Statistica Sinica Preprint No: SS-2024-0092	
Title	The Method of Limits and Its Application to The Analysis of Count Data in Genome-wide Association Studies
Manuscript ID	SS-2024-0092
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202024.0092
Complete List of Authors	Jiming Jiang, Leqi Xu, Yiliang Zhang and Hongyu Zhao
Corresponding Authors	Jiming Jiang
E-mails	jimjiang@ucdavis.edu

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0092

THE METHOD OF LIMITS AND ITS APPLICATION TO THE ANALYSIS OF COUNT DATA IN GENOME-WIDE ASSOCIATION STUDIES

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

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

Abstract: We propose a new method of statistical inference, called the method of limits (MoL), which may be viewed as an extension of the method of moments. This method is motivated by the need to analyze count data for genome wide association studies (GWAS), where the existing methods are hindered in statistical inference due to computational challenges. We establish consistency and asymptotic normality of the MoL estimator of heritability from GWAS data, which is seen as an advantage over the existing PQLseq method. Furthermore, we derived a consistent estimator of the proportion of causal SNPs. MoL also showed an advantage of both statistical and computational efficiency measured by average statistical efficiency (ASE) in our simulation studies compared to PQLseq. We also illustrate the usefulness of MoL through its application to the UK Biobank data to infer the heritability of weekly champagne consumption and weekly red wine consumption using the count data.

Key words and phrases: asymptotic distribution, Big GWAS data, consistency, computa-

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0092

tion, MoL, proportion of causal SNPs, relative average statistical efficiency

1. Introduction

The method of moments (MoM) is a classical statistical method known to produce consistent estimators of model parameters. Although the method is known to be less efficient compared to the maximum likelihood (ML), MoM often has a computational advantage over the ML (e.g., Jiang and Nguyen (2021)). The latter is an attractive feature, especially in the modern era of Big Data. In fact, in large samples, the difference between ML and MoM estimators may be ignorable from a practical standpoint.

Despite its popularity, difficulties are encountered in executing the MoM idea. To see this, note that an MoM equation can often be expressed as

$$S = \mathrm{E}_\theta(S), \tag{1.1}$$

where S is a vector of statistics, and $\mathrm{E}_\theta(S)$ is the vector of expected values, or moments, of S under the parameter vector θ. Sometimes, the expression of $\mathrm{E}_\theta(S)$ is not simple. A consequence of this is that equation (1.1) may not have an analytical solution, or even a unique solution. Furthermore, numerically solving the equation may encounter convergence issues, and this is especially likely to happen when the dimension of θ is relatively high.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0092

Moreover, the right side of (1.1) may not even exist. For example, suppose that observations $X_1, \ldots, X_n$ are independent with the pdf

$$f_n(x|\theta) = \begin{cases} (n^2-1)/2n^2, & \text{if } x \in [\theta-1, \theta+1], \\ [2n^2 c\pi\{1+(x-\theta)^2\}]^{-1}, & \text{otherwise,} \end{cases} \tag{1.2}$$

that is, X_i has a uniform distribution between $\theta - 1$ and $\theta + 1$, with weight $1-n^{-2}$, and a Cauchy distribution elsewhere, with weight n^{-2}. Here, $c = 1/2 - \arctan(1)/\pi$ is a normalizing constant. Clearly, $\mathrm{E}(X_i)$ does not exist for any $1 \leq i \leq n$; thus, $\mathrm{E}_\theta(\bar{X})$ does not exist for $\bar{X} = n^{-1}\sum_{i=1}^n X_i$. On the other hand, it can be shown that $\bar{X} \xrightarrow{\mathrm{P}} \theta$. To see this, note that $\mathrm{P}_\theta(\max_{1\leq i\leq n}|X_i-\theta| > 1) \leq \sum_{i=1}^n \mathrm{P}_\theta(|X_i-\theta|>1) = n/n^2 = 1/n \to 0$. Thus, with probability tending to one $\bar{X} = n^{-1}\sum_{i=1}^n X_i 1_{(|X_i-\theta|\leq 1)}$, which can then be shown to converge to θ in probability.

In fact, the consistency of the MoM estimator involves showing that the left side of (1.1) converges in probability to a limit, which is a function of θ. When the right side of (1.1) exists, the limit is typically the same as $\mathrm{E}_\theta(S)$, or the limit of $\mathrm{E}_\theta(S)$. However, as the last example shows, it is possible that $\mathrm{E}_\theta(S)$ does not exist; and yet, S still has a limit as the sample size increases. Given that, it would seem necessary, anyway, to obtain the limit of the left side of (1.1) as a function of θ, regardless of whether the limit is equal to that of the right side [of (1.1)]. This motivates the following method: (i) Obtain the limit of the left

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0092

side of (1.1), say, in the sense of convergence in probability. (ii) Suppose that the limit in (i) is a known function of θ, say, $L(\theta)$. Then, an estimator of θ is obtained by solving

$$S \;=\; L(\theta). \tag{1.3}$$

We call the method described above the method of limits (MoL). A main advantage of MoL over MoM is that, the limiting function on the right side of (1.3) is often (much) simpler than the right side of (1.1), because certain lower-order terms disappear in the limit, leading to simpler, sometimes closed-form, solutions.

In the statistical literature, there are plenty examples of using limits as a (powerful) tool to obtain simplified solutions to difficult, sometime intractable problems (Jiang (2022)). A good example is the central limit theorem, which is often used to establish asymptotic normality of an estimator or test. Under such a simplification, a centralized estimator is asymptotically normal with mean zero and an asymptotic variance. While in many cases the asymptotic variance is fairly simple, there are situations in which the asymptotic variance is complicated. For example, in genome-wide association studies (GWAS), linear mixed models (LMM; e.g., Jiang and Nguyen (2021)) have been widely used since the seminal paper of Yang *et al.* (2010). As noted by Jiang *et al.* (2016), the LMM used in the GWAS context may be viewed as misspecified in the sense

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0092

that a large portion of the random effects associated with the single nucleotide polymorphisms (SNPs), which are assumed normal, are zero. Nevertheless, the latter authors were able to establish consistency and asymptotic normality of the restricted maximum likelihood (REML) estimators of genetic parameters of interest, such as the heritability, under the misspecified LMM. However, the variance of the limiting normal distribution is too complicated to be useful for inference. Due to such a concern, Dao *et al.* (2021) considered MoM estimators of the genetic parameters, which have much simpler asymptotic variances that can be used for inference.

