Statistica Sinica Preprint No: SS-2021-0060	
Title	High-Dimensional Asymptotic Behavior of Inference Based on Gwas Summary Statistics
Manuscript ID	SS-2021-0060
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202021-0060
Complete List of Authors	Jiming Jiang, Wei Jiang, Debashis Paul, Yiliang Zhang and Hongyu Zhao
Corresponding Author	Jiming Jiang
E-mail	jimjiang@ucdavis.edu

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0060

Statistica Sinica

High-dimensional Asymptotic Behavior of Inference Based on GWAS Summary Statistics

Jiming Jiang, Wei Jiang, Debashis Paul, Yiliang Zhang and Hongyu Zhao

University of California, Davis and Yale University

Abstract: We study the high-dimensional asymptotic behavior of inferences based on summary statistics that are widely used in genome-wide association studies (GWAS) under model misspecification. The high dimensionality is in the sense that the number of single-nucleotide polymorphisms (SNPs) under consideration may be much larger than the sample size. The model misspecification is in the sense that the number of causal SNPs may be much smaller than the total number of SNPs under consideration. Specifically, we establish two parameters of genetic interest, namely, the consistency and asymptotic normality of the estimators of the heritability and genetic covariance. Our theoretical results are supported by the findings of empirical studies involving simulated and real data.

Key words and phrases: asymptotic normality, Bernoulli, consistency, genetic covariance, heritability, martingale, model misspecification, random matrix theory

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0060

1. Introduction

Over the past 15 years, genome-wide association studies (GWAS) have identified tens of thousands of single-nucleotide polymorphisms (SNPs) associated with complex human traits and diseases (Buniello et al. (2019)). In addition to the success in finding risk loci, estimations of heritability and genetic covariance based on collected GWAS data also provide insights into the genetic basis of complex traits/diseases (Tenesa and Haley (2013); van Rheenen et al. (2019)). Heritability is the proportion of phenotypic variance due to genetic effects, and genetic covariance is the covariance of genetic effects contributing to two phenotypes. Methods based on the linear mixed model (LMM) and the restricted maximum likelihood (REML) algorithm have been developed to estimate these two quantities of significant genetic interest (Yang et al. (2010); Lee et al. (2012)). Compared with traditional family-based approaches for estimating these two quantities, these methods do not need to collect related samples and can use large GWAS samples for estimation. Moreover, they do not require the studied phenotypes to be measured on the same individuals when estimating the genetic covariance, which makes it possible to study a spectrum of human complex traits/diseases simultaneously by using different cohorts. With regard to the statistical properties of these estimates, the high-dimensional

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0060

asymptotic theory of the REML for heritability estimation has been recently established, even under a misspecified LMM, which provides theoretical support for the robustness of the REML estimator (Jiang et al. (2016)).

However, LMM-based methods require individual-level genotype and phenotype data, which are usually difficult to obtain, owing to policy and privacy concerns. Increasingly accessible marginal association statistics from GWAS and advances in analytical methods that rely only on these summary statistics have circumvented challenges in data sharing and greatly accelerated research in complex trait/disease genetics. Owing to its computational efficiency, the linkage disequilibrium (LD) score regression (LDSC; Bulik-Sullivan et al. (2015a,b)) is currently the most popular method for estimating heritability and genetic covariance using GWAS summary statistics. Based on this method, bioinformatics servers have been built to improve the computation and visualization of the heritability and genetic covariance of a wide range of phenotypes (Zheng et al. (2017)).

In a typical GWAS data set, the total number of SNPs, p (e.g., $10^6 \sim 10^7$), is often much larger than the sample size, n (e.g., $10^3 \sim 10^6$), that is, $p \gg n$. In addition, more SNPs can be observed when more subjects are recruited in GWAS, that is, p increases with n. In other words, GWAS

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0060

data analyses are high dimensional. Despite the polygenicity of many phenotypes, such as anthropometric characteristics (Berndt et al. (2013)) and psychiatric disorders (Sullivan, Daly and O'donovan (2012)), the SNPs that have biological effects on the phenotypes (causal SNPs) are still only a small portion of all the SNPs. However, heritability and genetic covariance estimation methods based on summary statistics, such as the LDSC, often assume that the effects of all SNPs are nonzero, while the true underlying model might be sparse; that is, the assumed model is misspecified in the LDSC. Although the LDSC has become a routine past of post-GWAS analyses for estimating the heritability and genetic covariance, the high-dimensional asymptotic behavior of the LDSC under a model misspecification has not yet been rigorously justified. Therefore, there is a pressing need for a theoretical justification for the LDSC. In this paper, we establish the consistency and asymptotic normality of the heritability and genetic covariance estimators of the LDSC in a regime of high-dimensional statistics, as both the sample size n and the dimension of the random effects p tend to infinity. Our results indicate that the misspecified LDSC estimators converge to the desired true values of the genetic quantities. We also provide their convergence rates (in probability) and asymptotic variances. Our theoretical results are fully supported by our empirical studies.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0060

1.1 LDSC estimation under model misspecification

We first explain how to estimate heritability using an LDSC and GWAS summary statistics (Bulik-Sullivan et al. (2015a)). Based on the LMM, phenotypes are modeled as

$$\phi = X\beta + \epsilon, \tag{1.1}$$

where ϕ is an $n \times 1$ vector of (quantitative) phenotypes, X is an $n \times p$ random design matrix of genotypes normalized to mean zero and variance one, β is a $p \times 1$ vector of random effects following a $N[0, (h^2/p)I_p]$ distribution, in which I_p denotes the p-dimensional identity matrix, and ϵ is an $n \times 1$ vector of errors that is distributed as $N[0, (1-h^2)I_n]$. Here, X, β, and ϵ are mutually independent. We further assume that the genotypes of different subjects are independent of each other. Before the normalization, the genotypes are coded as 0, 1, and 2, which are the allelic dosages (number of minor alleles) of the variants. Denote f_j as the known minor allele frequency (MAF) of SNP j. According to the Hardy Weinberg equilibrium (HWE), the probabilities of the genotype being 0, 1, and 2 for SNP j are $(1-f_j)^2$, $2f_j(1-f_j)$, and f_j^2, respectively. Thus, after the normalization, we have $-2f_j/\sqrt{2f_j(1-f_j)}$, $(1-2f_j)/\sqrt{2f_j(1-f_j)}$, and $(2-2f_j)/\sqrt{2f_j(1-f_j)}$,

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0060

respectively, in X. It follows that $\mathrm{E}(XX') = pI_n$. Hence,

$$\mathrm{Var}(\phi) = \mathrm{Var}(X\beta) + \mathrm{Var}(\epsilon) = \frac{h^2}{p}\mathrm{E}(XX') + (1-h^2)I_n = I_n. \qquad (1.2)$$

Heritability is defined as the proportion of phenotypic variance attributed to genetic factors. Based on this definition, the heritability of a phenotype is the sum of the random effect variances, which is h^2.

Owing to the existence of LD, genotypes of different SNPs are correlated, especially for SNPs located nearby (Stephens et al. (2001)). We denote r_{jk} as the genotypic correlation between SNP j and SNP k, that is, $r_{jk} = \mathrm{E}(X_{ij}X_{ik})$, which does not depend on i. The pairwise correlations between SNPs are stored in an LD matrix R, that is, for any subject i, for $1 \leq i \leq n$, $\mathrm{cov}(X_{[i]}) = R$, where $X_{[i]}$ is the ith row of X. The correlations usually decay with an increase in the pairwise distances, and hence the LD matrix is C-dependent, overall (discussed in detail). The LD score of an SNP is defined as $l_j = \sum_{k=1}^{p} r_{jk}^2$, where the sum is taken over all the variants, including SNP j itself ($r_{jj} = 1$). As a special case, when SNP j is independent of the other SNPs, we have $l_j = 1$. In practice, the LD matrix and LD scores can be obtained from a public external reference panel constituting individual-level genotype data (e.g., the 1000 Genomes Project Clarke et al. (2017)). Following the arguments in Bulik-Sullivan et al. (2015a), we replace r_{jk}^2 in the definition of l_j with an approximately

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0060

unbiased estimator given by $r^2_{jk,adj} = \hat{r}^2_{jk} - \left(1 - \hat{r}^2_{jk}\right) / (N-2)$, where N is the sample size of the reference panel and $\hat{r}^2_{jk}$ denotes the square of the sample Pearson correlation coefficient.

