a GWAS model can be expressed as

$$y = X\beta + \tilde{Z}\alpha + \epsilon, \tag{2.1}$$

where X is an $n \times q$ matrix of known covariates such that $\mathrm{rank}(X) = q$, β is a vector of unknown fixed effects, $\tilde{Z} = p^{-1/2}Z$, $\alpha = (\alpha_j)_{1\le j\le p}$, $\epsilon = (\epsilon_i)_{1\le i\le n}$. Here α corresponds to the (unobservable) SNP-specific random effects and ϵ the environmental errors. For now we assume that the entries of Z are independent $N(0,1)$; $\alpha_j = b_j\zeta_j$, where $b_j, 1 \le j \le p$ are independent $\mathrm{Bernoulli}(\omega)$, $\zeta_j, 1 \le j \le p$ are independent $N(0,\sigma^2)$; $\epsilon_i, 1 \le i \le n$ are independent $N(0,\tau^2)$; and Z, $b = (b_j)_{1\le j\le p}$, $\zeta = (\xi_j)_{1\le j\le p}$, and ϵ are independent.

Again, for simplicity, here we present the results for the special case of $X = 0$; an extension will be considered in the sequel. Denote $S_z = \tilde{Z}\tilde{Z}' = p^{-1}ZZ'$ and $S_{z,\mathrm{o}} = S_z - dS_z$, where for any matrix $A = (a_{jk})_{1\le j,k\le p}$, $dA = \mathrm{diag}(a_{jj}, 1 \le j \le p)$. Then, consistent estimators of $\omega\sigma^2$ and τ^2 are given, respectively, by

$$\hat{\sigma}^2 = \frac{y'S_{z,\mathrm{o}}y}{\mathrm{tr}(S_{z,\mathrm{o}}^2)} \quad \text{and} \quad \hat{\tau}^2 = \frac{y'y - \hat{\sigma}^2\mathrm{tr}(S_z)}{n}. \tag{2.2}$$

A consistent estimator of the heritability, $h^2 = \omega\sigma^2/(\tau^2 + \omega\sigma^2)$, is given by

$$\hat{h}^2 = \frac{\hat{\sigma}^2}{\hat{\tau}^2 + \hat{\sigma}^2}. \tag{2.3}$$

Furthermore, as $n, p \to \infty$ such that

$$\frac{n}{p} \to \gamma \in (0, \infty), \tag{2.4}$$

we have

$$\sqrt{n}\begin{pmatrix} \hat{\tau}^2 - \tau^2 \\ \hat{h}^2 - h^2 \end{pmatrix} \xrightarrow{\mathrm{d}} N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \gamma_{11} & \gamma_{12} \\ \gamma_{12} & \gamma_{22} \end{pmatrix}\right], \tag{2.5}$$

where $\gamma_{11} = 2\{\tau^4 + (\omega\sigma^2)^2 + \gamma^{-1}(\tau^2 + \omega\sigma^2)^2\}$, $\gamma_{12} = -2\{\omega\sigma^2 + \gamma^{-1}(\tau^2 + \omega\sigma^2\}$, and $\gamma_{22} = 2\left\{\frac{1}{\gamma} + h^2 + h^2(1-h^2)\right\} + \gamma\left(\frac{3}{\omega} - 1\right)h^4(1-h^2)^2$.

It is seen that $\psi = \omega^{-1}$ is involved in the asymptotic variance of $\hat{h}^2$. A consistent estimator of ψ is given by $\hat{\psi} = \tilde{\psi} \vee 1$, where

$$\tilde{\psi} = \frac{1}{\hat{\sigma}^4 c_3}\left\{\frac{nM}{n+2} - (\hat{\tau}^4 + 2\hat{\tau}^2\hat{\sigma}^2 c_1 + \hat{\sigma}^4 c_2)\right\} \tag{2.6}$$

with $M = (p/3n^2)\sum_{j=1}^{p} r_j^4$, r_j being the jth component of $r = \tilde{Z}'y$, and

$$c_1 = \frac{n}{p} + 1 + \frac{3}{p},$$
$$c_2 = \left(1 - \frac{1}{p}\right)\left\{2\left(\frac{n}{p}\right) + 1 + \frac{6}{p}\right\},$$
$$c_3 = \left(\frac{n}{p}\right)^2 + 10\left(\frac{n}{p^2}\right) + \frac{3}{p} + \frac{21}{p^2}.$$

In contrast, the REML estimators of τ^2 and $\omega\sigma^2$ do not have closed-form expressions. The asymptotic variances of the REML estimators have no analytical expressions, involving complex integrals that are difficult to analyze either theoretically or numerically (Jiang *et al.* 2016). In fact, asymptotic normality of the REML estimator of the heritability was not established, because it relies on the joint asymptotic normality of the REML estimators of τ^2 and $\omega\sigma^2$, which was not established. Also, the variance of the asymptotic distribution of the REML estimator of h^2 is expected to depend on the additional unknown parameter ω (or ψ), whose consistent estimator was not available (Jiang *et al.* 2016).

A modification of the ANOVA estimators, which is computationally further simplified, can be obtained as follows. It can be shown via the RMT (e.g., Jiang 2022, Corollary 16.2) that $n^{-1}\mathrm{tr}(S_z) \xrightarrow{\text{a.s.}} 1$ and $n^{-1}\mathrm{tr}(S_{z,\mathrm{o}}^2) \xrightarrow{\text{P}} 1 + \gamma$. Thus, again by (2.4), it can be shown that the ANOVA estimators, (2.2), are asymptotically equivalent to the following simplified ANOVA estimator:

$$\hat{\sigma}_*^2 = \left(\frac{p}{n^2}\right) y' S_{z,\mathrm{o}} y, \quad \hat{\tau}_*^2 = \frac{y'y}{n} - \hat{\sigma}_*^2. \tag{2.7}$$

(2.7) avoids the trace calculations in (2.2), especially $\mathrm{tr}(S_{z,\mathrm{o}}^2)$, which may still present some computationally challenge when n is large.

Furthermore, the above consistency and asymptotic results are extended to allow correlations in the SNPs. See the subsequent sections for details.

3. ANOVA estimation

As it turns out, much of the development can be done by focusing on the simpler case with $X = 0$. An extension can be made relatively straightforwardly to the general case of model (2.1). We therefore follow the strategy by first focusing on the simpler case.

3.1 Simple case

Let $Z = (z_{ij})_{1\le i\le n, 1\le j\le p}$, and $A = (a_{i_1 i_2})_{1\le i_1, i_2\le n}$ be an $n \times n$ matrix that depends on Z only. Then, we have

$$\mathrm{E}(y'Ay|Z) = \sum_{i_1,i_2=1}^{n} a_{i_1 i_2}\mathrm{E}(y_{i_1}y_{i_2}|Z). \tag{3.8}$$

Furthermore, by (2.1) with $X = 0$, we have

$$\mathrm{E}(y_{i_1}y_{i_2}|Z) = \sigma^2 s_{z,i_1,i_2} + \tau^2 1_{(i_1=i_2)}, \tag{3.9}$$

where $s_{z,i_1,i_2} = p^{-1}\sum_{j=1}^{p} z_{i_1 j}z_{i_2 j}$. Combining (3.8), (3.9), we have

$$\mathrm{E}(y'Ay|Z) = \sigma^2 \sum_{i_1,i_2=1}^{n} a_{i_1 i_2}s_{z,i_1,i_2} + \tau^2\mathrm{tr}(A). \tag{3.10}$$

ANOVA estimators of σ^2, τ^2 are obtained by solving two ANOVA equations, each of which is in the form of (3.10) with the conditional expectation sign on the left side removed. If we choose $A = I_n$, we obtain the first ANOVA equation:

$$\sigma^2 \mathrm{tr}(S_z) + \tau^2 n = y'y, \tag{3.11}$$

where $S_z = p^{-1}ZZ' = (s_{z,i_1,i_2})_{1\le i_1,i_2\le n}$. If we let $A = S_{z,\mathrm{o}}$, defined above (2), we obtain the second ANOVA equation:

$$\sigma^2 \sum_{i_1\neq i_2} s^2_{z,i_1,i_2} = \sum_{i_1\neq i_2} s_{z,i_1,i_2} y_{i_1} y_{i_2}. \tag{3.12}$$

The estimators (2.2), (2.3) are obtained by solving (3.11), (3.12).