The main motivation of the current paper is also GWAS. However, we are interested in situations where the phenotype data are counts. There have been extensions of the LMM to discrete responses in genetic studies such as binary observations in case-control studies (Golan *et al.* (2014) and counts (Sun *et al.* (2019)). In particular, the latter authors proposed inference based penalized quasi-likelihood (PQL; Breslow and Clayton (1993)) under a generalized linear mixed model (GLMM; e.g., Jiang and Nguyen (2021)). While PQL is computationally attractive, it is known to produce inconsistent estimators of the model parameters (Jiang (1998), Booth and Hobert (1999)), including variance components of genetic interest. On the other hand, ML estimation is known to be computationally infeasible under such a GLMM (e.g., Sun *et al.* (2019), Jiang

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0092

and Nguyen (2021)). It remains a challenging task to produce estimators that are computationally attractive as well as have good asymptotic behaviors.

The last sentence highlights the main contribution of our current paper. We demonstrate both theoretically and empirically the validity of statistical inference using MoL, especially in situations of big data. Specifically, we establish consistency and asymptotic normality of the MoL estimators of parameters of genetic interests, including variances associated with the genetic and environmental factors, as well as the heritability, under Big GWAS count data. Here, the term Big data refers to a data set whose sample size is beyond the computational capability of PQLseq (Sun *et al.* (2019)).

Furthermore, we obtain a consistent estimator of the proportion of causal SNPs, that is, the proportion of nonzero random effects associated with the SNPs. This proportion is involved in the asymptotic distribution of some genetic parameters of interest. It therefore plays an important role in deriving inferential methods, such as confidence intervals, based on the asymptotic distribution. To our knowledge, consistency of an estimator of such a proportion has not been rigorously established in the literature.

Finally, computational efficiency has become increasingly important in the era of Big data. It is desirable to consider performance of an estimator in terms of both statistical and computational efficiency. We introduce a notion, called aver-

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0092

age statistical efficiency (ASE), which combines the two types of efficiency into a single measure. In classical statistical inference, the reciprocal of the variance of the asymptotic distribution is viewed as a measure of statistical efficiency. The statistical efficiency is based purely on statistical considerations, which, in particular, has not taken into consideration the views of other professions in today's data science, such as those of computer scientists. When computing time is taken into consideration, it is reasonable to divide the statistical efficiency over the computing time. This leads to

$$\text{ASE} = \frac{1/\sigma^2}{c} = \frac{1}{\sigma^2 c}, \tag{1.4}$$

where σ^2 is the asymptotic variance, and c is the computing time needed in order to compute an estimator with the statistical efficiency, $1/\sigma^2$. Of course, c depends on the time unit and, more importantly, the computing facility. Therefore, the ASE is more useful when comparing two estimation methods under the same time unit and computing facility. This leads to the relative ASE, or RASE. Let σ_1^2, σ_2^2 denote the asymptotic variances of two estimation methods, say, Method 1 and Method 2, respectively, and c_1, c_2 be their corresponding computing times under the same unit and computing facility. Then, the RASE of Method 1 over Method 2 is defined as

$$\text{RASE} = \frac{1/(\sigma_1^2 c_1)}{1/(\sigma_2^2 c_2)} = \frac{\sigma_2^2 c_2}{\sigma_1^2 c_1}. \tag{1.5}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0092

In empirical studies, the asymptotic variance is typically replaced by the empirical (or simulated) variance of the estimator. This allows an investigator to compare performance of different methods under RASE. We show via extensive simulation that MoL has significant advantage over PQLseq in terms of RASE.

The asymptotic theory for MoL is established in Section 2, which also includes a consistent estimator of the proportion of causal SNPs. In Section 3, we present results of simulation studies, in which we compare MoL and PQLseq in terms of finite-sample performance. A real-life example of Big GWAS count data is discussed in Section 4. Technical proofs are deferred to Supplementary Material.

2. Asymptotic theory

We begin by first using a simple example to illustrate the idea of MoL for GWAS with count data. We then consider a general setting with continuous (normal) covariates. Finally, we consider a more general situation with both continuous and categorical covariates.

2.1 A simple-case demonstration

Similar to the setting of Sun *et al.* (2019), we assume that, given an $n \times p$ genotype matrix, Z, a $p \times 1$ vector of SNP-specific random effects, α, and an

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0092

$n \times 1$ vector of errors, ϵ, phenotype counts, $y_1, \ldots, y_n$ are conditionally independent such that $y_i|W \sim \text{Poisson}(Ne^{\eta_i})$. Here, $W = (Z, \alpha, \epsilon)$ and N is a known positive integer; furthermore, e^{η_i} is an unknown fraction that is assumed to satisfy

$$\eta_i = \gamma_i + \epsilon_i = \tilde{z}_i'\alpha + \epsilon_i, \tag{2.1}$$

where $\tilde{z}_i = z_i/\sqrt{p}$, z_i' is the ith row of Z, and ϵ_i is the ith component of ϵ. It is furthermore assumed that Z, α, ϵ are independent, the entries of Z are independent and standard normal, $\alpha \sim N(0, \sigma_1^2 I_p)$, and $\epsilon \sim N(0, \sigma_0^2 I_n)$, where σ_0^2, σ_1^2 are unknown variances. See, for example, Yang *et al.* (2010) for the model setting in the linear case; the current model can be interpreted similarly.

Our immediate goal is to estimate σ_0^2 and σ_1^2. For that we need to construct two statistics, S_0 and S_1, and find their limits. The first seems to be obvious: $S_0 = \bar{y} = n^{-1}y_\cdot$ with $y_\cdot = \sum_{i=1}^n y_i$, and we have the following result.