The design matrix X may be difficult to access, owing to privacy and security issues. The advantage of the LDSC is that it needs only more accessible GWAS summary statistics as input. In GWAS summary statistics, we have z-score for each SNP that reflects the marginal association between the phenotype and the SNP. Because the marginal heritability explained by one SNP is usually small, the z-score of SNP j, z_j, can be approximated by $z_j = X_j'\phi/\sqrt{n}$, where X_j denotes the jth column of X. In an LDSC (Bulik-Sullivan et al. (2015b)), the heritability can be estimated by solving the following linear regression:

$$\mathrm{E}(z_j^2) \;=\; 1 + h^2(n/p)l_j, \quad j = 1, \ldots, p, \tag{1.3}$$

where the intercept is fixed as one.

The model has been generalized to estimate the genetic covariance between phenotypes (Bulik-Sullivan et al. (2015a)). Genetic covariance analysis can provide new insights into the shared genetics of many phenotypes, with numerous downstream applications (van Rheenen et al. (2019), Zhang et al. (2021)), and so has become a popular post-GWAS analysis tool. Let us assume that there are two GWAS for two different phenotypes

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0060

with sample sizes n_1 and n_2, respectively. The two GWAS share the same set of p SNPs, but are not necessarily performed within the same cohort. In practice, two different GWAS may share a subset of subjects. Denote the number of shared subjects as n_{o} (the subscript o refers to "overlap"), $0 \leq n_{\mathrm{o}} \leq n_1 \wedge n_2 \equiv \min(n_1, n_2)$. The phenotypes are modeled as

$$\begin{aligned} \phi_1 &= X\beta + \epsilon \\ \phi_2 &= Y\gamma + \delta, \end{aligned} \tag{1.4}$$

where ϕ_1 and ϕ_2 are $n_1 \times 1$ and $n_2 \times 1$ vectors, respectively, of phenotypes, X and Y are $n_1 \times p$ and $n_2 \times p$ random design matrices, respectively, of genotypes normalized to have mean zero and variance one with the same LD matrix, β and γ are two $p \times 1$ vectors of random effects jointly normally distributed so that $\mathrm{E}(\beta) = \mathrm{E}(\gamma) = 0$ and

$$\mathrm{Var}\begin{pmatrix} \beta \\ \gamma \end{pmatrix} = \frac{1}{p}\begin{pmatrix} h_1^2 I_p & \rho_g I_p \\ \rho_g I_p & h_2^2 I_p \end{pmatrix},$$

and ϵ and δ are $n_1 \times 1$ and $n_2 \times 1$ vectors, respectively, of random errors. The marginal distributions of ϵ and δ are $N[0, (1-h_1^2)I_{n_1}]$ and $N[0, (1-h_2^2)I_{n_2}]$, respectively. Here, (X, Y), (β, γ), and (ϵ, δ) are independent. Without loss of generality, we assume that the first n_{o} samples in each study are shared. In addition, ϵ and δ are correlated because of the non-genetic correlation

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0060

introduced through the overlapping samples:

$$\mathrm{cov}(\epsilon_i, \delta_j) = \begin{cases} \rho_{\mathrm{e}}, & 1 \le i = j \le n_{\mathrm{o}} \\ 0, & \text{otherwise} \end{cases}.$$

Similarly to (1.3), to estimate the genetic covariance, ρ_{g}, one can fit the following linear regression model:

$$\mathrm{E}(z_{1j} z_{2j}) \quad = \quad \rho n_{\mathrm{o}} / \sqrt{n_1 n_2} + \rho_{\mathrm{g}} (\sqrt{n_1 n_2}/p) l_j, \quad j = 1, \ldots, p, \qquad (1.5)$$

where $\rho = \rho_{\mathrm{g}} + \rho_{\mathrm{e}}$. As a special case, if study 1 and study 2 are the same study, which means that we have $n_1 = n_2 = n_{\mathrm{o}}$, $\rho_{\mathrm{g}} = h_1^2 = h_2^2 = h^2$, and $\rho_{\mathrm{e}} = 1 - h^2$, then model (1.5) reduces to model (1.3).

A basic assumption in the above LMMs is that all SNPs contribute to the phenotypic variance. In reality, however, only a subset of the SNPs are causal SNPs. Let $S, T_1, T_2 \subset \{1, 2, \ldots, p\}$ represent the indices of causal SNPs shared in both traits, those presented only in trait 1, and those presented only in trait 2, respectively. In other words, $S \cup T_k$ are the indices of the causal SNPs for trait k (where $k = 1, 2$). Note that S, T_1, and T_2 are mutually exclusive subsets. Let β_S and γ_S be the vectors of random effects corresponding to the SNPs in S for both phenotypes. Similarly, β_{T_k} and γ_{T_k} are defined as the random effect vectors corresponding to the SNPs in T_k, for $k = 1, 2$. Let $m = |S|$ (cardinality), $m_1 = |T_1|$, and $m_2 = |T_2|$. Under

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0060

the true model, the distribution of β_j is $N[0, h_1^2/(m+m_1)]$ for $j \in S \cup T_1$, and $\beta_j = 0$ for $j \notin S \cup T_1$. Similarly, we have $\gamma_j \sim N[0, h_2^2/(m+m_2)]$ when $j \in S \cup T_2$, and $\gamma_j = 0$ when $j \notin S \cup T_2$. The true LMMs can then be expressed as

$$\begin{aligned} \phi_1 &= X_S\beta_S + X_{T_1}\beta_{T_1} + \epsilon \\ \phi_2 &= Y_S\gamma_S + Y_{T_2}\gamma_{T_2} + \delta \end{aligned} \tag{1.6}$$

[compare with (1.4)], where X_A is a normalized genotype matrix for the SNPs in set A (where $A = S$ or T_1), and Y_A is defined similarly ($A = S$ or T_2). The joint distribution for the effects of the SNPs in S is given by

$$\begin{pmatrix} \beta_S \\ \gamma_S \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \frac{h_1^2}{m+m_1} I_m & \frac{\rho_g}{m} I_m \\ \frac{\rho_g}{m} I_m & \frac{h_2^2}{m+m_2} I_m \end{pmatrix} \right].$$

Here, h_1^2, h_2^2, and ρ_g are the heritability of phenotype 1, heritability of phenotype 2, and genetic covariance between phenotype 1 and 2, respectively, under the true model. Detailed assumptions about the distributions of the genotype matrices and the random effects under the true model are given in the following.

In practice, it is impossible to determine whether an SNP is causal for a phenotype. Therefore, we have to follow the assumption of the LDSC that all SNPs are causal in order to estimate the heritability and genetic

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0060

covariance, which actually leads to the misspecified model. The main goal of this study is to show that the consistency and asymptotic normality properties of misspecified LDSC estimators are valid when n, p, and m tend to infinity.

We conclude this section with a couple of numerical illustrations.