Note 1. It might be thought that one could, perhaps, search over different pairs of A matrices in order to find an "optimal" pair of equations. However, this likely will increase the complexity of the resulting estimator, as there seems to be no simple "optimal choice". As our intention is to develop methods that are suitable to Big data, we are satisfied with the estimators (2.2), (2.3), not only for their simplicity but also for their sound theoretical properties, as stated below.

Theorem 3.1. Suppose that $n, p \to \infty$ such that (2.4) holds, then the following hold: (I) $\hat\tau^2 \xrightarrow{\mathrm{P}} \tau^2$, $\hat\sigma^2 \xrightarrow{\mathrm{P}} \omega\sigma^2$; and (II)

$$\sqrt{n}\begin{pmatrix} \hat\tau^2 - \tau^2 \\ \hat\sigma^2 - \omega\sigma^2 \end{pmatrix} \xrightarrow{\mathrm{d}} N(0, \Sigma), \tag{3.13}$$

where $\Sigma = (\sigma_{st})_{1\le s,t\le 2}$ with

$$\sigma_{11} = \gamma_{11} = 2\left\{\tau^4 + (\omega\sigma^2)^2 + \frac{(\tau^2+\omega\sigma^2)^2}{\gamma}\right\}, \tag{3.14}$$

$$\sigma_{12} = \sigma_{21} = -2\left\{2(\omega\sigma^2)^2 + \frac{(\tau^2+\omega\sigma^2)^2}{\gamma}\right\}, \tag{3.15}$$

$$\sigma_{22} = \left[4\tau^2 + \left\{8 + \left(\frac{3}{\omega} - 1\right)\gamma\right\}\omega\sigma^2\right]\omega\sigma^2 + 2\frac{(\tau^2+\omega\sigma^2)^2}{\gamma}. \tag{3.16}$$

Corollary 3.1. Under the conditions of Theorem 3.1, we have

$$\sqrt{n}(\hat\tau^2 - \tau^2) \xrightarrow{\mathrm{d}} N(0,\sigma_{11}) \text{ and } \sqrt{n}(\hat\sigma^2 - \omega\sigma^2) \xrightarrow{\mathrm{d}} N(0,\sigma_{22}),$$

where σ_{11} and σ_{22} are given by (3.14) and (3.16), respectively.

Note 2. It is seen that the asymptotic variance of $\hat\tau^2$, σ_{11}, and the asymptotic covariance between $\hat\tau^2$ and $\hat\sigma^2$, σ_{12}, depend only on τ^2, $\omega\sigma^2$ and γ. By (I) of Theorem 3.1, consistent estimators of τ^2 and $\omega\sigma^2$ are already available; note that γ can be approximated by n/p. On the other hand, the asymptotic variance of $\hat\sigma^2$, σ_{22}, involves ω, in addition to τ^2, $\omega\sigma^2$ and γ. Thus, in order to obtain a consistent estimator of the asymptotic variance of $\hat\sigma^2$, a consistent estimator of ω, or $\psi = \omega^{-1}$, is needed. This is discussed in the next section.

In addition to τ^2, another variance component of genetic interest is the heritability h^2. By applying the delta method, we obtain the following.

Corollary 3.2. Under the conditions of Theorem 3.1, we have $\hat h^2 \xrightarrow{\mathrm{P}} h^2$, where $\hat h^2$ is given by (2.3). Furthermore, (2.5) holds.

Note 3. It is also seen that the asymptotic covariance of $\hat{\tau}^2$ and $\hat{h}^2$ depends on τ^2, $\omega\sigma^2$ and γ only; on the other hand, the asymptotic variance of $\hat{h}^2$ depends on ψ, in addition to h^2 and γ. Thus, again, a consistent estimator of ψ is in need.

3.2 General case

Now let us extend the method to allow covariates. Suppose that, in (1), X is non-random. Consider $P_X = X(X'X)^{-1}X'$ and let

$$P_{X^\perp} = I_n - P_X = T\mathrm{diag}(1, \ldots, 1, 0, \ldots, 0)T'$$

be the eigenvalue decomposition of $P_{X^\perp}$, where the first $n-q$ eigenvalues are 1, and T is an orthonormal matrix. Note that $P_{X^\perp}$ is idempotent, whose eigenvalues are therefore 0 or 1. Let $T = (t_1 \ \cdots t_n)$ (t_j being the jth column of T), and $L = (t_1 \ \cdots t_{n-q})$. Consider a REML-style transformation:

$$u = L'y = L'\tilde{Z}\alpha + L'\epsilon = \tilde{W}\alpha + \delta, \tag{3.17}$$

where $\tilde{W} = p^{-1/2}W$ with $W = L'Z = [w_{ij}]_{1\le i\le n-q, 1\le j\le p}$. Note that $L'X = 0$ and $L'L = I_{n-q}$. Following the same steps, we have, for any symmetric matrix $B = (b_{i_1 i_2})_{1\le i_1, i_2\le n-q}$ that depends only on Z,

$$\mathrm{E}(u'Bu|Z) = \sigma^2 \sum_{i_1, i_2=1}^{n-q} b_{i_1 i_2} s_{w, i_1, i_2} + \tau^2 \mathrm{tr}(B), \tag{3.18}$$

where s_{w,i_1,i_2} is s_{z,i_1,i_2} with z replaced by w. By letting $B = I_{n-q}$, we obtain

$$\sigma^2 \mathrm{tr}(S_w) + \tau^2(n-q) = u'u, \tag{3.19}$$

where $S_w = p^{-1}WW' = L'S_zL$. By letting $B = S_w - dS_w \equiv S_{w,\mathrm{o}}$, we obtain

$$\sigma^2 \sum_{i_1 \neq i_2} s^2_{w,i_1,i_2} = \sum_{i_1 \neq i_2} s_{w,i_1,i_2} u_{i_1} u_{i_2}. \tag{3.20}$$

(3.19), (3.20) lead to the following estimators:

$$\hat{\sigma}^2 = \frac{u' S_{w,\mathrm{o}} u}{\mathrm{tr}(S^2_{w,\mathrm{o}})}, \tag{3.21}$$

$$\hat{\tau}^2 = \frac{u'u - \hat{\sigma}^2 \mathrm{tr}(S_w)}{n-q}. \tag{3.22}$$

Note that $\tilde{W} = p^{-1/2}W$ with $W = (W_1 \ \cdots \ W_p)$ and $W_j = L'Z_j, 1 \leq j \leq p$. Note that $Z_1, \ldots, Z_p$ are independent $N(0, I_{n-q})$; in other words, the entries of W are independent $N(0,1)$. Furthermore, we have $\delta = L'\epsilon \sim N(0, \tau^2 I_{n-q})$. Thus, with y replaced by u, and Z replaced by W, we are in the same situation as in the simple case, with n replaced by $n-q$. Therefore, Theorem 3.1 and its corollaries extend immediately, provided that

$$q = o(n). \tag{3.23}$$

We state the results below for the sake of completeness.