Lemma 1. As, $n, p \to \infty$, we have $\bar{y} = n^{-1}\sum_{i=1}^n y_i \xrightarrow{\text{P}} Ne^{(\sigma_0^2+\sigma_1^2)/2}$.

As for the next candidate, it is less obvious. However, it is seen in Lemma 1 that the limit of $\bar{y}$ is a function of $\sigma_0^2 + \sigma_1^2$; therefore, one needs something whose limit is not another function of $\sigma_0^2 + \sigma_1^2$ (otherwise, one cannot separate the two variances). The following lemma shows what may be a right candidate.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0092

Lemma 2. Suppose that $n, p \to \infty$ such that

$$\frac{p}{n^2} \quad \to \quad 0. \tag{2.2}$$

Then, we have $T_1 = n^{-2} \sum_{i_1 \neq i_2} z'_{i_1} z_{i_2} y_{i_1} y_{i_2} \xrightarrow{\text{P}} N^2 \sigma_1^2 e^{\sigma_0^2 + \sigma_1^2}$.

Note. Condition (2.2) requires that $p << n^2$. This is typically reasonable in GWAS applications. 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). This means n is approximately 500,000 and p is about 1,000,000; therefore, $p/n^2 \approx 4 \times 10^{-6}$. From a theoretical standpoint, a standard assumption in random matrix theory is that $n, p \to \infty$ such that $p/n \to \gamma \in (0, \infty)$. See, for example, Jiang (2022) [ch. 16, in particular, (16.13)]. Clearly, assumption (2.2) is weaker than the standard assumption.

In view of these lemmas, it is natural to consider $S_1 = T_1/\bar{y}^2$, whose limit is precisely σ_1^2. This leads to the following MoL equations:

$$S_0 \quad = \quad e^{(\sigma_0^2 + \sigma_1^2)/2}, \tag{2.3}$$

$$S_1 \quad = \quad \sigma_1^2. \tag{2.4}$$

(2.3) and (2.4) lead to closed-form solutions, which are the MoL estimators:

$$\hat{\sigma}_1^2 = \frac{1}{y^2} \sum_{i_1 \neq i_2} z'_{i_1} z_{i_2} y_{i_1} y_{i_2}, \quad \hat{\sigma}_0^2 = 2(\log \bar{y} - \log N) - \hat{\sigma}_1^2. \tag{2.5}$$

Note. One advantage of MoL over MoM is that it is not always possible to obtain analytical expressions for the moments, or that the analytical expressions

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0092

are too complicated that a closed-form expression of the estimator is not possible. For example, one can derive an analytical expression for $\mathrm{E}(S_0)$; however, an analytical expression of $\mathrm{E}(S_1)$ is not available. If, instead, one consider the MoM equations $\bar{y} = \mathrm{E}(\bar{y})$ and $T_1 = \mathrm{E}(T_1)$, the expectations have analytical forms, but they are too complex that the MoM equations do not have a closed-form solution. Although, for data of moderate size, a closed-form expression of the estimator may not make much difference, so far as computation is concerned, for Big data there can be a major difference, as we shall show below.

The asymptotic theory, to be established in the sequel, implies that (2.5) are consistent estimators of σ_1^2, σ_0^2, respectively (the results follow by applying similar arguments to parts of the proof of Theorem 1, given in the supplementary material). To have some idea about how these estimators perform empirically, a small-scale simulation study was run. We consider two scenarios: (I) $\sigma_0^2 = 0.7, \sigma_1^2 = 0.3$; (II) $\sigma_0^2 = 0.4, \sigma_1^2 = 0.6$. $N = 10$ in both cases. The results, based on 100 simulation runs, are presented in Table 1. The good performance of MoL can be seen in this table.

2.2 Continuous covariates

We now make several extensions of the simple model described above in (2.1). First, we allow the total count, N, to vary among the individuals; in other words,

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0092

Table 1: **Empirical Performance of MoL Estimators**

	(I)				(II)			
True Parameter	$\sigma_0^2 = 0.7$		$\sigma_1^2 = 0.3$		$\sigma_0^2 = 0.4$		$\sigma_1^2 = 0.6$	
Performance Measure	Mean	s.d.	Mean	s.d.	Mean	s.d.	Mean	s.d.
$n = 200, p = 500$	0.726	0.458	0.253	0.443	0.450	0.439	0.589	0.492
$n = 200, p = 1000$	0.744	0.644	0.265	0.649	0.376	0.581	0.585	0.582
$n = 500, p = 500$	0.703	0.195	0.296	0.219	0.447	0.149	0.532	0.186
$n = 500, p = 1000$	0.686	0.253	0.307	0.244	0.448	0.270	0.545	0.298
$n = 1000, p = 1000$	0.700	0.158	0.286	0.154	0.399	0.165	0.591	0.177
$n = 1000, p = 2000$	0.693	0.218	0.310	0.216	0.401	0.202	0.606	0.229

N is replaced by N_i for y_i, where N_i is a known positive integer, $1 \leq i \leq n$. Second, we allow some random effects to be (exactly) zero; the nonzero random effects thus correspond to the causal SNPs. To do so, let $\alpha = b \circ \xi = (b_j \xi_j)_{1 \leq j \leq p}$, where $b = (b_j)_{1 \leq j \leq p}$, $\xi = (\xi_j)_{1 \leq j \leq p}$ so that $b_j \sim \text{Bernoulli}(\omega)$, where $\omega \in (0, 1]$ is an unknown probability known as the proportion of causal SNPs, and $\xi_j \sim N(0, \sigma_1^2)$. We further assume that $b_j, \xi_j, j = 1, \ldots, p$ are independent, and Z, α, ϵ are independent. Finally, the entries of Z are assumed to be independent sub-Gaussian (but remain standardized, that is, with mean 0 and variance 1), as in Jiang *et al.* (2016). As noted by the latter authors, the asymptotic results can be extended to the case where the Z matrix is standardized.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0092

A further extension is made by replacing the η_i in (2.1) by

$$\tilde{\eta}_i = \beta_0 + x_i'\beta + \tilde{z}_i'\alpha + \epsilon_i = \beta_0 + x_i'\beta + \eta_i, \quad i = 1, \ldots, n, \tag{2.6}$$

where β_0 is an unknown intercept, β is a vector of unknown parameters, x_i is a vector of observed covariates. It is assumed that, conditional on $X = (x_i')_{1\le i\le n}$ and $N = (N_i)_{1\le i\le n}$, we have $y_i|W \sim \text{Poisson}(e^{\tilde{\eta}_i}N_i)$, where $W = (Z, \alpha, \epsilon)$, and the distribution of Z does not depend on X and N.