1.2 Numerical illustrations

In GWAS, SNPs are high-density bi-allelic genetic markers. Each SNP can be considered as a binomial random variable with two trials, and the probability of "success" is the minor allele frequency f_j. In each sample, the SNP genotypes are correlated, which is known as LD. To mimic the LD matrix of the human genome, the LD matrix we use to simulate genotype data has a block structure, which is a special case of the C-dependent relationship formally introduced later. In this simulation, because the LD matrix is known, we directly calculate the LD scores based on the true LD matrix. Please note that we usually rely on an external reference panel to estimate LD scores in practice, owing to the unavailability of the true LD matrix. Later, we discuss the effect of randomness in the estimation of the LD scores. According to the genome partition software LDetect (Berisa and Pickrell (2016)), there are ∼2,000 independent blocks for ∼5,000,000

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0060

SNPs in the human genome of European ancestry. We randomly selected 200 blocks (three blocks were later removed because of their small size), and scaled down the number of SNPs in each block to make the total number of SNPs 20,000. In the following illustrative examples, the number of SNPs in each block ranges from 1 to 502, with a mean of 102. The SNP genotypes in different blocks are independent. We further assume that the local LD matrix for each block follows an AR(1) structure, that is, if SNP j and SNP $(j+d)$ are in the same block, the genotypic correlation between these two SNPs is α^d. We use the AR(1) correlation structure to mimic the observation that LD decays with distance in a real genome. The AR(1) model coefficient α for each block is independently sampled from $\{0.1, \ldots, 0.9\}$ with equal probability. After generating the LD matrix, we fix it in the remainder of the experiment. In our simulations, the SNPs in the same block share the same MAF, which is sampled from the Uniform$(0.05, 0.5)$ distribution. CorBin is a highly efficient R-package for generating high-dimensional binary/binomial data with a specified correlation structure, including exchangeable, AR(1), and K-dependent structures (Jiang et al. (2020)). We use CorBin to generate correlated genotype data for each individual.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0060

Heritability. In this illustrative simulation for heritability estimation, we fix $p = 20,000$ and the true value of heritability $h^2 = 0.6$. Let $b_j = 1$ if SNP j is causal, the corresponding effect size of which follows $N(0, h^2/m)$; otherwise, $b_j = 0$ and the effect of SNP j is zero. The indicators $b_1, \ldots, b_p$ are independent Bernoulli random variables such that $\mathrm{P}(b_j = 1) = \omega \in (0, 1)$. Note that $m = \sum_{j=1}^{p} b_j$. We use τ to represent the ratio of the sample size to the SNP number (i.e., $\tau = n/p$). We examine the behavior of the LDSC heritability estimator for different ω and τ (Figure 1). In the first scenario, we fix $\tau = 0.1$ and vary ω from 0.005 to 1. In the second scenario, we fix $\omega = 0.05$ and vary τ from 0.05 to 0.5. To avoid having no causal SNPs being generated when the expected causal SNP proportion ω is small, we set $(\omega/2)p$ as a lower bound for m. The genotype data, SNP effect sizes, and error terms are generated independently. We use an LDSC in which all SNPs are implicitly assumed to be causal, to estimate the heritability of the phenotype. The process is repeated 100 times for each setting of ω and τ. As shown in Figure 1, there is almost no bias in the estimated h^2, regardless of the sample size or the underlying true model. This suggests that the LDSC works well in terms of providing an unbiased estimator of heritability, even in the case of a model misspecification.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0060

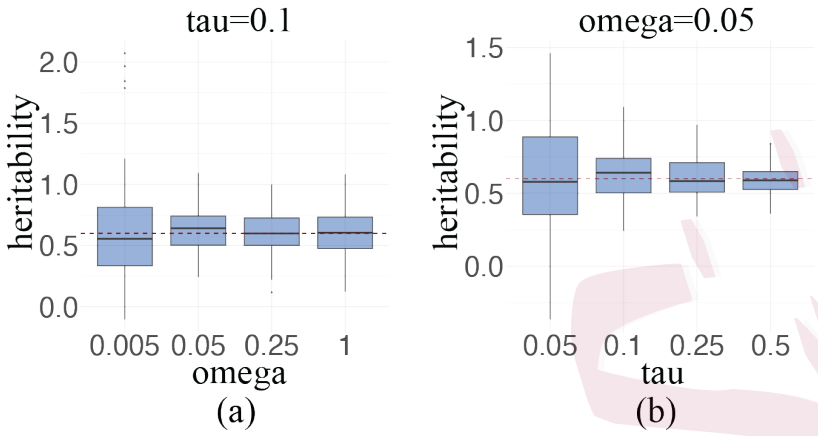

Figure 1: Heritability Estimation: C-dependent SNPs for $p = 20,000$; (a) $\tau = 0.1$ ($n = 2,000$), and different ω; (b) $\omega = 0.05$, and different τ.

Genetic covariance. We also conducted simulations for the genetic covariance estimation (Figure 2). Here, we set $p = 20,000$, $\rho_{\mathrm{g}} = 0.15$, and $\rho_{\mathrm{e}} = 0.1$. We assume that study 1 and study 2 are performed on the same cohort. Thus, $n_1 = n_2 = n_{\mathrm{o}}$ and $X = Y$. We define $\tau = n_1/p = n_2/p$. We use the Bernoulli random variables b_{1j} and b_{2j} to indicate whether SNP j is a causal SNP for phenotypes 1 and 2, respectively, such that $\mathrm{P}(b_{1j} = 1) = \omega_1$, $\mathrm{P}(b_{2j} = 1) = \omega_2$, and $\mathrm{P}(b_{1j}b_{2j} = 1) = \omega$. We have $m = \sum_{j=1}^{p} b_{1j}b_{2j}$, $m + m_1 = \sum_{j=1}^{p} b_{1j}$, and $m + m_2 = \sum_{j=1}^{p} b_{2j}$. As an illustration, we assume the causal SNPs for the two phenotypes are the same set, that is, $\omega_1 = \omega_2 = \omega$ and $m_1 = m_2 = 0$. However, the consistency from this illustrative experiment is also evident when ω_1, ω_2, and ω are

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0060

not equal. Here, we use $(\omega/2)p$ as a lower bound for m to avoid no causal SNPs being generated. After independently generating the genotype data, SNP effect sizes, and error terms, we use the LDSC to estimate the genetic covariance of the phenotypes. All SNPs are misspecified as causal during the LDSC estimation. The process is repeated 100 times for each setting of ω and τ. We first fix $\tau = 0.1$ and vary ω from 0.005 to 1. We then fix $\omega = 0.05$ and vary τ from 0.05 to 0.5. Figure 2 shows that the LDSC estimator for genetic covariance remains unbiased under the misspecified models and different scenarios and sample sizes.

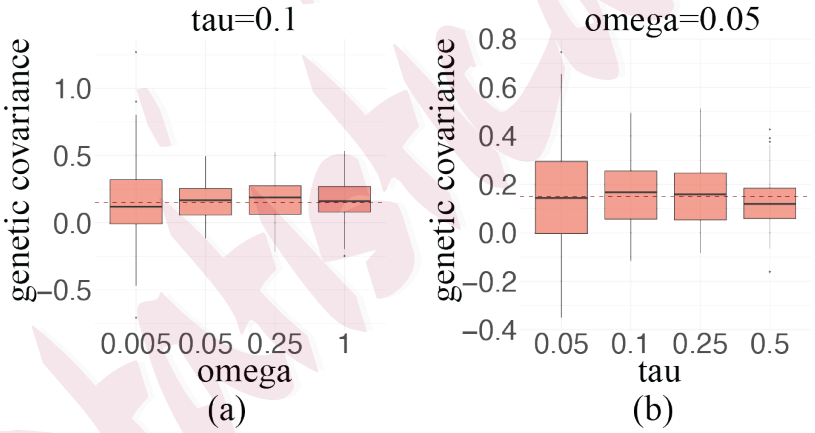

Figure 2: Genetic Covariance Estimation: C-dependent SNPs for $p = 20,000$; (a) $\tau = 0.1$ $(n = 2,000)$, $\omega_1 = \omega_2 = 0$, and different ω; (b) $\omega_1 = \omega_2 = 0$, $\omega = 0.05$, and different τ.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0060

2. Asymptotic theory

As noted in the previous section, there are two main quantities of genetic interest, namely, heritability and genetic covariance. We first study the asymptotic behavior of the LDSC heritability estimator under a suitable framework. Later, we extend the framework to study the asymptotic behavior of the LDSC genetic covariance estimator. We begin with some preparation.

2.1 Definition, key lemmas, and corollary

For any two subsets of indexes $\mathcal{B}_r \subset \{1, \ldots, p\}$, for $r = 1, 2$, the distance between $\mathcal{B}_1$ and $\mathcal{B}_2$ is defined as

$$d(\mathcal{B}_1, \mathcal{B}_2) = \min_{j_1 \in \mathcal{B}_1, j_2 \in \mathcal{B}_2} |j_1 - j_2|.$$

Definition 2.1. The columns of X, denoted by $X_1, \ldots, X_p$, 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 $d(J_r, J_s) > C$, for $1 \leq r \neq s \leq t$, $[X_j]_{j \in J_1}, \ldots, [X_j]_{j \in J_t}$ are independent.