Theorem 3.2. Suppose that X is non-random and (3.23) holds. Then, as $n, p \to \infty$ such that (2.4) holds, the conclusions (I), (II) of Theorem 3.1 hold with $\hat{\sigma}^2$ and $\hat{\tau}^2$ given by (3.21) and (3.22), respectively, and the same Σ as in Theorem 3.1. In particular, the conclusions of Corollary 3.1 and Corollary 3.2 hold with the new $\hat{\tau}^2, \hat{\sigma}^2$, and $\hat{h}^2$ correspondingly, and the same $\gamma_{st}, 1 \leq s \leq t \leq 2$ as in (2.5).

4. Estimation of proportion of causal SNPs and inference

Again, we first consider $X = 0$ and then generalize the results.

4.1 Simple case

The vector r defined below (2.6) is associated with the SNP-wide correlation between y and Z_j. We first establish the following result.

Lemma 4.1. For the jth component of r, r_j, we have

$$\frac{\mathrm{E}(r_j^4|Z)}{3} = \frac{\tau^4}{n^2}(Z_j'Z_j)^2 + 2\frac{\tau^2\omega\sigma^2}{n^2p}Z_j'Z_jZ_j'ZZ'Z_j$$
$$+\frac{(\omega\sigma^2)^2}{n^2p^2}\sum_{1\le k\ne l\le p}(Z_j'Z_k)^2(Z_j'Z_l)^2 + \frac{(\omega\sigma^2)^2\psi}{n^2p^2}\sum_{k=1}^{p}(Z_j'Z_k)^4.$$

Recall the M defined below (2.6). Furthermore, let

$T_1 = (n^2p)^{-1}\sum_{j=1}^p (Z_j'Z_j)^2$,

$T_2 = (n^2p^2)^{-1}\sum_{j=1}^p Z_j'Z_jZ_j'ZZ'Z_j$,

$T_3 = (n^2p^3)^{-1}\sum_{j=1}^p\sum_{1\le k\ne l\le p}(Z_j'Z_k)^2(Z_j'Z_l)^2$,

and $T_4 = (n^2p^3)^{-1}\sum_{j,k=1}^p (Z_j'Z_k)^4$. By Lemma 4.1, we have

$$\mathrm{E}(M|Z) = \tau^4T_1 + 2\tau^2\omega\sigma^2T_2 + (\omega\sigma^2)^2T_3 + (\omega\sigma^2)^2\psi T_4. \tag{4.24}$$

equation (4.24) leads to an empirical method of moments (EMM; e.g., Jiang 2003) estimator of ψ: $\hat{\psi}_{\mathrm{c}} = \tilde{\psi}_{\mathrm{c}} \vee 1$, where

$$\tilde{\psi}_{\mathrm{c}} = \frac{M - (\hat{\tau}^4T_1 + 2\hat{\tau}^2\hat{\sigma}^2T_2 + \hat{\sigma}^4T_3)}{\hat{\sigma}^4T_4}. \tag{4.25}$$

Here, the subscript c refers to conditional estimator, because (24) is based on the expectation conditional on Z.

The computation of $\hat{\psi}_{\rm c}$ can be intensive for large n, and especially large p, due to the evaluations of $T_s, s = 1, 2, 3, 4$. Alternatively, one may consider the unconditional moment by taking another expectation of both sides of (24) (with respect to Z). This leads to (detailed computations are omitted)

$$\mathrm{E}(M) = \left(1 + \frac{2}{n}\right)\{\tau^4 + 2\tau^2\omega\sigma^2 c_1 + (\omega\sigma^2)^2 c_2 + (\omega\sigma^2)^2\psi c_3\}, \qquad (4.26)$$

where the constants c_1, c_2, c_3 are given in Section 1 [below (2.6)]. Equation (4.26) leads to the estimator $\hat{\psi}$ given in Section 2.

The following result establishes consistency of $\hat{\psi}$.

Theorem 4.1. Under the assumptions of Theorem 3.1 we have $\hat{\psi} \xrightarrow{\rm P} \psi$; in other words, $\hat{\psi}$ is a consistent estimator of ψ.

Note. From the proof given in the supplement, it is evident that the conditional estimator, $\hat{\psi}_{\rm c}$, is also consistent. Furthermore, simulation results (see Section 6.2) show that the two estimators, $\hat{\psi}_{\rm c}$ and $\hat{\psi}$, perform almost identically. We thus prefer the unconditional estimator, $\hat{\psi}$, based on its computational advantage (for big data) and nearly identical performance as the conditional estimator.

We now consider confidence intervals for τ^2 and h^2. By Corollary 3.1, we have $\sqrt{n/\hat{\sigma}_{11}}(\hat{\tau}^2 - \tau^2) \xrightarrow{\rm d} N(0, 1)$, where $\hat{\sigma}_{11}$ is the σ_{11} in (3.14) with $\tau^2, \omega\sigma^2$ replaced by $\hat{\tau}^2, \hat{\sigma}^2$, respectively, and γ replaced by n/p. It follows that an asymp-

totic $100(1-a)\%$ $(0<a<1)$ confidence interval for τ^2 is given by

$$\left[\hat{\tau}^2 - z_{a/2}\sqrt{\frac{\hat{\sigma}_{11}}{n}},\ \ \hat{\tau}^2 + z_{a/2}\sqrt{\frac{\hat{\sigma}_{11}}{n}}\right], \tag{4.27}$$

where $z_{a/2}$ is the critical value such that $\mathrm{P}(\zeta > z_{a/2}) = a/2$ for $\zeta \sim N(0,1)$.

By Corollary 3.2 and Theorem 4.1, we can construct a confidence interval for h^2 that has the asymptotically correct coverage probability. This is given by

$$\left[\hat{h}^2 - z_{a/2}\sqrt{\frac{\hat{\gamma}_{22}}{n}},\ \ \hat{h}^2 + z_{a/2}\sqrt{\frac{\hat{\gamma}_{22}}{n}}\right], \tag{4.28}$$

where $\hat{h}^2$ is the same as in (2.3), and $\hat{\gamma}_{22}$ is given below (2.5) with h^2, ω^{-1} and γ replaced by $\hat{h}^2$, $\hat{\psi}$ and n/p, respectively.

We summarize the results as follows.

Theorem 4.2. Under the conditions of Theorem 3.1, (4.27) and (4.28) have asymptotically correct coverage probability, $1-a$, respectively, for τ^2 and h^2.

4.2 General case

We can extend the results on estimation of ψ and confidence intervals to the general case of (1) in the same way as we did in Section 3.2. Namely, all we have to do is to replace the y in the previous subsection by u of (3.17). Note that the vector r is now equal to $\tilde{W}'u$. It can also be expressed in terms of y and Z, that is $r = \tilde{W}'u = \tilde{Z}'P_{X^\perp}y$. Note that $P_{X^\perp} = LL'$. We summarize the corresponding extended theoretical results for the sake of completeness.