As for X and N, it is assume that $x_1, \ldots, x_n$ are independent following a q-dimensional multivariate normal distribution with mean vector b and covariance matrix B, where b, B are unknown and B is positive definite. Some extension beyond the normality is possible, although normality simplifies the results considerably. Furthermore, we assume that $N_1, \ldots, N_n$ are i.i.d. with a finite 4th moment, and X, N are independent. Note that we can write $x_i = b + B^{1/2}\tilde{x}_i$, where $B^{1/2}$ is the symmetric square root of B, and $\tilde{x}_i = B^{-1/2}(x_i - b) \sim N(0, I_q)$. Then, (2.6) can be written as

$$\tilde{\eta}_i = \beta_0 + b'\beta + \tilde{x}_i'B^{1/2}\beta + \eta_i = \mu + \tilde{x}_i'\tilde{\beta} + \eta_i, \tag{2.7}$$

where $\mu = \beta_0 + b'\beta$ and $\tilde{\beta} = B^{1/2}\beta$. Let $\sigma_\alpha^2 = \omega\sigma_1^2$ and $\tau^2 = \beta' B\beta$. Similar to Lemma 1 and Lemma 2, we have the following results.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0092

Lemma 3. Under the assumed model, the following limits can be obtained:

$$\bar{y} \xrightarrow{\mathrm{P}} e^{\mu+(\sigma_0^2+\sigma_\alpha^2+\tau^2)/2}\mathrm{E}(N_1), \tag{2.8}$$

$$T_1 \xrightarrow{\mathrm{P}} \sigma_\alpha^2 e^{2\mu+\sigma_0^2+\sigma_\alpha^2+\tau^2}\{\mathrm{E}(N_1)\}^2, \tag{2.9}$$

$$\frac{1}{n}\sum_{i=1}^{n} y_i(y_i-1) \xrightarrow{\mathrm{P}} e^{2(\mu+\sigma_0^2+\sigma_\alpha^2+\tau^2)}\mathrm{E}(N_1^2), \tag{2.10}$$

$$\frac{1}{n^2}\sum_{i_1\neq i_2} \hat{x}_{i_1}'\hat{x}_{i_2} y_{i_1} y_{i_2} \xrightarrow{\mathrm{P}} \tau^2 e^{2\mu+\sigma_0^2+\sigma_\alpha^2+\tau^2}\{\mathrm{E}(N_1)\}^2, \tag{2.11}$$

provided that $n, p \to \infty$ such that (2.2) holds, where $\hat{x}_i$ is $\tilde{x}_i$ with b and B replaced by $\bar{x} = n^{-1}\sum_{i=1}^n x_i$ and $S_x = (n-1)^{-1}\sum_{i=1}^n (x_i-\bar{x})(x_i-\bar{x})'$, respectively.

(2.8)–(2.11) lead to the following MoL estimators of $\sigma_\alpha^2, \tau^2, \sigma_0^2$ and μ:

$$\hat{\sigma}_\alpha^2 = \frac{1}{y_\cdot^2}\sum_{i_1\neq i_2} z_{i_1}' z_{i_2} y_{i_1} y_{i_2}, \tag{2.12}$$

$$\hat{\tau}^2 = \frac{1}{y_\cdot^2}\sum_{i_1\neq i_2} \hat{x}_{i_1}'\hat{x}_{i_2} y_{i_1} y_{i_2}, \tag{2.13}$$

$$\hat{\sigma}_0^2 = \log\{y(y-1)\}_\cdot - 2\log y_\cdot + 2\log N_\cdot - \log(N^2)_\cdot - \hat{\sigma}_\alpha^2 - \hat{\tau}^2, \tag{2.14}$$

$$\hat{\mu} = \log y_\cdot - \log N_\cdot - \frac{1}{2}(\hat{\sigma}_0^2 + \hat{\sigma}_\alpha^2 + \hat{\tau}^2), \tag{2.15}$$

where $\{y(y-1)\}_\cdot = \sum_{i=1}^n y_i(y_i-1)$, $N_\cdot = \sum_{i=1}^n N_i$ and $(N^2)_\cdot = \sum_{i=1}^n N_i^2$. Note that all these estimators have closed-form expressions. Not only that, the theorem below guarantees that they are consistent estimators.

Theorem 1. Under the assumed model, (2.12)–(2.15) are consistent estima-

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0092

tors of σ_α^2, τ^2, σ_0^2 and μ, respectively, provided that $n, p \to \infty$ and (2.2) holds.

Among the parameters involved in Theorem 1, two are of genetic interest, namely, σ_α^2 and σ_0^2. The rest of this subsection is devoted to deriving asymptotic distribution of $\hat{\sigma}_\alpha^2$ and $\hat{\sigma}_0^2$. Such a result can be used for inference about these parameters. Define $\psi = \omega^{-1}$, $\sigma^2 = \sigma_0^2 + \sigma_\alpha^2 + \tau^2$. Denote the right sides of (2.8)–(2.11) by $b_r, r = 1, 2, 3, 4$, respectively. For any random variable u with nonzero mean, u_* is defined as $u/\mathrm{E}(u)$.