A standard example of C-dependency is the moving average process in time series (e.g., Shumway and Stoffer (2017)), and a special case is that of independent SNPs, which corresponds to $C = 0$; in other words, $X_1, \ldots, X_p$

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0060

are independent. In practice, even without such a cut-off C, if the correlation decays reasonably fast as the distance between the SNPs increases, after a certain distance, the correlation may be treated as approximately zero. Therefore, the C-dependent notion is not unreasonable from a practical standpoint.We show that as long as C is $O(1)$, the asymptotic results do not depend on the actual value of C.

The technical lemmas and their corollaries are given in the Supplementary Material.

2.2 Heritability

We assume that the locations of the causal SNPs are characterized by a set of independent Bernoulli random variables, $b_1, \ldots, b_p$, such that $\mathrm{P}(b_j = 1) = \omega \in (0, 1]$. Let $S = \{1 \leq j \leq p : b_j = 1\}$, $m = |S|$ (cardinality), $X_S = [X_j]_{j \in S}$, and $\beta_S = (\beta_j)_{j \in S}$. Note that there is a nonzero probability that $m = 0$, in which case, some of the quantities introduced below involving m in the denominators are not well defined. However, we can (slightly) modify the definition of m as $m^* = m \vee \{(\omega/2)p\}$, without affecting the consistency or asymptotic normality we examine here. For example, let ζ_N denote a random variable involving m, and let $\tilde{\zeta}_N$ be the same quantity,

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0060

but with m replaced by m^*. Then, we have, for any Borel set B,

$$\begin{aligned}
\mathrm{P}(\zeta_N \in B) &= \mathrm{P}[\zeta_N \in B, m \geq (\omega/2)p] + \mathrm{P}[\zeta_N \in B, m < (\omega/2)p] \\
&= \mathrm{P}[\tilde{\zeta}_N \in B, m \geq (\omega/2)p] + \mathrm{P}[\zeta_N \in B, m < (\omega/2)p] \\
&= \mathrm{P}(\tilde{\zeta}_N \in B) - \mathrm{P}[\tilde{\zeta}_N \in B, m < (\omega/2)p] \\
&\quad + \mathrm{P}[\zeta_N \in B, m < (\omega/2)p] \\
&= \mathrm{P}(\tilde{\zeta}_N \in B) + o(1).
\end{aligned}$$

It can be shown that $\mathrm{P}(m^* \neq m) = o(n^{-K})$, for any positive integer K. Therefore, without loss of of generality, we can replace m by m^*; however, for notational simplicity, we still denote it by m. We assume that the following hold:

(i) X and $b = (b_j)_{1 \leq j \leq p}$ are independent;

(ii) (X, b, β) is independent of ϵ;

(iii) $\beta_S | X, b \sim N\,[0, (h^2/m) I_m]$, and

(iv) $\epsilon \sim N[0, (1 - h^2) I_n]$.

The true underlying model can be expressed as

$$\phi = X_S \beta_S + \epsilon = \sum_{j \in S} \beta_j X_j + \epsilon. \tag{2.1}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0060

Let $W = (X, b)$. For any n-dimensional constant λ, we have

$$\begin{aligned} \mathrm{E}(e^{\lambda'\phi}|W) &= \mathrm{E}\{e^{\lambda' X_S \beta_S}\mathrm{E}(e^{\lambda'\epsilon}|W,\beta)|W\} \\ &= e^{(1-h^2)\lambda'\lambda/2}\mathrm{E}(e^{\lambda' X_S\beta_S}|W) \\ &= e^{\lambda'\Sigma\lambda/2}, \end{aligned}$$

where $\Sigma = (1-h^2)I_n + (h^2/m)X_S X_S'$. It follows that $\phi|W \sim N(0,\Sigma)$.

It can be shown that the heritability estimator of LDSC, $\hat{h}^2$, can be expressed as

$$\hat{h}^2 = \frac{\sum_{j=1}^p u_j(z_j^2 - 1)}{\sum_{j=1}^p u_j^2}, \tag{2.2}$$

where $u_j = (n/p)l_j$. Furthermore, we have $z_j^2 = \phi'(n^{-1}X_j X_j')\phi$. Thus, we have

$$\hat{h}^2 = \phi' A\phi - \frac{u_\cdot}{u_\cdot^2}, \tag{2.3}$$

where $u_\cdot = \sum_{j=1}^p u_j$, $u_\cdot^2 = \sum_{j=1}^p u_j^2$, and $A = (nu_\cdot^2)^{-1}\sum_{j=1}^p u_j X_j X_j'$.

We establish the consistency of the heritability estimator in the following theorem. The proof is given in the Supplementary Material.

Theorem 1. Suppose that $X_1, \cdots, X_p$ are C-dependent, and $n/p \longrightarrow \tau \in (0,1]$. Then, we have $\hat{h}^2 = h^2 + o_{\mathrm{P}}(1)$.

We now consider the asymptotic normality of the heritability estimator. First, define

$$r_{j_1j_2j_3} = \mathrm{E}(X_{1j_1}X_{1j_2}^2X_{1j_3}), \quad r_{1,j_1j_2j_3j_4} = \mathrm{E}(X_{1j_1}X_{1j_2}X_{1j_3}^2X_{1j_4}^2),$$

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0060

$$r_{2,j_1j_2j_3} = \mathrm{E}(X_{1j_1}^2 X_{1j_2}^2 X_{1j_3}^2), \quad r_{2,j_1j_2j_3j_4} = \mathrm{E}(X_{1j_1}^2 X_{1j_2}^2 X_{1j_3}^2 X_{1j_4}^2),$$

$v_t = \sum_{j=1}^p u_j r_{tj}^2, 1 \le t \le p$, $v_\cdot^2 = \sum_{t=1}^p v_t^2$, and $(uv)_\cdot = \sum_{t=1}^p u_t v_t$.

Theorem 2. Suppose that, in addition to the conditions of Theorem 1, the following limits exist: $\tau_2 = \lim(u_\cdot^2/p) > 0, \lambda = \lim(u_\cdot/u_\cdot^2), \phi = \lim\{(uv)_\cdot/u_\cdot^2\}$,

$\gamma = \lim\{(u_\cdot^2)^2/nv_\cdot^2\} > 0$,

$\gamma_1 = \lim p^{-1} \sum_{j_1,j_2,j_3,j_4=1}^p u_{j_1}u_{j_3}r_{j_1j_2}r_{j_2j_3}r_{j_3j_4}r_{j_4j_1}$,

$\gamma_2 = \lim p^{-2} \sum_{j_1,j_2,j_3,j_4=1}^p u_{j_1}u_{j_3}r_{j_1j_2j_3}r_{j_3j_4}r_{j_4j_1}$,

$\gamma_3 = \lim p^{-3} \sum_{j_1,j_2,j_3,j_4=1}^p u_{j_1}u_{j_3}r_{j_1j_2}r_{1,j_1j_2j_3j_4}$,

$\gamma_4 = \lim p^{-3} \sum_{j_1,j_2,j_3,j_4=1}^p u_{j_1}u_{j_3}r_{j_1j_2j_3}r_{j_3j_4j_1}$,

$\gamma_5 = \lim p^{-4} \sum_{j_1,j_2,j_3,j_4=1}^p u_{j_1}u_{j_3}r_{2,j_1j_2j_3j_4}$,

$\gamma_6 = \lim p^{-1} \sum_{j_1,j_2,j_3=1}^p u_{j_1}u_{j_2}r_{j_1j_2}r_{j_2j_3}r_{j_3j_1}$,

$\gamma_7 = \lim p^{-2} \sum_{j_1,j_2,j_3=1}^p u_{j_1}u_{j_2}(r_{j_1j_2}r_{j_2j_3j_1} + r_{j_2j_3}r_{j_3j_1j_2} + r_{j_3j_1}r_{j_1j_2j_3})$,

and $\gamma_8 = \lim p^{-3} \sum_{j_1,j_2,j_3=1}^p u_{j_1}u_{j_3}r_{2,j_1j_2j_3}$, as $p \to \infty$. Then, we have

$$\sqrt{n}(\hat{h}^2 - h^2) \quad \xrightarrow{\mathrm{d}} \quad N(0, \sigma^2), \tag{2.4}$$

as $p \to \infty$, where the asymptotic variance has the following expression:

$$\begin{aligned}\sigma^2 &= 3h^4\tau\left(\frac{1-\omega}{\omega}\right)\left(\frac{\tau}{\gamma}-1\right)\\ &+2\left[h^4\left\{\frac{\gamma_1\tau^3+6\gamma_2\tau^2+(4\gamma_3+2\gamma_4)\tau+\gamma_5}{\tau_2^2}+\left(\frac{1-\omega}{\omega}\right)\tau(\lambda+1)^2\right\}\right.\\ &\left.+2h^2(1-h^2)\frac{\gamma_6\tau^2+\gamma_7\tau+\gamma_8}{\tau_2^2}+(1-h^2)^2\left(\frac{\phi\tau}{\tau_2}+\lambda^2\right)\right].\end{aligned} \tag{2.5}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0060

An inspection of the limits defined in Theorem 2, in terms of the order of the sum involved in the limit, suggests that they can all be reasonably expected. The proof of Theorem 2 is given in the Supplementary Material.