Theorem 4.3. Under the assumptions of Theorem 3.2, we have $\hat{\psi} \xrightarrow{\mathrm{P}} \psi$, where $\hat{\psi}$ is defined the same way as in the previous subsection, with y, Z replaced by u, W, respectively. In other words, $\hat{\psi}$ is a consistent estimator of ψ.

Theorem 4.4. Under the conditions of Theorem 3.2, (4.27) and (4.28), with the new estimators $\hat{\tau}^2$, $\hat{\sigma}^2$, $\hat{h}^2$ and $\hat{\psi}$, have asymptotically correct coverage probability, $1 - a$, respectively, for τ^2 and h^2.

Similarly, one can derive critical values for testing hypotheses regarding the variance components that have asymptotically correct levels of significance.

5. Extension

We refer to Section 7 of the supplementary material for extensions of the results in other directions. In particular, an extension is considered in the supplement to relax the assumption of independent SNPs to the assumption that one has the decomposition $Z = \Gamma W$, where Γ is the square root of a covariance matrix, and W is a random matrix whose entries are independent $N(0, 1)$. This condition holds if the entries of Z are normal. In GWAS, however, the entries of Z are standardized $\mathrm{Bimonial}(2, f_j)$ variables, where f_j is the allele frequency corresponding to the jth SNP ($1 \leq j \leq p$; e.g., Jiang *et al.* 2016, sec. 1.2). Clearly, such entries are not even approximately normal.

Due to such a concern, we consider below another type of extension in terms

of the C-dependence (e.g., Jiang *et al.* 2023). The columns of Z, $Z_1, \ldots, Z_p$, are said to be C-dependent, where C is a constant, which may not be known, if for any subsets of $\{1, \ldots, p\}$, $J_1, \ldots, J_t$, such that $d(J_r, J_s) = \min_{j_r \in J_r, j_s \in J_s} |j_r - j_s| > C$, for $1 \leq r \neq s \leq t$, $[Z_j]_{j \in J_1}, \ldots, [Z_j]_{j \in J_t}$ are independent. Extension of the current results to the C-dependence situation is entirely possible, although the results are going to be lengthy, and "messy" (see below). The reason is that, under the assumption of independent SNPs, most of the limits involved are 0; the rest are 1 or something very simple. This is no longer the case in the C-dependent case; see below for some examples. Due to this reason, as well as the space limit, it is not possible to provide full extension of all of the results previously obtained here in this paper. Our strategy is to extend one of the main results, namely, Theorem 3.1, to demonstrate that the extension can be done, and showcase the main tools used in the extension via the proof given in the supplement. The rest of the extensions is, again, deferred to a future publication.

Define $\mathcal{S}_{1,p} = \{(j_1, j_2) : 1 \leq j_1, j_2 \leq p, |j_1 - j_2| \leq C\}$, $\bar{\mathcal{S}}_{1,p} = \{(j_1, j_2) : 1 \leq j_1, j_2 \leq p, |j_1 - j_2| > C\}$, and $d_{1,p} = |\bar{\mathcal{S}}_{1,p}|$, where $|S|$ denotes the cardinality of set S. We assume the following.

A1. $\mathrm{E}(z_{ij}) = 0$, $\mathrm{E}(z_{ij}^2) = 1$ and $\mathrm{E}(z_{ij}^8) \leq c$ for some constant $c > 0$, and $Z_1, \ldots, Z_p$ are C-dependent.

A2. The following expectations do not depend on the index i: $\mathrm{E}(z_{ij_1} z_{ij_2}) =$

$\mathrm{E}(z_{1j_1}z_{1j_2}) \equiv r_{2,j_1j_2}$, $\mathrm{E}(z_{ij_1}^2 z_{ij_2}^2) = \mathrm{E}(z_{1j_1}^2 z_{1j_2}^2) \equiv r_{4,j_1j_2}$, $\mathrm{E}(z_{ij_1}z_{ij_2}^2 z_{ij_3}) = \mathrm{E}(z_{1j_1}z_{1j_2}^2 z_{1j_3}) \equiv r_{4,j_1j_2j_3}$; $\mathrm{E}(z_{ij_1}z_{ij_2}z_{ij_3}z_{ij_4}) = \mathrm{E}(z_{1j_1}z_{1j_2}z_{1j_3}z_{1j_4}) \equiv r_{4,j_1j_2j_3j_4}$,

$\mathrm{E}(z_{ij_1}^2 z_{ij_2}^2 z_{ij_3}^2) = \mathrm{E}(z_{1j_1}^2 z_{1j_2}^2 z_{1j_3}^2) \equiv r_{6,j_1j_2j_3}$,

$\mathrm{E}(z_{ij_1}z_{ij_2}z_{ij_3}^2 z_{ij_4}^2) = \mathrm{E}(z_{1j_1}z_{1j_2}z_{1j_3}^2 z_{1j_4}^2) \equiv r_{6,j_1j_2j_3j_4}$,

$\mathrm{E}(z_{ij_1}^2 z_{ij_2}^2 z_{ij_3}^2 z_{ij_4}^2) = \mathrm{E}(z_{1j_1}^2 z_{1j_2}^2 z_{1j_3}^2 z_{1j_4}^2) \equiv r_{8,j_1j_2j_3j_4}$,

A3. As $n, p \to \infty$, (2.4) holds; furthermore, we have $p^{-2}d_{1,p} \to d_1$;

$$\frac{1}{p}\sum_{j_1,j_2,j_3,j_4=1}^{p} r_{2,j_1j_2}r_{2,j_2j_3}r_{2,j_3j_4}r_{2,j_4j_1} \longrightarrow f_1;\ \frac{1}{p}\sum_{j_1,j_2=1}^{p} r_{2,j_1j_2}^2 \longrightarrow f_2;$$

$$\frac{1}{p^2}\sum_{j_1,j_2,j_3,j_4=1}^{p} r_{4,j_1j_2j_3}r_{2,j_3j_4}r_{2,j_4j_1} \longrightarrow f_3;\ \frac{1}{p^2}\sum_{j_1,j_2=1}^{p} r_{4,j_1j_2} \longrightarrow f_4;$$

$$\frac{1}{p^3}\sum_{j_1,j_2,j_3,j_4=1}^{p} r_{2,j_1j_2}r_{6,j_1j_2j_3j_4} \longrightarrow f_5;$$

$$\frac{1}{p}\sum_{j_1,j_2,j_3=1}^{p} r_{2,j_1j_2}r_{2,j_2j_3}r_{2,j_3j_1} \longrightarrow f_6;$$

$$\frac{1}{p^2}\sum_{j_1,j_2,j_3=1}^{p} (r_{2,j_1j_2}r_{4,j_2j_3j_1} + r_{2,j_2j_3}r_{4,j_3j_1j_2} + r_{2,j_3j_1}r_{4,j_1j_2j_3}) \longrightarrow f_7;$$

$$\frac{1}{p^3}\sum_{j_1,j_2,j_3=1}^{p} r_{6,j_1j_2j_3} \longrightarrow f_8;\ \frac{1}{p^3}\sum_{j_1,j_2,j_3,j_4=1}^{p} r_{4,j_1j_2j_3}r_{4,j_3j_4j_1} \longrightarrow f_9;$$

$$\frac{1}{p^3}\sum_{j_1,j_2,j_3,j_4=1}^{p} r_{4,j_1j_2j_3j_4}^2 \longrightarrow f_{10};\ \frac{1}{p^4}\sum_{j_1,j_2,j_3,j_4=1}^{p} r_{8,j_1j_2j_3j_4} \longrightarrow f_{11};$$

$$\frac{1}{p}\sum_{j_1=1}^{p}\left(\sum_{j_2=1}^{p} r_{2,j_1j_2}^2\right)^2 \longrightarrow f_{12};\ \frac{1}{p^2}\sum_{j_1,j_2=1}^{p}\left(\sum_{j_3=1}^{p} r_{2,j_1j_3}^2\right) r_{4,j_1j_2} \longrightarrow f_{13};$$

$$\frac{1}{p^3}\sum_{j_1=1}^{p}\left(\sum_{j_2=1}^{p} r_{4,j_1j_2}\right)^2 \longrightarrow f_{14}.$$