Theorem 2. Under the conditions of Theorem 1, with (2.2) strengthened to

$$\frac{n}{p} \;\to\; \gamma \in (0, \infty), \tag{2.16}$$

we have (I) $\sqrt{n}(\hat{\sigma}_\alpha^2 - \sigma_\alpha^2) \xrightarrow{\mathrm{d}} N(0, v_1^2)$, where

$$v_1^2 = \sigma_\alpha^4 \left[\gamma(3\psi - 1) + 4 \left\{ e^{\sigma^2} \mathrm{E}(N_1)_*^2 + \frac{b_1 + b_3}{b_2} \right\} \right]; \tag{2.17}$$

and (II) $\sqrt{n}(\hat{\sigma}_0^2 - \sigma_0^2) \xrightarrow{\mathrm{d}} N(0, v_0^2)$, where

$$\begin{aligned} v_0^2 \;=\; & 4 \left[\{(\sigma_\alpha^2 + \tau^2 + 1)^2 + \sigma_\alpha^2 + \tau^2\} e^{\sigma^2} - 1 \right] \mathrm{E}\{(N_1)_*^2\} \\ & - 4\{(2\sigma_\alpha^2 + 2\tau^2 + 1) e^{2\sigma^2} - 1\} \mathrm{E}\{(N_1)_* (N_1^2)_*\} \\ & + (e^{4\sigma^2} - 1) \mathrm{E}\{(N_1^2)_*^2\} \\ & + \frac{4}{b_1} \left[e^{\sigma^2} \frac{\mathrm{E}(N_1)\mathrm{E}(N_1^3)}{\{\mathrm{E}(N_1^2)\}^2} - (\sigma_\alpha^2 + \tau^2 + 1) \right] + \frac{2}{b_3}. \end{aligned} \tag{2.18}$$

Note 1. To see the right side of (2.18) is nonnegative, define random variables $Y_r, r = 1, 2, 3$ as independent, and independent with N_1, such that $Y_1 \sim$

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0092

$N(0,\sigma_0^2)$, $Y_2 \sim N(0,\sigma_\alpha^2)$, $Y_3 \sim N(0,\tau^2)$, and that, conditional on $N_1, Y_r, r = 1,2,3$, $Y \sim \text{Poisson}(N_1 e^\eta)$ with $\eta = \mu + Y_1 + Y_2 + Y_3$. Then, it can be shown that the right side of (2.18) is equal to

$$\text{E}\left\{\frac{2}{b_1}(\sigma_\alpha^2+\tau^2-1-Y_2-Y_3)Y + \frac{Y(Y-1)}{b_3} + 2(N_1)_* - (N_1^2)_*\right\}^2. \quad (2.19)$$

Note 2. Unlike v_1^2, v_0^2 does not depend on either ω or γ (note that $\psi = \omega^{-1}$).

So far, all of the unknown parameters involved in $v_s^2, s = 0,1$, except $\psi = \omega^{-1}$, have their consistent estimators. Namely, μ, σ_0^2, σ_α^2, τ^2, γ, and $\text{E}(N_1^k), k = 1,2,3,4$ can be consistently estimated by $\hat\mu$, $\hat\sigma_0^2$, $\hat\sigma_\alpha^2$, $\hat\tau^2$, n/p, and $\overline{N^k} = n^{-1}\sum_{i=1}^n N_i^k, k = 1,2,3,4$, respectively. A consistent estimator of ψ is given below.

Lemma 4. We have $T_2 \xrightarrow{\text{P}} 3\psi b_2^2$, where

$$T_2 = \frac{p}{n(n-1)(n-2)(n-3)} \sum_{j=1}^{p} \sum_{i_1,i_2,i_3,i_4 \text{ distinct}} z_{i_1 j} z_{i_2 j} z_{i_3 j} z_{i_4 j} y_{i_1} y_{i_2} y_{i_3} y_{i_4}$$

under the conditions of Theorem 2.

Combining Lemma 4 and (2.9), the following result immediately follows.

Theorem 3. $\hat\psi = T_2/3T_1^2 \xrightarrow{\text{P}} \psi$ under the conditions of Theorem 2.

Now all of the parameters involved in (2.17) and (2.18) have their consistent estimators. Thus, by replacing the asymptotic variance by its consistent estimator, inference about σ_α^2 or σ_0^2, such as confidence intervals and tests, can be made. Note that such inferential methods are not available for PQLseq (see Sun

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0092

et al. (2019)).

Theorem 2 gives the asymptotic distribution of $\hat{\sigma}^2_\alpha$ and $\hat{\sigma}^2_0$ separately. In fact, a joint asymptotic distribution of these two estimators can be obtained. This is considered under a more general setting in the final subsection of this section regarding heritability estimation.

2.3 Continuous and categorical covariates

In practice, the covariates X may not be all normally distributed or, at least, continuous. For example, some of the covariates may be binary indicators. Thus, in still another extension, we assume that the covariates are divided into two groups; the first group are continuous; the second group are discrete or categorical. The continuous covariates are assumed to be jointly multivariate normal as above, and the discrete/categorical covariates are assumed to be multinomial. Such a setting has been proposed in the literature, when the distribution of the covariates is considered. See, for example, Little and Rubin (2002).

Specifically, for the discrete/categorical covariates, there is a set of different combinations of the their values that are present in the data, denoted by $c_1, \ldots, c_K$. Let y_{ik}, x_{ik} denote the phenotype count and vector of continuous covariates of the ith individual in the kth group corresponding to c_k, $i = 1, \ldots, n_k$, where n_k is the total number of individuals in the kth group. The c_k's are con-

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0092

sidered fixed and known. The continuous covariates are $X = (X_k)_{1\le k\le K}$ with $X_k = (x'_{ik})_{1\le i\le n_k}$. Similarly, we have $Z = (Z_k)_{1\le k\le K}$ with $Z_k = (z'_{ik})_{1\le i\le n_k}$, where $z_{ik} = (z_{ijk})_{1\le j\le p}$. Let α be defined as at the beginning of this section, and $\epsilon = (\epsilon_k)_{1\le k\le K}$ with $\epsilon_k = (\epsilon_{ik})_{1\le i\le n_k}$. Also let $N = (N_k)_{1\le k\le K}$, where $N_k = (N_{ik})_{1\le i\le n_k}$ and N_{ik} are known positive integers. We assumed that, conditional on X, N and $W = (Z, \alpha, \epsilon)$, $y_{ik}, i = 1, \ldots, n_k$, $k = 1, \ldots, K$ are conditionally independent such that $y_{ik}|X, N, W \sim \text{Poisson}(N_{ik}e^{\eta_{ik}})$, where

$$\eta_{ik} = c_k + x'_{ik}\beta + \tilde{z}'_{ik}\alpha + \epsilon_{ik}, \tag{2.20}$$