A special case is when the SNPs are independent, that is, $C = 0$. In this case, it can be verified that $\tau_2 = \tau^2$, $\lambda = \tau^{-1}$, $\gamma = \tau$, $\phi = 1$, and $\gamma_s = \tau^2$, if $1 \leq s \leq 8$ and $s \neq 7$, and $\gamma_7 = 3\tau^2$, where τ is given by Theorem 2. Thus, we have the following result.

Corollary 2. In the case of independent SNPs, (2.4) holds under $n/p \longrightarrow \tau \in (0, 1]$, where

$$\sigma^2 \equiv 2\left[\frac{1}{\tau} + h^4\left\{\frac{\tau}{\omega}\left(1+\frac{1}{\tau}\right)^2 + \left(2+\frac{1}{\tau}\right)^2\right\} + 2h^2(1-h^2)\left(1+\frac{1}{\tau}\right)^2 + \frac{(1-h^2)^2}{\tau^2}\right]. \tag{2.6}$$

Note that because $\omega, \tau \in (0, 1]$, we have $\sigma^2 = O(1)$, and hence the convergence rate of $\hat{h}^2$ is $n^{-1/2}$. Because m and n are of the same order, the convergence rate can be expressed in terms of either n or m. In fact, because the asymptotic depends on both ω and τ, the asymptotic variance depends on the (limit) ratio of m/n, which makes sense.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0060

2.3 Genetic covariance

Let b_{rj}, for $1 \le j \le p, r = 1, 2$, be such that

(I) (b_{1j}, b_{2j}), for $j = 1, \ldots, p$, are independent;

(II) $b_{rj} \sim \text{Bernoulli}(\omega_r)$, for $j = 1, 2$, where $\omega_1, \omega_2 \in (0, 1]$;

(III) $\omega = \text{P}(b_{1j}b_{2j} = 1) > 0$.

Note that the definition allows a correlation between b_{1j} and b_{2j} for the same j. Note that $\omega = \omega_1\omega_2$ if b_{1j}, and b_{2j} are independent. Denote $b_r = (b_{rj})_{1\le j\le p}$, $r = 1, 2$, and $b = (b_1, b_2)$. Then, we have $S = \{1 \le j \le p : b_{1j}b_{2j} = 1\}$, and $m = |S|$, and $S \cup T_r = \{1 \le j \le p : b_{rj} = 1\}$, and $m_r = |S \cup T_r| - |S| = |S \cup T_r| - m$. Thus, $|S \cup T_r| = m + m_r$, for $r = 1, 2$. For any subset of indices $J \subset \{1, \ldots, p\}$, let $X_J = [X_j]_{j\in J}$ and Y_J be defined similarly. Let $\beta = (\beta_S', \beta_{T_1}', \beta_{T_2}', \beta_U')'$, and $\gamma = (\gamma_S', \gamma_{T_1}', \gamma_{T_2}', \gamma_U')'$. We assume that the following conditions hold:

(a) (X, Y) and b are independent;

(b) (X, Y, β, γ, b) is independent of (ϵ, δ);

(c) $(\beta, \gamma)|b \sim N(0, \Omega)$, where Ω is the covariance matrix described in Section 1;

(d) $(\epsilon, \delta) \sim$ the distribution specified in Section 1.

It is more convenient to define $\beta_j = 0$, for $j \notin S \cup T_1$, and $\gamma_j = 0$, for $j \notin S \cup T_2$. Let $\xi_j = (\xi_{1j}, \xi_{2j})' = (\sqrt{m + m_1}b_{1j}\beta_j, \sqrt{m + m_2}b_{2j}\gamma_j)'$, for

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0060

$1 \leq j \leq p$. Then, given b, ξ_j, for $1 \leq j \leq p$ are independent vectors with $\xi_j \sim N(0, \Sigma_b)$, where

$$\Sigma_b = \begin{bmatrix} h_1^2 & \rho_g \sqrt{(1+m_1/m)(1+m_2/m)} \\ \rho_g \sqrt{(1+m_1/m)(1+m_2/m)} & h_2^2 \end{bmatrix},$$

$j \in S$; $\xi_j = (\xi_{1j}, 0)'$ with $\xi_{1j} \sim N(0, h_1^2)$, $j \in T_1$; $\xi_j = (0, \xi_{2j})'$ with $\xi_{2j} \sim N(0, h_2^2)$, $j \in T_2$; and $\xi_j = 0$, $j \notin S \cup T_1 \cup T_2$. Let $\xi_b = (\xi_j)_{1 \leq j \leq p}$, the column vector that combines all ξ_j. Note that for β_S, and γ_S to have the joint distribution specified in Section 1, it is necessary that the following holds:

$$\frac{\rho_g}{m} = |\mathrm{cov}(\beta_j, \gamma_j)| \leq \sqrt{\mathrm{var}(\beta_j)\mathrm{var}(\gamma_j)} = \frac{h_1 h_2}{\sqrt{(m+m_1)(m+m_2)}},$$

$j \in S$, if $S \neq \emptyset$. It follows that the following inequality must be satisfied:

$$\left(1 + \frac{m_1}{m}\right)\left(1 + \frac{m_2}{m}\right) \leq \left(\frac{h_1 h_2}{\rho_g}\right)^2. \tag{2.7}$$

Therefore, we modify the definition of the covariance matrix of β and γ so that, when (2.7) does not hold, the covariance matrix of β_S and γ_S is 0 (matrix). As a result, the covariance matrix of ξ_j is $\mathrm{diag}(h_1^2, h_2^2)$, for $1 \leq j \leq p$, when (2.7) does not hold. It can be seen that (2.7) holds with probability tending to one, provided that

$$\frac{\omega_1 \omega_2}{\omega^2} < \left(\frac{h_1 h_2}{\rho_g}\right)^2. \tag{2.8}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0060

Therefore, asymptotically, (β_S, γ_S) still has the distribution described in Section 1.

The LDSC estimator of the genetic covariance is defined differently under independent SNPs than it is under correlated SNPs. We consider these cases separately.

1. Independent SNPs. In this case, we assume $n_{\mathrm{o}} = 0$ in order to ensure identifiability. Then, the LDSC estimator of ρ_{g} is simplified to

$$\hat{\rho}_{\mathrm{g}} = \frac{1}{n_1 n_2} \sum_{j=1}^{p} \phi_1' X_j Y_j' \phi_2 = \phi' A \phi, \tag{2.9}$$

where $\phi = (\phi_1', \phi_2')'$ and $A = (2n_1 n_2)^{-1} \sum_{j=1}^{p} \Psi_j$, with

$$\Psi_j = \begin{pmatrix} 0 & X_j Y_j' \\ Y_j X_j' & 0 \end{pmatrix}. \tag{2.10}$$

The following result is proved in the Supplementary Material.

Theorem 3. Suppose that the SNPs are independent, (2.8) holds with $\omega > 0$, and

$$\frac{n_r}{p} \to \tau_r \in (0, 1], \ r = 1, 2. \tag{2.11}$$

Then, we have $\hat{\rho}_{\mathrm{g}} = \rho_{\mathrm{g}} + o_{\mathrm{P}}(1)$.