Regarding assumption *A1*, as noted in Jiang *et al.* (2023), the C-dependence may be viewed as an approximation to a correlation structure that is fast decaying as the genetic distance between the SNPs increases. For example, it is known that, under an auto-regressive (AR) correlation structure, the correlation decays exponentially fast so that, beyond a certain distance (the number C), the correlation can be practically ignored. The rest of the assumption *A1* are standardization and moment conditions. Assumption *A2* may be regarded as a weaker condition than the i.i.d. assumption. Namely, if the rows, $Z_i = (Z_{ij})_{1\le j\le p}$, $1 \le i \le n$ are i.i.d., then, all of the expectations involved in *A2* do not depend on i. Assumption *A3* is more of technical nature. Basically, these are some limits involved in the asymptotic covariance matrix (ACM) if the asymptotic distribution of the ANOVA estimators of the variance components. Among these limits, f_2 plays a particularly important role, and its positivity is essential because this quantity appears in the denominators of some elements involved in the ACM (i.e., λ_2). Note that, because $\sum_{j_1,j_2=1}^{p} r_{2,j_1j_2}^2 \ge \sum_{j=1}^{p} r_{2,jj}^2 = p$, by *A1*, we have $f_2 \ge 1$.

We can now state an extension of Theorem 3.1.

Theorem 5.1. Under assumptions *A1–A3*, the conclusions of Theorem 3.1, (I) and (II), hold with the new Σ given below:

$$\Sigma = \begin{pmatrix} \lambda_0 - 2\lambda_1 + \lambda_2 & \lambda_1 - \lambda_2 \\ \lambda_1 - \lambda_2 & \lambda_2 \end{pmatrix}, \tag{5.29}$$

where the definition of $\lambda_s, s = 0, 1, 2$ are given below:

$$\lambda_0 = \lambda_0(\sigma^2, \tau^2, \omega, \gamma, f, d_1)$$

$$\omega\{3\gamma(1-\omega) + 2(f_2\gamma + d_1)\omega\}\sigma^4 + 4\omega\sigma^2\tau^2 + 2\tau^4;$$

$$\lambda_1 = \lambda_1(\sigma^2, \tau^2, \omega, \gamma, f, d_1) =$$

$$\omega[3f_2\gamma(1-\omega) + (2/\gamma)\{f_6\gamma^2 + (f_7 - f_2)\gamma + f_8 - d_1\}\omega]\sigma^4 + 4\omega\sigma^2\tau^2;$$

$$\lambda_2 = \lambda_2(\sigma^2, \tau^2, \omega, \gamma, f, d_1) =$$

$$\omega[(1-\omega)(3/\gamma)\{f_{12}\gamma^2 + (f_2 - f_{13})\gamma + d_1 - 2f_4 + f_{14}\} + \omega(2/\gamma^2)\{f_1\gamma^3$$

$$+2(3f_3 - f_6)\gamma^2 + (f_2 + 4f_5 - 2f_7 + 2f_9 + f_{10})\gamma + d_1 - 2f_8 + f_{11}\}]\sigma^4$$

$$+(4/f_2^2\gamma^2)\{f_6\gamma^2 + (f_7 - 2f_2)\gamma + f_4 + f_8 - 2d_1\}\omega\sigma^2\tau^2 + (2/f_2\gamma)\tau^4.$$

The elements involved in the new asymptotic covariance matrix (ACM), Σ, are extensions of those involved in the ACM in Theorem 3.1 for the case of independent SNPs. Specifically, $\lambda_0 - 2\lambda_1 + \lambda_2$ is the extension of σ_{11} in Theorem 3.1, which is the asymptotic variance of $\sqrt{n}(\hat{\tau}^2 - \tau^2)$; $\lambda_1 - \lambda_2$ is the extension of σ_{12}, which is the asymptotic covariance between $\sqrt{n}(\hat{\tau}^2 - \tau^2)$ and $\sqrt{n}(\hat{\sigma}^2 - \omega\sigma^2)$; and λ_2 is the extension of σ_{22}, which is the asymptotic variance of $\sqrt{n}(\hat{\sigma}^2 - \omega\sigma^2)$. To further interpret λ_0 and λ_s, note that it is easy to verify the following alternative expression of Σ:

$$\Sigma = \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \lambda_0 & \lambda_1 \\ \lambda_1 & \lambda_2 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -1 & 1 \end{pmatrix}. \tag{5.30}$$

Also note that $\mathrm{var}(y_i) = \omega\sigma^2 + \tau^2 \equiv \sigma_y^2$ is the total variance of any individual component of y. Let $\hat{\sigma}_y^2 = \hat{\sigma}^2 + \hat{\tau}^2$. By Theorem 5.1 and (5.30), it follows that

$$\sqrt{n}\begin{pmatrix} \hat{\sigma}_y^2 - \sigma_y^2 \\ \hat{\sigma}^2 - \omega\sigma^2 \end{pmatrix} \xrightarrow{\mathrm{d}} N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \lambda_0 & \lambda_1 \\ \lambda_1 & \lambda_2 \end{pmatrix}\right]. \tag{5.31}$$

By (5.31), we can interpret λ_0 as the asymptotic variance of $\sqrt{n}(\hat{\sigma}_y^2 - \sigma_y^2)$, and λ_1 as the asymptotic covariance between $\sqrt{n}(\hat{\sigma}_y^2 - \sigma_y^2)$ and $\sqrt{n}(\hat{\sigma}^2 - \omega\sigma^2)$. As noted already, λ_2 is the asymptotic variance of $\sqrt{n}(\hat{\sigma}^2 - \omega\sigma^2)$.

6. Simulation studies

In GWAS, one usually uses linear/logistic regression or LMM to handle the covariates by adding them as fixed effect terms. It is then assumed that the LMM model under consideration is derived from a model already accounting for the covariates. In other words, one may understand the procedure as firstly projecting the phenotypes/genotypes into a space spanned by the covariates, then conducting the regression between the residuals of phenotypes and genotypes after the projection. Thus, one may focus on a LMM without the fixed effects (or the fixed effects are zeros). We follow this consideration in our simulation settings. Furthermore, the simulation studies intend to study the empirical inferential performance of the ANOVA estimator and to compare the empirical performance of ANOVA and BOLT-REML as well as a moment-matching method (mmhe, Ge *et*

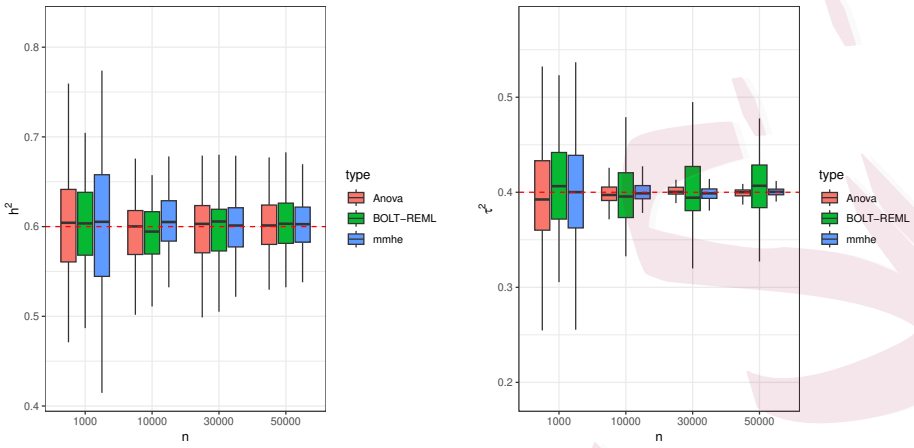
al. 2017). In addition, the empirical performance of the proposed ω estimator is also investigated. Our simulation design is similar to that of Jiang *et al.* (2016); see Section 5.1 of the Supplementary Material for details.