$\tilde{z}_{ik} = z_{ik}/\sqrt{p}$, and β is a vector of unknown fixed effects. Furthermore, assume that, conditional on (X, N), Z, α, ϵ are independent; the entries of Z are i.i.d. sub-Gaussian; and the entries of ϵ are independent $N(0, \sigma_0^2)$. Finally, assume that x_{ik}, N_{ik}, $i = 1, \ldots, n_k$, $k = 1, \ldots, K$ are independent such that $x_{ik} \sim N(b_k, B_k)$, where b_k is an unknown mean vector and B_k an unknown nonsingular covariance matrix, and $N_{ik}, i = 1, \ldots, n_k$ are i.i.d. with a finite fourth moment.

An observation is that, within each group k, we are in the same situation as the one considered previously, that is, the right side of (2.7), with μ replaced by $\mu_k = c_k + b'_k\beta$ and $\tilde{\beta}$ replaced by $\tilde{\beta}_k = B_k^{1/2}\beta$. Thus, according to the earlier results, we have the MoL estimators of σ_α^2, $\tau_k^2 = \beta' B_k \beta$, σ_0^2, and μ_k from the kth

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0092

group. To repeat these expressions, the MoL estimators are given by

$$\hat{\sigma}^2_{\alpha,k} = \frac{1}{y^2_{\cdot k}} \sum_{i_1 \neq i_2} z'_{i_1 k} z_{i_2 k} y_{i_1 k} y_{i_2 k}, \tag{2.21}$$

$$\hat{\tau}^2_k = \frac{1}{y^2_{\cdot k}} \sum_{i_1 \neq i_2} \hat{x}'_{i_1 k} \hat{x}_{i_2 k} y_{i_1 k} y_{i_2 k}, \tag{2.22}$$

$$\hat{\sigma}^2_{0,k} = \log\{y_k(y_k - 1)\}_\cdot - 2 \log y_{\cdot k} + 2 \log N_{\cdot k} - \log(N^2_k)_\cdot. \\ -\hat{\sigma}^2_{\alpha,k} - \hat{\tau}^2_k, \tag{2.23}$$

$$\hat{\mu}_k = \log y_{\cdot k} - \log N_{\cdot k} - \frac{1}{2}(\hat{\sigma}^2_{0,k} + \hat{\sigma}^2_{\alpha,k} + \hat{\tau}^2_k), \tag{2.24}$$

where $y_{\cdot k} = \sum_{i=1}^{n_k} y_{ik}$, $N_{\cdot k} = \sum_{i=1}^{n_k} N_{ik}$, $\{y_k(y_k - 1)\}_\cdot = \sum_{i=1}^{n_k} y_{ik}(y_{ik} - 1)$, $(N^2_k)_\cdot = \sum_{i=1}^{n_k} N^2_{ik}$, and $\hat{x}_{ik} = \hat{B}_k^{-1/2}(x_{ik} - \overline{x_k})$ with $\overline{x_k} = n_k^{-1} \sum_{i=1}^{n_k} x_{ik}$ and

$$\hat{B}_k = \frac{1}{n_k - 1} \sum_{i=1}^{n_k} (x_{ik} - \overline{x_k})(x_{ik} - \overline{x_k})'.$$

Let $n = \sum_{k=1}^{K} n_k$ be the total sample size. As our main interest is σ^2_α and σ^2_0, we take weighted averages of (2.21) and (2.23) over $1 \leq k \leq K$ to get

$$\hat{\sigma}^2_\alpha = \frac{1}{n} \sum_{k=1}^{K} n_k \hat{\sigma}^2_{\alpha,k}, \quad \hat{\sigma}^2_0 = \frac{1}{n} \sum_{k=1}^{K} n_k \hat{\sigma}^2_{0,k}. \tag{2.25}$$

By Theorem 1, the following result immediately follows.

Theorem 4. Suppose that K is fixed, and the assumptions of Theorem 1 hold for every group k, $1 \leq k \leq K$. Then, the estimators (2.25) are consistent.

Note. Typically in GWAS, the mixed effects model is applied to the "residuals" after the fixed effects have been "subtracted", so that one can focus on

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0092

estimating the genetic and environmental variance components, that is, σ_α^2 and σ_ϵ^2 via the mixed effects model. This means that the c_k and x_{ik} on the right side of (2.20) are actually zeros. We allowed our method to be more general to involve continuous and discrete covariates, but, at least in real-life GWAS, we have not encountered a case with large K.

The next result is an extension of Theorem 2 under the more general setting.

Theorem 5. Suppose that (2.16) holds with n, γ replaced by n_k, γ_k, respectively, for $1 \le k \le K$. Then, we have (I) $\sqrt{n}(\hat{\sigma}_\alpha^2 - \sigma_\alpha^2) \xrightarrow{\mathrm{d}} N(0, v_1^2)$, where

$$v_1^2 = \sigma_\alpha^4 \left[\gamma_\cdot \left(\frac{3}{\omega} - 1 \right) + \frac{4}{\gamma_\cdot} \sum_{k=1}^{K} \gamma_k \left\{ e^{\sigma_k^2} \mathrm{E}(N_{11})_*^2 + \frac{b_{1k} + b_{3k}}{b_{2k}} \right\} \right], \quad (2.26)$$

$\gamma_\cdot = \sum_{k=1}^K \gamma_k$, $\sigma_k^2, b_{rk}, r = 1, 2, 3$ are, respectively, $\sigma^2, b_r^2, r = 1, 2, 3$ with τ^2 replaced by τ_k^2. Furthermore, we have (II) $\sqrt{n}(\hat{\sigma}_0^2 - \sigma_0^2) \xrightarrow{\mathrm{d}} N(0, v_0^2)$, where $v_0^2 = \gamma_\cdot^{-1} \sum_{k=1}^K \gamma_k v_{0k}^2$, v_{0k}^2 being the v_0^2 of (2.18) with N_1 replaced by N_{11}, and $\tau^2, \sigma^2, b_1, b_3$ replaced by $\tau_k^2, \sigma_k^2 = \sigma_0^2 + \sigma_\alpha^2 + \tau_k^2, b_{1k}, b_{3k}$, respectively, $1 \le k \le K$.