The next result relates to the asymptotic distribution of $\hat{\rho}_{\mathrm{g}}$. The proof is given in the Supplementary Material.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0060

Theorem 4. Under the conditions of Theorem 1, we have $\sqrt{n_\cdot}(\hat{\rho}_{\mathrm{g}} - \rho_{\mathrm{g}}) \xrightarrow{\mathrm{d}} N(0, \sigma^2)$, where $n_\cdot = n_1 + n_2$ and

$$\begin{aligned}\sigma^2 &= (\tau_1+\tau_2)\left\{h_1^2h_2^2\left(\frac{\omega}{\omega_1\omega_2}+\frac{\tau_1+\tau_2+1}{\tau_1\tau_2}\right)+\rho_{\mathrm{g}}^2\left(\frac{1}{\omega}+\frac{1}{\tau_1}+\frac{1}{\tau_2}\right)\right\}\\ &\quad+\left(\frac{1}{\tau_1}+\frac{1}{\tau_2}\right)\left\{h_1^2(1-h_2^2)\tau_1+(1-h_1^2)h_2^2\tau_2+1-h_1^2h_2^2\right\}. \quad (2.12)\end{aligned}$$

2. C-dependent SNPs. In this case, the genetic covariance, ρ_{g}, is estimated by fitting the following linear regression in the LDSC:

$$z_j = \beta_0 + \beta_1 u_j + e_j, \quad j = 1, \ldots, p, \quad (2.13)$$

where $z_j = z_{1j}z_{2j}$, $u_j = (\sqrt{n_1n_2}/p)l_j$, and $\beta_1 = \rho_{\mathrm{g}}$. The LDSC estimators, which are also the least squares (LS) estimators of the regression coefficients, are given by

$$\hat{\rho}_{\mathrm{g}} = \hat{\beta}_1 = \frac{\sum_{j=1}^p (u_j-\bar{u})(z_j-\bar{z})}{\sum_{j=1}^p (u_j-\bar{u})^2} = \phi' A \phi, \quad (2.14)$$

$$\hat{\beta}_0 = \bar{z} - \hat{\beta}_1\bar{u}, \quad (2.15)$$

where $A = (2\sqrt{n_1n_2}d_p)^{-1}\sum_{j=1}^p(u_j-\bar{u})\Psi_j$, $d_p = \sum_{j=1}^p(u_j-\bar{u})^2$, Ψ_j is given by (2.10), $\bar{u} = p^{-1}\sum_{k=1}^p u_k$, and $\bar{z} = p^{-1}\sum_{k=1}^p z_k$. Because our main interest lies in estimating ρ_{g}, we focus on $\hat{\beta}_1 = \hat{\rho}_{\mathrm{g}}$. Theorem 5 states the consistency of the estimator.

Theorem 5. Suppose that the SNPs are C-dependent, $\omega > 0$, (2.8), and (2.11) hold, and $d_p/\sqrt{p} \to \infty$. Then, we have $\hat{\rho}_{\mathrm{g}} = \hat{\beta}_1 = \rho_{\mathrm{g}} + o_{\mathrm{P}}(1)$.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0060

Note that, under (2.11), $d_p/\sqrt{p} \to \infty$ iff $\sum_{j=1}^{p}(l_j - \bar{l})^2/\sqrt{p} \to \infty$, where l_j is the LD score and $\bar{l} = p^{-1}\sum_{j=1}^{p} l_j$. The proof of Theorem 5 is given in the Supplementary Material.

Next, we consider the asymptotic distribution of $\hat{\rho}_{\mathrm{g}}$. The result is relatively simpler in terms of the asymptotic variance under the assumption that

$$n_{\mathrm{o}} \;=\; o(n_1 \wedge n_2). \tag{2.16}$$

Thus, we consider this special case first. Define the following quantities: $\rho_b = \mathrm{cov}(b_{1j}, b_{2j}) = \omega - \omega_1\omega_2$, $\psi_0 = d_p/p$, $\psi_1 = p^{-1}\sum_{j,k=1}^{p}(u_j - \bar{u})(u_k - \bar{u})r_{jk}^2$, $\psi_{2,s} = \sum_{j,k=1}^{p}(u_j - \bar{u})(u_k - \bar{u})r_{jk}r_{ks}r_{sj}$, $\psi_{3,s} = p^{-1}\sum_{j,k=1}^{p}(u_j - \bar{u})(u_k - \bar{u})r_{jk}r_{ksj}$; $\psi_1(s,t) = \mathrm{E}(h_{1,s,t}^2)$, and $\psi_2(s,t) = \mathrm{E}(h_{1,s,t}h_{2,s,t})$, where $h_{1,s,t}$ and $h_{2,s,t}$ are the (s,t) elements of $H_1 = X'XDY'Y$ and $H_2 = Y'YDX'X$, respectively, and $D = \mathrm{diag}(u_j - \bar{u}, 1 \le j \le p)$.

Theorem 6. Suppose that the SNPs are C-dependent, $\omega > 0$, and (2.8), (2.11), and (2.16) hold. Further suppose that the following identities hold for all j, k, and s:

$$\mathrm{E}(X_{1j}X_{1k}^2X_{1s}) = \mathrm{E}(Y_{1j}Y_{1k}^2Y_{1s}),\ \mathrm{E}(X_{1j}^2X_{1k}^2X_{1s}^2) = \mathrm{E}(Y_{1j}^2Y_{1k}^2Y_{1s}^2).$$

Furthermore, suppose that the following limits exist as $p \to \infty$: $\phi_0 =$

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0060

$\lim \psi_0 > 0$, $\phi_1 = \lim \psi_1$, $\phi_r = \lim(n_./p^2) \sum_{s=1}^{p} \psi_{r,s}$, $r = 2, 3$, and

$$\vartheta_1 = \lim \frac{n_. n_1 n_2}{p^4} \sum_{t=1}^{p} \left\{ \sum_{s=1}^{p} (u_s - \bar{u}) r_{st}^2 \right\}^2,$$

$$\lambda_r = \lim \frac{n_.}{n_1 n_2 p^4} \sum_{s,t=1}^{p} \psi_r(s,t), \quad \mu_r = \lim \frac{n_.}{n_1 n_2 p^4} \sum_{s=1}^{p} \psi_r(s,s),$$

$r = 1, 2$. Then, we have $\sqrt{n_.}(\hat{\rho}_{\mathrm{g}} - \rho_{\mathrm{g}}) \xrightarrow{\mathrm{d}} N(0, \sigma^2)$, where $\sigma^2 = \sigma_1^2 + \sigma_2^2$ with

$$\sigma_1^2 = \rho_{\mathrm{g}}^2 \left(\frac{1}{\omega} - 1 \right) \left(\frac{\vartheta_1}{\phi_0^2} - \tau_1 - \tau_2 \right), \tag{2.17}$$

$$\sigma_2^2 = \frac{1}{\phi_0^2} \left\{ h_1^2 h_2^2 \left(\lambda_1 - \mu_1 + \frac{\omega \mu_1}{\omega_1 \omega_2} \right) + h_1^2 (1 - h_2^2)(\tau_1 \phi_2 + \phi_3) \right. \tag{2.18}$$

$$+ (1 - h_1^2) h_2^2 (\tau_2 \phi_2 + \phi_3) + (1 - h_1^2)(1 - h_2^2)(\tau_1 + \tau_2) \phi_1$$

$$\left. + \rho_{\mathrm{g}}^2 \left(\lambda_2 - \mu_2 + \frac{\mu_2}{\omega} \right) \right\}.$$

When b_{1j} and b_{2j} are uncorrelated, that is, $\omega = \omega_1 \omega_2$, the above asymptotic variance, σ^2, depends on ω^{-1}, but not on ω_1 and ω_2. A similar observation is made for (2.12). The proof of Theorem 6 is given in the Supplementary Material.

Finally, we extend Theorem 6 to not require (2.16). First, define the following additional quantities: $\psi_r(i,t) = \mathrm{E}(h_{r,i,t}^2), r = 3, 4$, where $h_{3,i,t}$ and $h_{4,i,t}$ are the (i,t) elements of $H_3 = XDY'Y$ and $H_4 = YDX'X$, respectively, and $\psi_5(i,t) = \mathrm{E}(h_{3,i,t} h_{4,i,t})$. Furthermore, define $\psi_6(i_1, i_2) = \mathrm{E}(h_{5,i_1,i_2}^2)$, where h_{5,i_1,i_2} is the (i_1, i_2) element of $H_5 = XDY'$. We now have

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0060

the extension of Theorem 6.