Our first simulation study is on the comparison of the ANOVA, BOLT-REML, and mmhe estimators. Figure 1 presents the side-by-side boxplots for the estimation of the heritability, h^2, and error variance, τ^2, where we fix $p = 1000$ and let n increase from 1000 to 50000. We set $h_{true} = 0.6$, $\tau^2 = 0.4$, $\sigma^2 = 6$, and $\omega = 0.1$. The results are based on 500 simulation runs.

It is seen that for the heritability estimation, the three methods perform similarly. More specifically, ANOVA and mmhe seem to have larger variation than BOLT-REML when $n = 1000$, but that seems to improve as n increase. In fact, for $n = 50000$, ANOVA seems to perform slightly better than BOLT-REML in both bias and variation. Overall, ANOVA and mmhe perform similarly.

A bigger difference is seen when it comes to estimation of τ^2. Again, ANOVA seems to have slightly larger variation when $n = 1000$, but it improves quickly, and dramatically, as n increases. For $n = 10000$, 30000 and 50000, ANOVA is seen to significantly outperform BOLT-REML in terms of the variation. In fact, the performance of BOLT-REML barely changes once n is larger than 1000, while the performance of ANOVA keeps improving, both in terms of bias and, more significantly, in terms of variation. Again, ANOVA and mmhe

estimators show a similar performance as n increases. It is worth noting that the computational time of ANOVA is much less than BOLT-REML and mmhe, especially when n is large. See the next section for a concrete demonstration.



(a) (b)

Figure 1: Comparison of ANOVA, BOLT-REML, and mmhe: $p = 1000$. (a) Heritability; (b) Environmental Variance (τ^2).

We also considered a scenario when all SNPs being causal and obtained similar results. See Section 5.2 of the Supplementary Material.

Our next simulation study focuses on performance of the confidence intervals based on the asymptotic theory. Because BOLT-REML does not provide variance estimation, or standard errors, we are unable to obtain the comparing results for BOLT-REML. Table 1 reports the results based on 500 simulation runs. It is seen that the coverage probability (CP) for both τ^2 and h^2 are close to

the nominal level of 95%. On the other hand, the mean length (ML) decreases with n for τ^2 but not for h^2. This can be explained by our asymptotic theory, namely, (2.5). Note that the ML of the confidence interval is associated with the asymptotic variance of the estimator. By (2.5), the asymptotic variance of $\hat{\tau}^2$ is γ_{11}/n, which decreases with n; as a result, the ML for τ^2 decreases with n, which is what is observed in Table 1. Again by (2.5), the asymptotic variance of $\hat{h}^2$ is γ_{22}/n, which is approximately equal to $(3/\omega - 1)h^4(1-h^2)^2/p$ plus something that is positive. Because p is fixed at 1,000 in this simulation setting, the asymptotic covariance has a constant lower bound. This explains why the ML for h^2 does not necessarily decrease with n in Table 1.

As noted (see Section 1, third paragraph), the BOLT-REML method is not equipped with tools for inferential analysis such as confidence intervals. It should be noted the mmhe method also provides standard errors for its estimators of τ^2 and h^2, which was used to obtain the p-values for hypothesis testing about h^2 in Ge *et al.* (2017). For comparison, we also constructed the mmhe confidence intervals for τ^2 and h^2 using the mmhe estimators and the associated standard errors. The results are also presented in Table 1. It can be seen that the CPs of the mmhe confidence intervals are seriously inaccurate (under-coverage) compared to those the of ANOVA confidence intervals.

Finally, we report some empirical results regarding the statistical perfor-

Table 1: Empirical Length and Coverage Probability of 95% Confidence Interval for τ^2 and h^2 with Fixed $p = 1000$, $\omega = 0.1$ and $h^2 = 0.6$

n	Anova				mmhe			
	τ^2		h^2		τ^2		h^2	
	ML	CP	ML	CP	ML	CP	ML	CP
1000	0.214	0.924	0.285	0.950	0.071	0.460	0.175	0.803
10000	0.044	0.946	0.168	0.978	0.007	0.248	0.018	0.210
30000	0.024	0.972	0.162	0.962	0.002	0.100	0.006	0.053
50000	0.018	0.958	0.161	0.984	0.001	0.115	0.004	0.028

mance of $\hat{\omega} = \hat{\psi}^{-1}$, the proposed estimator of ω, the proportion of causal SNPs. We have also studied the performance of $\hat{\omega}_c = \hat{\psi}_c^{-1}$, the conditional estimator (see Section 3.1), which performed almost identically as $\hat{\omega}$. Thus, only the simulation results for $\hat{\omega}$ are reported. Here, we consider scenarios where p is much larger than n, namely, $n = 2000$, $p = 20000$ and $n = 5000$, $p = 50000$. The summary statistics, based on 100 simulation runs, are presented in Tables 2 and 3. It is seen that the means are close to the true values in all cases. However, in terms of the mean, median and interquartile range ($= \text{Q3} - \text{Q1}$), the large n, p scenario is seen to lead to better performance. This is consistent with our asymptotic theory (Theorems 3.1 and 3.3).

Table 2: Estimation of $\hat{\omega}$ with $n = 2000$, $p = 20000$, and $h^2 = 0.6$

ω	Min	Q1	Median	Mean	Q3	Max
0.05	0.018	0.039	0.046	0.050	0.067	0.303
0.1	0.032	0.082	0.103	0.111	0.158	1.000
0.25	0.033	0.135	0.191	0.252	0.571	1.000
0.5	0.097	0.325	0.877	0.400	1.000	1.000

Table 3: Estimation of $\hat{\omega}$ with $n = 5000$, $p = 50000$, and $h^2 = 0.6$

ω	Min	Q1	Median	Mean	Q3	Max
0.05	0.021	0.041	0.051	0.049	0.059	0.100
0.1	0.049	0.095	0.109	0.107	0.142	0.503
0.25	0.093	0.178	0.236	0.250	0.384	1.000
0.5	0.133	0.298	0.529	0.427	1.000	1.000

Additional simulation results involving correlated SNPs (see Theorem 5.1) are deferred to the supplementary material.

7. A Big-data application

In this study, we applied our proposed ANOVA method to the UK Biobank, a large prospective study designed to elucidate the causes of complex diseases among middle-aged adults (Sudlow *et al.* 2015). We focused on estimating heritability for two traits of interest, BMI and height, using both ANOVA and BOLT-REML methods.