Notes. Similar to Note 1 following Theorem 2, the positivity of v_{0k}^2 can be shown in a similar way [see (2.19)], with $\tau^2, Y_3, \eta, Y, b_1, b_3$ replaced by τ_k^2, Y_{3k}, η_k, Y_k, b_{1k}, b_{3k}, respectively, $1 \le k \le K$, and $N_1 \sim N_{11}$. Also note that, unlike v_1^2, v_0^2 does not depend on ω, but it does depend on the γs (this is different from the v_0^2 of Theorem 2, unless $K = 1$).

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0092

For estimating ψ, let z_{ijk} be the jth component of z_{ik}. Define

$$T_{1k} = \frac{1}{n_k^2} \sum_{1 \le i_1 \ne i_2 \le n_k} z'_{i_1 k} z_{i_2 k} y_{i_1 k} y_{i_2 k},$$

$$T_{2k} = \frac{p \sum_{j=1}^{p} \sum_{1 \le i_1, i_2, i_3, i_4 \le n_k, \text{ distinct}} z_{i_1 jk} z_{i_2 jk} z_{i_3 jk} z_{i_4 jk} y_{i_1 k} y_{i_2 k} y_{i_3 k} y_{i_4 k}}{n_k (n_k - 1)(n_k - 2)(n_k - 3)}.$$

Also define $T_1^2 = n^{-1} \sum_{k=1}^{K} n_k T_{1k}^2$ and $T_2 = n^{-1} \sum_{k=1}^{K} n_k T_{2k}$. By Lemma 4 and (2.9), a consistent estimator of ψ can be obtained. We state the result formally.

Theorem 6. $\hat{\psi} = T_2 / 3T_1^2 \xrightarrow{\text{P}} \psi$ under the conditions of Theorem 5.

As noted, the results of Theorem 5 and Theorem 6 can be used for inferential purposes. In this regard, the following computational note may be useful.

Computational note. To compute the inner summation over four distinct indexes involved in the T_2 in Lemma 4, or the T_{2k} above, note that, for example, by letting $\lambda_i = z_{ij} y_i$ for fixed j, the inner summation is in the form of

$$\begin{aligned} & \sum_{1 \le i_1, i_2, i_3, i_4 \le n, i_1, i_2, i_3, i_4 \text{ distinct}} \lambda_{i_1} \lambda_{i_2} \lambda_{i_3} \lambda_{i_3} \\ = \; & s_{\lambda,1}^4 - 6 s_{\lambda,1}^2 s_{\lambda,2} + 8 s_{\lambda,1} s_{\lambda,3} + 3 s_{\lambda,2}^2 - 6 s_{\lambda,4}, \end{aligned} \tag{2.27}$$

where $s_{\lambda,r} = \sum_{i=1}^{n} \lambda_i^r, r = 1, 2, 3, 4$. The right side of (2.27) is much easier to compute.

2.4 Heritability estimation

Sun *et al.* (2019)) defines the heritability as the ratio of genetic variation over total variation in the scale of the linear predictor under the GLMM, conditional

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0092

on X. Under this definition, the heritability is simply $h^2 = \sigma_\alpha^2/(\sigma_\alpha^2+\sigma_0^2)$. Define $\hat{h}^2 = \hat{\sigma}_\alpha^2/(\hat{\sigma}_\alpha^2+\hat{\sigma}_0^2)$, where $\hat{\sigma}_\alpha^2, \hat{\sigma}_0^2$ are the MoL estimators of $\sigma_\alpha^2, \sigma_0^2$, respectively.

First consider the setting of Section 2.2. Theorem 1 immediately implies the following.

Corollary 1. Under the assumptions of Theorem 1, we have $\hat{h}^2 \xrightarrow{\mathrm{P}} h^2$.

To obtain the asymptotic distribution of $\hat{h}$, we need to first strengthen Theorem 2 to obtain the joint asymptotic distribution of $\hat{\sigma}_\alpha^2$ and $\hat{\sigma}_0^2$. We have the following result.

Theorem 7. Under the assumptions of Theorem 2, we have

$$\sqrt{n}\begin{pmatrix} \hat{\sigma}_\alpha^2 - \sigma_\alpha^2 \\ \hat{\sigma}_0^2 - \sigma_0^2 \end{pmatrix} \xrightarrow{\mathrm{d}} N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} v_1^2 & v_{10} \\ v_{10} & v_0^2 \end{pmatrix}\right], \tag{2.28}$$

where v_1^2, v_0^2 are given in Theorem 2, and

$$v_{10} = 4e^{\sigma^2}\left[\sigma_\alpha^2 e^{\sigma^2}\mathrm{E}\{(N_1)_*(N_1^2)_*\} - (2\sigma_\alpha^4+\tau^2+1)\mathrm{E}\{(N_1)_*^2\}\right].$$

By Theorem 7 and the delta method (e.g., Jiang 2022, p. 94), the following result immediately follows.

Corollary 2. Under the assumptions of Theorem 2, we have

$$\sqrt{n}(\hat{h}^2 - h^2) \xrightarrow{\mathrm{d}} N(0, \nu^2),$$

where $\nu^2 = (\sigma_\alpha^2+\sigma_0^2)^{-4}(v_0^2\sigma_\alpha^4 - 2v_{10}\sigma_\alpha^2\sigma_0^2 + v_1^2\sigma_0^4)$.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0092

Corollary 2 and Theorem 3 can be utilized to make inference about h^2. Finally, under the more general setting of Section 2.3. The following result can be established.