Theorem 7. Suppose that the conditions of Theorem 6 without (2.16) hold. In addition, suppose that the following limits exist:

$$\lambda_r = \lim \frac{n_\cdot}{n_1 n_2 p^3} \sum_{i=1}^{n_{r-2}} \sum_{t=1}^{p} \psi_r(i,t), \quad r = 3, 4,$$
$$\lambda_{5,\mathrm{o}} = \lim \frac{n_\cdot}{n_1 n_2 p^3} \sum_{i=1}^{n_\mathrm{o}} \sum_{t=1}^{p} \psi_5(i,t),$$
$$\lambda_6 = \frac{n_\cdot}{n_1 n_2 p^2} \sum_{i_1=1}^{n_1} \sum_{i_2=1}^{n_2} \psi_6(i_1,i_2), \ \lambda_{6,\mathrm{o}} = \frac{n_\cdot}{n_1 n_2 p^2} \sum_{i_1,i_2=1}^{n_\mathrm{o}} \psi_6(i_1,i_2).$$

Then, the conclusion of Theorem 6 holds with σ_2^2 replaced by the following:

$$\begin{aligned} \sigma_2^2 \ = \ & \frac{1}{\phi_0^2} \left[h_1^2 h_2^2 \left(\lambda_1 - \mu_1 + \frac{\omega \mu_1}{\omega_1 \omega_2} \right) + \rho_\mathrm{g}^2 \left(\lambda_2 - \mu_2 + \frac{\mu_2}{\omega_2} \right) \right. \\ & + (1-h_1^2) h_2^2 \lambda_3 + h_1^2 (1-h_2^2) \lambda_4 + 2 \rho_\mathrm{e} \rho_\mathrm{g} \lambda_{5,\mathrm{o}} \\ & \left. + (1-h_1^2)(1-h_2^2) \lambda_6 + \left\{ \rho_\mathrm{e} + \frac{(h_1^2 - h_2^2)^2}{2} \right\} \lambda_{6,\mathrm{o}} \right]. \end{aligned}$$

The proof of Theorem 7 is given in the Supplementary Material.

3. Simulation studies

We carried out comprehensive simulations to numerically validate our theoretical results. In these experiments, we evaluated the consistency of the summary-statistics-based heritability and genetic covariance estimators under a model misspecification. We also compared the empirical distributions of these estimators with the asymptotic distributions derived from our the-

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0060

ory. Unless explicitly stated, the heritability and genetic covariance estimators refer to those described in the previous sections.

3.1 Heritability

Figure 1.2 shows that the heritability estimator is nearly unbiased, even under a model misspecification. Following the same settings introduced in Section 1.2, we computed the observed standard deviation of the heritability estimator during 100 runs for each combination of the underlying model parameters (i.e., τ and ω). Then, we computed the corresponding theoretical standard errors of the estimators under different model settings using the formula derived in Theorem 2. All the limits presented in Theorem 2 are computed based on their corresponding observed values. For example, τ_2 is replaced by $u_{\cdot}^2/p$. As shown in Table 3.1, the values of the standard errors derived from our theory are very close to the observed standard errors under different combinations of τ and ω. Further evaluation results for the consistency and asymptotic normality of the heritability estimator are included in the Supplementary Material. The consistency and asymptotic normality of the heritability estimators under different settings look good.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0060

$\tau = 0.1$			$\omega = 0.05$		
ω	observed	theoretical	τ	observed	theoretical
0.005	0.41	0.39	0.05	0.37	0.36
0.05	0.17	0.20	0.1	0.17	0.20
0.25	0.17	0.18	0.25	0.13	0.13
1	0.18	0.17	0.5	0.10	0.11

Table 1: Observed and Theoretical Standard Errors of the Heritability Estimator

3.2 Genetic covariance

In the numerical examples presented in Section 1.2, we demonstrated the approximate unbiasedness of the genetic covariance estimator when the two studies share the same set of subjects. However, in practice, there are often few, if any, subjects shared between two GWAS, especially when they come from different cohorts. Here, we set $n_o/n = 0.1$, where $n = n_1 = n_2$. All other settings are the same as those described in Section 1.2. We calculated the observed standard deviation of the genetic covariance estimator based on 100 simulation runs under each parameter setting. We then computed the theoretical standard errors derived from Theorem 7. Similarly, the limits presented in the theorem are determined by their corresponding observed values. The comparisons of the observed and theoretical standard errors are compared in Table 3.2. The theoretical standard errors are close to the observed standard deviation, confirming our results for Theorems 6

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0060

and 7. We further investigated the consistency and asymptotic normality of the genetic covariance estimator. The empirical results are presented in the Supplementary Material. The consistency and asymptotic normality of the genetic covariance estimator look good under all settings. We also conducted additional simulations to investigate the relationship between the efficiency loss of the LDSC and model sparsity; the results are provided in the Supplementary Material. We found that the sparser the true model is, the greater is the efficiency loss. This makes intuitive sense, because the LD score regression is developed based on a polygenetic assumption.

$\tau = 0.1$, $n_o/n = 0.1$			$\omega = 0.05$, $n_o/n = 0.1$		
ω	observed	theoretical	τ	observed	theoretical
0.005	0.28	0.26	0.05	0.29	0.24
0.05	0.17	0.17	0.1	0.17	0.17
0.25	0.17	0.15	0.25	0.14	0.12
1	0.17	0.15	0.5	0.11	0.11

Table 2: Observed and Theoretical Standard Errors of Genetic Covariance Estimator

Next, we provide a real-data example that applies the LDSC to estimate the heritability and genetic covariance among four lipid traits: high-density lipoprotein (HDL) cholesterol, low-density lipoprotein (LDL) cholesterol, total cholesterol (TC), and triglyceride (TG). In the example, we compare the results from the LDSC with the REML estimates; see the Supplemen-

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0060

tary Material.

4. Discussion

The LDSC has become a popular method for estimating heritability and genetic correlation, owing to its efficiency and simplicity. We have examined the consistency and asymptotic normality of the LDSC under a misspecified model. Although the LDSC is based on a random-effects model, several methods have been proposed that estimate heritability and genetic correlation based on a fixed-effects model (Shi et al. (2017), Shi, Kichaev and Pasaniuc (2016), Wang and Li (2021), Guo et al. (2019)). It has been shown that under the assumption of a random-effects model, the estimator of the fixed-effects model converges to the estimator of the LDSC, almost surely (Wang and Li (2021)). When the assumption does not hold, neither model holds an advantage. One benefit of the random-effects model is that it incorporates implicit and automatic regularization of the regression coefficients, unlike in the case of a sparse fixed-effects model. The latter requires a careful choice of the penalty/thresholding parameters in orders to be effective. In addition, the random-effects model provides a systematic mechanism for carrying out statistical inference. In essence, this is achieved using the asymptotic distribution of the estimated heritability

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0060

and genetic correlation. Furthermore, methods based on the random-effects model are, in general, more computationally efficient. The fixed-effects model involves calculating the inverse of the LD matrix, which needs the additional assumption that the LD matrix is block-diagonal. On the other hand, methods based on the fixed-effects model require fewer assumptions on the genetic effects. Therefore, some researchers believe it is more robust across a wide range of genetic architectures, such as sparse causal SNPs (Wang and Li (2021)). However, we have proved that the LDSC can also provide a consistent estimator under a model misspecification.