After excluding individuals with missing phenotype data and those of non-European ancestries, we retained 310,456 and 310,798 individuals with complete genotype and phenotype information for BMI and height, respectively. Since the ANOVA heritability estimation method inherently assumes a specific SNP correlation structure, SNP pruning was necessary before analysis. To accomplish this, we conducted GWAS to derive marginal association P-values for each SNP, identifying marginally significant associations using a threshold of 0.01. These significant SNPs served as representative markers for estimating heritability. Subsequently, we performed a sensitivity analysis to explore the impact of different SNP correlation thresholds (0, 2×10^{-5}, 4×10^{-5}, 6×10^{-5}, 8×10^{-5}, 1×10^{-4}) within a 100kb pruning window. Although these correlation

thresholds appear minor, they influenced the number of retained SNPs, thereby affecting the heritability estimates from both ANOVA and BOLT-REML.

A critical question arising from this analysis was determining the optimal correlation threshold for SNP pruning prior to applying ANOVA. To address this, we propose a principled and practical approach based on aligning ANOVA and BOLT-REML estimates. This approach leverages the established accuracy of BOLT-REML for SNP heritability estimation alongside the ability of ANOVA to provide confidence intervals, a feature not available with BOLT-REML. Specifically, we selected the correlation threshold at which the BOLT-REML heritability estimate fell within the 99% confidence interval of the ANOVA estimate.

This "matching" approach can be justified theoretically. According the theory established in this paper, when the SNPs are independent, the ANOVA estimator of the heritability is consistent. On the other hand, under the independent SNPs assumption, the REML estimator is also consistent (Jiang *et al.* 2016). Here, an important assumption is that the SNPs are independent. Although such an assumption has been relaxed for the ANOVA estimator (see the new Theorem 5.1), we do not yet know is this can be relaxed for the REML estimator. So, up to our current knowledge, we known when the SNPs are independent, the two estimators, ANOVA and BOLT-REML, are both consistent. If they are both consistent estimators, their values should be similar. More precisely, the

two estimators are consistent if the setting is right. Here, the right setting is that the SNPs are independent. In other words, if the right pruning threshold is chosen such that the SNPs are (nearly) independent, the values of the ANOVA and BOLT-REML estimates should be similar.

In Section 5.4 of the supplementary material, we carry out an additional simulation study to investigate the empirical performance of the matching strategy. The results suggests that the ANOVA estimator based on the matching-prune strategy performs similarly to the that based on an optimal pruning strategy known under the simulation study but not necessarily known in practice.

As illustrated in Figure 2, the ANOVA heritability estimate was lower than that from BOLT-REML at SNP correlation thresholds below 4×10^{-5}, whereas it exceeded the BOLT-REML estimate at thresholds above 4×10^{-5}. At exactly 4×10^{-5}, the estimates from ANOVA and BOLT-REML closely aligned, with the BOLT-REML estimate falling within the ANOVA 99% confidence interval. We therefore conclude that a correlation threshold of approximately 4×10^{-5} represents the appropriate pruning level for applying ANOVA to both traits.

Detailed ANOVA estimation results for SNP correlation at 4E-05 are also summarized in Table 4. The table includes the number of SNPs used for heritability estimation (n_SNPs), the heritability estimation results of ANOVA (h^2), the associated standard errors [sd(h^2)] and estimated proportion of causal SNPs

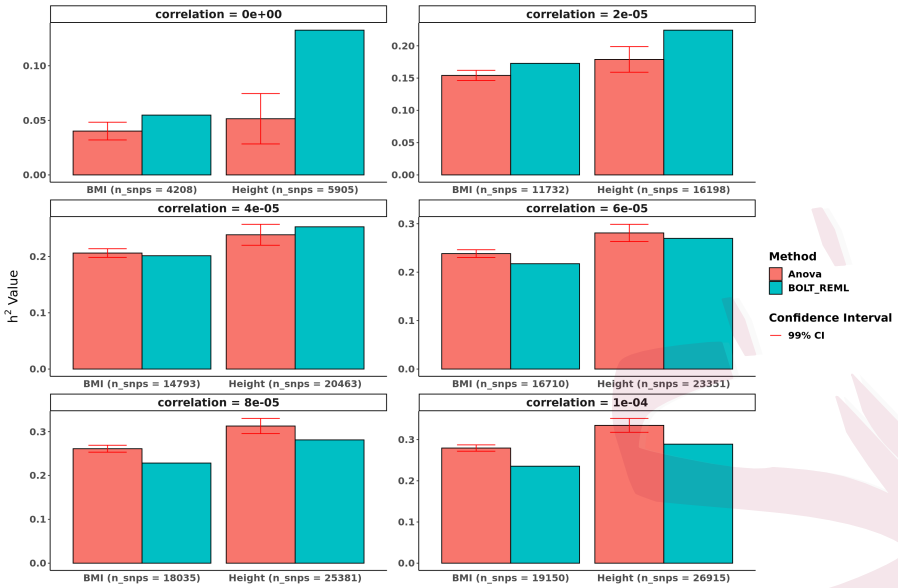


Figure 2: Heritability Estimation from ANOVA and BOLT-REML in BMI and Height

(ω). The results quantify the genetic influences on both BMI and height.

Importantly, the estimated heritability values reported here reflect only the genetic contributions of the selected SNPs filtered at a correlation threshold of 4×10^{-5}, and marginally significant associations identified by applying a threshold of 0.01, which likely underestimates the overall heritability attributable to all available SNPs. Although the proportion of causal SNPs (ω) for height (0.095) is smaller than for BMI (0.673), indicating fewer causal SNPs identified within significant SNPs for height, height is generally known to be highly polygenic with many causal SNPs of smaller effects, many of which may have been filtered out by our stringent p-value threshold. Future studies are required to extend

our method to consider all SNPs and accurately estimate the overall heritability attributable to genetic variation.

Table 4: ANOVA Real Data Results

trait	n_SNPs	h^2	$\mathrm{sd}(h^2)$	ω	trait	n_SNPs	h^2	$\mathrm{sd}(h^2)$	ω
BMI	14793	0.206	0.003	0.673	Height	20463	0.239	0.007	0.095

To further evaluate both statistical accuracy and computational efficiency, we compared heritability estimates obtained from ANOVA and BOLT-REML with those derived from the moment-matching method (mmhe). As illustrated in Figure 3, the heritability estimates from BOLT-REML consistently fell within the confidence intervals provided by ANOVA but not within those generated by mmhe. Given the established accuracy of BOLT-REML as a benchmark, the observed discrepancy raises questions regarding the reliability of the confidence intervals provided by mmhe. Moreover, this real-data application highlights the computational advantages of the ANOVA approach. As previously discussed, the computational demands of BOLT-REML primarily arise from large sample sizes (n), rather than the number of SNPs (p). With sample sizes exceeding 300,000 individuals per trait, our analysis confirmed that ANOVA substantially outperformed BOLT-REML in terms of computational speed. Additionally, ANOVA exhibited significantly improved runtime efficiency compared to

mmhe, which required approximately 10 hours to complete the same analysis.