The LDAK model (Speed and Balding (2019)) assumes that the variances of the SNP effects of the standardized SNPs are proportional to a set of known parameters $q_1, q_2, \ldots, q_p$, where p is the number of SNPs. This model can be viewed as a generalization of the LDSC. Indeed, when $q_1 = q_2 = \ldots = q_p$, the LDAK model reduces to the LDSC. In practice, the value of q_i for SNP i is a function of the MAF of SNP i, f_i. Our results can be extended to the LDAK model. Under the model of the LDSC, we have $E\left(z_j^2\right) = 1 + h^2\left(n/p\right) l_j$, where $j = 1, 2, \ldots, p$. Instead, in the LDAK, the regression problem changes to $E\left(z_j^2\right) = 1 + nh^2\left(\sum_{k=1}^{p} r_{jk}^2 q_k\right) / \sum_{k=1}^{p} q_k$, where r_{jk} is the correlation between SNP j and SNP k. Under an appropriate assumption for $q_1, q_2, \ldots, q_p$, the term $\left(\sum_{k=1}^{p} r_{jk}^2 q_k\right) p / \sum_{k=1}^{p} q_k$ is

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0060

interchangeable with l_j. However, for simplicity and to conserve space, we leave this extension to future work.

There are certain limitations in our theoretical assumptions. First, in practice, because the true LD matrix is unknown, we have to use an external reference panel to estimate the LD score. If the external data source used to estimate the LD scores is of higher order than n, which is the sample size of the GWAS, neither the consistency nor the asymptotic normality are affected. If the external sample size is of the same order as n, the consistency is not affected, but the asymptotic distribution will change. Second, we assume that the constant C in the C-dependent assumption is of $O(1)$. Actually, it is possible to allow C to increase, slowly, with n, so that the asymptotic results do not change. However, if the order of C exceeds a certain threshold, the asymptotic distribution, and even the consistency result, may change.

Supplementary Material

The online Supplementary Material contains our proofs and additional empirical results.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0060

Acknowledgments

The authors would like to thank the anonymous referees, associate editor, and editor for their constructive comments. Jiming Jiang and Debashis Paul's research was supported by NSF grant DMS-1713120. Wei Jiang and Yiliang Zhang's research was supported by NSF grant DMS-1902903 and NIH grant R01 GM134005. Hongyu Zhao's research was supported by NSF grants DMS-1713120 and DMS-1902903 and NIH grant R01 GM134005. We conducted the research using the UKBB resource under approved data requests (access ref: 29900).

References

Berisa, T. and Pickrell, J. K. (2016), Approximately independent linkage disequilibrium blocks in human populations, *Bioinformatics* 32, 283.

Berndt, S. I., Gustafsson, S., Mägi, R., Ganna, A., Wheeler, E., Feitosa, M. F., Justice, A. E., Monda, K. L., Croteau-Chonka, D. C., Day, F. R. and others (2013), Genome-wide meta-analysis identifies 11 new loci for anthropometric traits and provides insights into genetic architecture, *Nature Genetics* 45, 501–512.

Bulik-Sullivan, B. K., Loh, P.-R., Finucane, H. K., Ripke, S., Yang, J., Patterson, N., Daly, M. J., Price, A. L., and Neale, B. M. (2015a), LD Score regression distinguishes confounding from polygenicity in genome-wide association studies, *Nature Genetics* 47, 291–295.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0060

REFERENCES

Bulik-Sullivan, B., K., Finucane, H. K., Anttila, V., Gusev, A., Day, F. R., Loh, P.-R., Duncan, L., Perry, J. R. B., Patterson, N., Robinson, E. B. and others (2015b), An atlas of genetic correlations across human diseases and traits, *Nature Genetics* 47, 1236.

Buniello, A., MacArthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E. and others (2019), The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019, *Nucleic Acids Research* 47, D1005–D1012.

Clarke, L., Fairley, S., Zheng-Bradley, X., Streeter, I., Perry, E., Lowy, E., Tassé, A.-M., and Flicek, P. (2017), The international Genome sample resource (IGSR): A worldwide collection of genome variation incorporating the 1000 Genomes Project data, *Nucleic Acids Research* 45, D854–D859.

Guo, Z., Wang, W., Cai, T.T. and Li, H., (2019), Optimal estimation of genetic relatedness in high-dimensional linear models, *Journal of the American Statistical Association*, 114(525), pp.358-369.

Jiang, J. (2010). *Large Sample Techniques for Statistics*, Springer, New York.

Jiang, J., Li, C., Paul, D., Yang, C., and Zhao, H. (2016), On high-dimensional misspecified mixed model analysis in genome-wide association study, *The Annals of Statistics* 44, 2127–2160.

Jiang, W., Song, S., Hou, L., and Zhao, H. (2020), A set of efficient methods to generate high-dimensional binary data with specified correlation structures, arXiv:2007.14080.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0060

REFERENCES

Lee, S. H., Yang, J., Goddard, M. E., Visscher, P. M., and Wray, N. R. (2012), Estimation of pleiotropy between complex diseases using SNP-derived genomic relationships and restricted maximum likelihood, *Bioinformatics* 28, 2540–2542.

Shi, H., Kichaev, G. and Pasaniuc, B., (2016), Contrasting the genetic architecture of 30 complex traits from summary association data. *The American Journal of Human Genetics*, 99(1), pp.139-153.

Shi, H., Mancuso, N., Spendlove, S. and Pasaniuc, B., (2017), Local genetic correlation gives insights into the shared genetic architecture of complex traits, *The American Journal of Human Genetics*, 101(5), pp.737-751.

Shumway, R. H. and Stoffer, D. S. (2017), *Time Series Analysis and Its Applications*, Springer International Publishing AG.

Speed, D. and Balding, D.J., (2019), SumHer better estimates the SNP heritability of complex traits from summary statistics, *Nature genetics*, 51(2), pp.277-284.

Stephens, J. C., Schneider, J. A., Tanguay, D. A., Choi, J., Acharya, T., Stanley, S. E., Jiang, R., Messer, C. J., Chew, A., Han, J.-H. and others (2001), Haplotype variation and linkage disequilibrium in 313 human genes, *Science* 293, 489–493.

Sullivan, P. F., Daly, M. J., and O'donovan, M. (2012), Genetic architectures of psychiatric disorders: the emerging picture and its implications, *Nature Reviews Genetics* 13, 537–551.

Tenesa, A.and Haley, C. S. (2013), The heritability of human disease: estimation, uses and

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0060

abuses, *Nature Reviews Genetics* 14, 139–149.

van Rheenen, W., Peyrot, W. J., Schork, A. J., Lee, S. H., and Wray, N. R. (2019), Genetic correlations of polygenic disease traits: from theory to practice, *Nature Reviews Genetics* 20, 567–581.

Wang, J. and Li, H., (2021), Estimation of genetic correlation with summary association statistics, *Biometrika*.

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., Madden, P. A., Heath, A. C., Martin, N. G., Montgomery, G. W. and others (2010), Common SNPs explain a large proportion of the heritability for human height, *Nature Genetics* 42, 565–569.

Yang, J., Lee, S.H., Goddard, M.E. and Visscher, P.M., (2011), GCTA: a tool for genome-wide complex trait analysis, *The American Journal of Human Genetics*, 88(1), pp.76-82.

Zhang, Y., Cheng, Y., Jiang, W., Ye, Y., Lu, Q. and Zhao, H., (2021), Comparison of methods for estimating genetic correlation between complex traits using GWAS summary statistics, *Briefings in bioinformatics*, 22(5), p.bbaa442.

Zhao, B. and Zhu, H., (2021), On genetic correlation estimation with summary statistics from genome-wide association studies, *Journal of the American Statistical Association*, pp.1-11.

Zheng, J., Erzurumluoglu, A. M., Elsworth, B. L., Kemp, J. P., Howe, L., Haycock, P. C., Hemani, G., Tansey, K., Laurin, C., Pourcain, B. S. and others (2017), LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0060

of summary level GWAS data for SNP heritability and genetic correlation analysis, *Bioinformatics* 33, 272–279.

Department of Statistics, University of California, 399 Crocker Lane, Davis, CA 95616, USA.

E-mail: jimjiang@ucdavis.edu

Department of Biostatistics, Yale School of Public Health, 60 College Street, New Haven, CT 06510, USA.

E-mail: w.jiang@yale.edu

Department of Statistics, University of California, 399 Crocker Lane, Davis, CA 95616, USA.

E-mail: debpaul@ucdavis.edu

Department of Biostatistics, Yale School of Public Health, 60 College Street, New Haven, CT 06510, USA.

E-mail: yiliang.zhang@yale.edu

Department of Biostatistics, Yale School of Public Health, 60 College Street, New Haven, CT 06510, USA.

E-mail: hongyu.zhao@yale.edu