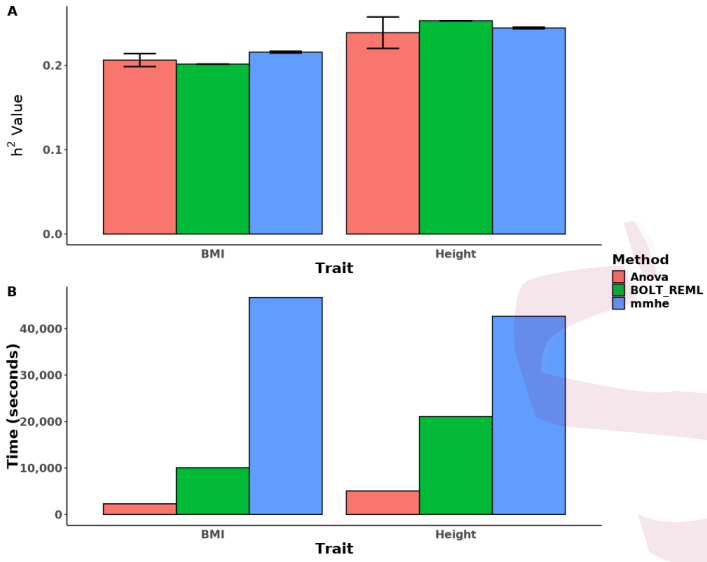


Figure 3: Heritability Estimates and Computation Times for BMI and Height Using ANOVA, BOLT-REML, and mmhe (SNP Correlation =4E-05)

8. Discussion

The increasing scale and complexity of GWAS datasets necessitate the development of statistical methods that are both computationally efficient and capable of delivering robust inferential analyses. In this study, we introduced a novel ANOVA-based framework for estimating heritability in the context of large-scale biobank datasets. Our method offers significant advancements over existing approaches, such as BOLT-REML, by enabling inferential analyses while

maintaining computational efficiency.

A key theoretical contribution of this work is the derivation of closed-form estimators, which ensures computational feasibility for biobank-scale data. Additionally, we established the consistency and asymptotic normality of these estimators, providing a solid theoretical foundation for their application in heritability estimation and inference. Through simulation studies, we demonstrated the finite-sample performance of our method, highlighting its advantages in statistical inference. Furthermore, our real-data application using the UK Biobank dataset underscores the practical utility of the proposed method, while traditional methods face computational and inferential challenges.

The current method may be extended to account for population stratification by incorporating genotype PCs similar in Lin *et al.* (2022). To account for the population substructure, we can incorporate the genotype PCs by first regressing them out of the phenotype using linear regression and subsequently using the residuals for heritability estimation. This approach effectively controls for confounding due to population stratification.

Despite its advantages, the proposed method has certain limitations. Like many statistical approaches in GWAS, it assumes weak correlations among genetic variants, an assumption that may not hold in the presence of linkage disequilibrium (LD). Although the extension of our results (see Section 4) allows

certain types of correlation structure, the conditions for the correlation structure may not be satisfied in a real-data situation. Future work should focus on extending the framework to account for highly correlated genetic variants, thereby enhancing its applicability to more complex genetic architectures.

In Section 5, we considered an extension of the consistency and asymptotic normality results to a case of correlated SNPs in terms of C-dependence. Extension beyond this setting is possible, although the development, and results, are expected to be more tedious. For example, a motivation for the C-dependence is when the between-SNPs correlation decays exponentially fast as the genetic distance between the SNPs increases. Thus, one way to extend the C-dependence is to assume that the correlation between two SNPs, which are distance d apart, is bounded in absolute value by $c\rho^d$, where $c > 0$ and $0 < \rho < 1$ are some constants. We shall explore such a further extension in our future work.

In conclusion, our ANOVA-based method represents a significant advancement in the analysis of variance components in GWAS. By addressing key computational and inferential challenges, it provides a robust and efficient solution for estimating genetic and environmental variance components in biobank-scale datasets. Its advantages over existing methods, such as BOLT-REML, in terms of statistical inference and computational efficiency make it a valuable tool for advancing genetic research. Moreover, we have shown that extension to correlated

SNPs under some realistic settings can be facilitated. Future work to broaden the extension to correlated genetic variants will further enhance its utility and robustness in diverse GWAS applications.

Supplementary Materials

The Supplementary Material contains proofs of the theoretical results.

Acknowledgements

Jiang's research was partially supported by the NSF grants DMS-1713120 and DMS-1914465. Zhao's research was partially supported by the NSF grant DMS-1713120 and NIH grants R01 GM134005 and R01 HG012735.

References

[1] Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., *et al.* (2018), The UK Biobank resource with deep phenotyping and genomic data, *Nature* 562, 203–209.

[2] Dao, C., Jiang, J., Paul, D. and Zhao, H. (2022), Variance estimation and confidence intervals from genome-wide association studies through high-dimensional misspecified mixed model analysis, *J. Stat. Plan. Infer.* 220, 15–23.

[3] Diggle, P. J., Heagerty, P., Liang, K. Y., and Zeger, S. L. (2002), *Analysis of Longitudinal Data*, 2nd ed., Oxford Univ. Press.

[4] Ge, T., Chen, CY., Neale, BM., Sabuncu, MR., Smoller, JW. (2017), Phenome-wide heritability

analysis of the UK Biobank. *PLOS Genetics* 13, e1006711.

[5] Hall, P. and Heyde, C. C. (1980), *Martingale Limit Theory and Its Application*, Academic Press, New York.

[6] Hansen, L. P. (1982), Large sample properties of generalized method of moments estimators, *Econometrica* 50, 1029–1054.

[7] Haseman, J. and Elston, R. (1972), The investigation of linkage between a quantitative trait and a marker locus, *Behav. Genet.* 2, 3–19.

[8] Hou, K., Burch, K. S., Majumdar, A., Shi, H., Mancuso, N., Wu, Y., Sankararaman, S., and Pasaniuc, B. (2019), Accurate estimation of SNP-heritability from biobank-scale data irrespective of genetic architecture, *Nature Genetics* 51, 1244–1251.

[9] Jiang, J. (2003), Empirical method of moments and its applications, *J. Statist. Planning Inference* 115, 69–84.

[10] Jiang, J. (2022), *Large Sample Techniques for Statistics*, 2nd ed., Springer, New York.

[11] Jiang, J. and Nguyen, T. (2021), *Linear and Generalized Linear Mixed Models and Their Applications*, 2nd ed., Springer, New York.

[12] 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.

[13] Jiang, J., Jiang, W., Paul, D., Zhang, Y., and Zhao, H. (2023), High-dimensional asymptotic behavior of inference based on GWAS summary statistics, *Stat. Sin.* 33, 1555–1576.

[14] Lin, Zhaotong, et al. (2022), Estimating SNP heritability in presence of population substructure in biobank-scale datasets, *Genetics* 220.4: iyac015.

[15] Loh, P. R., Bhatia, G., Gusev, A., Finucane, H. K., Bulik-Sullivan, B. K. *et al.* (2015), Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance components analysis, *Nature Genetics* 47, 1385–1392.

[16] McFadden, D. (1989), A method of simulated moments for estimation of discrete response models without numerical integration, *Econometrica* 57, 995–1026.

[17] Pazokitoroudi, A., Wu, Y., Burch, K.S. *et al.* (2020), Efficient variance components analysis across millions of genomes, *Nat. Commun.* 11, 4020. doi: 10.1038/s41467-020-17576-9.

[18] Searle, S. R., Casella, G., and McCulloch, C. E. (1992), *Variance Components*, Wiley, New York.

[19] Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A. , and Yang, J. (2017), 10 Years of GWAS Discovery: Biology, Function, and Translation, *Amer. J. Hum. Genet.* 101, 5–22.

[20] Wu, Y. and Sankararaman, S. (2018), A scalable estimator of SNP heritability for biobank-scale data, *Bioinformatics* 34, i187–i194.

[21] Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P. and others (2015), 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, e1001779.