Theorem 8. Under the assumptions of Theorem 5, we have (2.28), where $\hat{\sigma}_\alpha^2$, $\hat{\sigma}_0^2$, v_1^2, v_0^2 are the same as in Theorem 5, and $v_{01} = \gamma_\cdot^{-1} \sum_{k=1}^K \gamma_k v_{10,k}$, $v_{10,k}$ being the v_{10} in Theorem 7 with τ^2, σ^2, and N_1 replaced by τ_k^2, σ_k^2, and N_{11}, respectively, $1 \leq k \leq K$.

Under the more general setting, $\hat{h}^2$ is defined in the same way with newly defined $\hat{\sigma}_\alpha^2$ and $\hat{\sigma}_0^2$. Thus, by the same arguments, Corollary 2 holds under this setting with v_0^2, v_1^2, v_{10} given by Theorem 8. Theorem 6 can then be utilized to make inference about h^2.

3. Simulation studies

We performed comprehensive simulation studies to validate the theoretical properties of our proposed MoL method, including the consistency and asymptotic normality of the variance estimators, as well as the consistency of the estimator of the proportion of causal SNPs. The heritability estimation results derived from our MoL approach were compared with those from the PQLseq method in these experiments.

The genotype matrix Z was initial generated from a $\mathrm{Binomial}(2, 0.5 * p)$

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0092

distribution, where p follows a Beta distribution, $\mathrm{Beta}(0.5, 0.5)$. Then, each entry in the matrix was standardized to have a mean of 0 and a variance of 1. Recall that we use ω to represent the probability of each SNP being causal; thus, we have $\sigma_\alpha^2 = \omega \cdot \sigma_1^2$, $h^2 = \sigma_\alpha^2/(\sigma_\alpha^2 + \sigma_0^2)$. For this study, we considered $\sigma_0^2 = 0.6$, $\sigma_1^2 = 0.4$, $\omega = 0.5$, $\sigma_\alpha^2 = 0.2$, $h^2 = 0.25$, and $\mu = 0.2$. The total count N_i for each individual i, $1 \leq i \leq n$ was assumed to follow a $\mathrm{Poisson}(N)$ distribution with $N = 10$.

For the number of SNPs, p, we considered p = 500, 1000, 2000, 5000. As for the number of individuals, n, it is different due to the computational limitation of PQLseq, which we compared. For PQLseq, we fixed n/p to be 1 (so for PQLseq, n = 500, 1000, 2000, 5000). For MoL, we considered n/p to be 1,10, and 100 (so for MoL, n = 500, 1000, 2000, 5000; or 5000, 10000, 20000, 50000; or 50000, 100000, 200000, 500000). Note that the sample size for PQLseq was not as large as that for the MoL due to the computational limitation of the former. Also, the PQLseq algorithm may not converge, even if the entries of the genetic matrix are i.i.d.. Thus, we ran 200 simulations and pick the first 100 simulation results that PQLseq algorithm did converge. For the MoL method, to ensure a fair comparison, we chose either the corresponding 100 results, or the initial 100 results if the PQLseq results were not available.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0092

3.1 Unbiasedness of MoL estimators

Table 2 summarizes the MoL estimation results for $\sigma_0, \sigma_\alpha, \omega$. The empirical standard deviation (emp.s.d.), computed from all simulations, and the estimated standard error (est.s.d.), derived from the asymptotic theory, are also presented, along with the percentage of simulation runs where the true value of the parameter fell within the 95% confidence interval (CI) from the MoL method. By comparing the mean from the 100 simulations to the true value, we can conclude that the estimations for $\sigma_0, \sigma_\alpha, \omega$ are approximately unbiased. Furthermore, the accuracy of the estimation results improved when n/p increased, or when n/p was held constant while n and p increased. Additionally, by fixing $p = 500$, the effect of increasing n (with values $n = 500, 5000, 50000$) shows that larger n improves parameter estimation for the same p. Similarly, by fixing $n = 5000$ and varying p (with values $p = 500, 5000$), it is evident that increasing p makes the estimation more challenging for the same n. These conclusions are consistent across different values of p and n.

3.2 Asymptotic normality of MoL estimators

The MoL estimators for $\sigma_0^2, \sigma_\alpha^2$ were then standardized by subtracting the true variances and dividing the differences by the corresponding estimated standard deviations obtained from the asymptotic theory. This standardized value

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0092

Table 2: **MoL Estimation (Mean, S.D. and C.I.) for** $\sigma_0, \sigma_\alpha, \omega$

True Parameter		sigma02 = 0.6				sigmaa2 = 0.2				omega = 0.5	
Performance Measure		Mean	emp.s.d.	est.s.d.	95% CI	Mean	emp.s.d.	est.s.d	95% CI	Mean	emp.s.d.
$n/p = 1$	p=500, n=500	0.603	0.178	0.180	0.950	0.183	0.163	0.072	0.590	0.138	0.271
	p=1000, n=1000	0.596	0.170	0.127	0.890	0.193	0.128	0.051	0.490	0.175	0.315
	p=2000, n=2000	0.586	0.102	0.090	0.940	0.199	0.094	0.036	0.550	0.150	0.290
	p=5000, n=5000	0.591	0.066	0.057	0.890	0.202	0.058	0.023	0.520	0.309	0.325
$n/p = 10$	p=500, n=5000	0.598	0.058	0.057	0.980	0.200	0.033	0.030	0.910	0.523	0.170
	p=1000, n=10000	0.601	0.045	0.040	0.920	0.200	0.023	0.021	0.920	0.504	0.102
	p=2000, n=20000	0.596	0.028	0.028	0.930	0.199	0.018	0.015	0.850	0.511	0.074
	p=5000, n=50000	0.599	0.017	0.018	0.970	0.201	0.012	0.009	0.870	0.509	0.045
$n/p = 100$	p=500, n=50000	0.602	0.016	0.018	0.990	0.198	0.019	0.021	0.960	0.511	0.050
	p=1000, n=100000	0.599	0.012	0.013	0.970	0.198	0.013	0.015	0.940	0.504	0.040
	p=2000, n=200000	0.598	0.008	0.009	0.960	0.200	0.010	0.011	0.950	0.505	0.030
	p=5000, n=500000	0.600	0.005	0.006	0.980	0.200	0.006	0.007	0.950	0.501	0.020

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0092

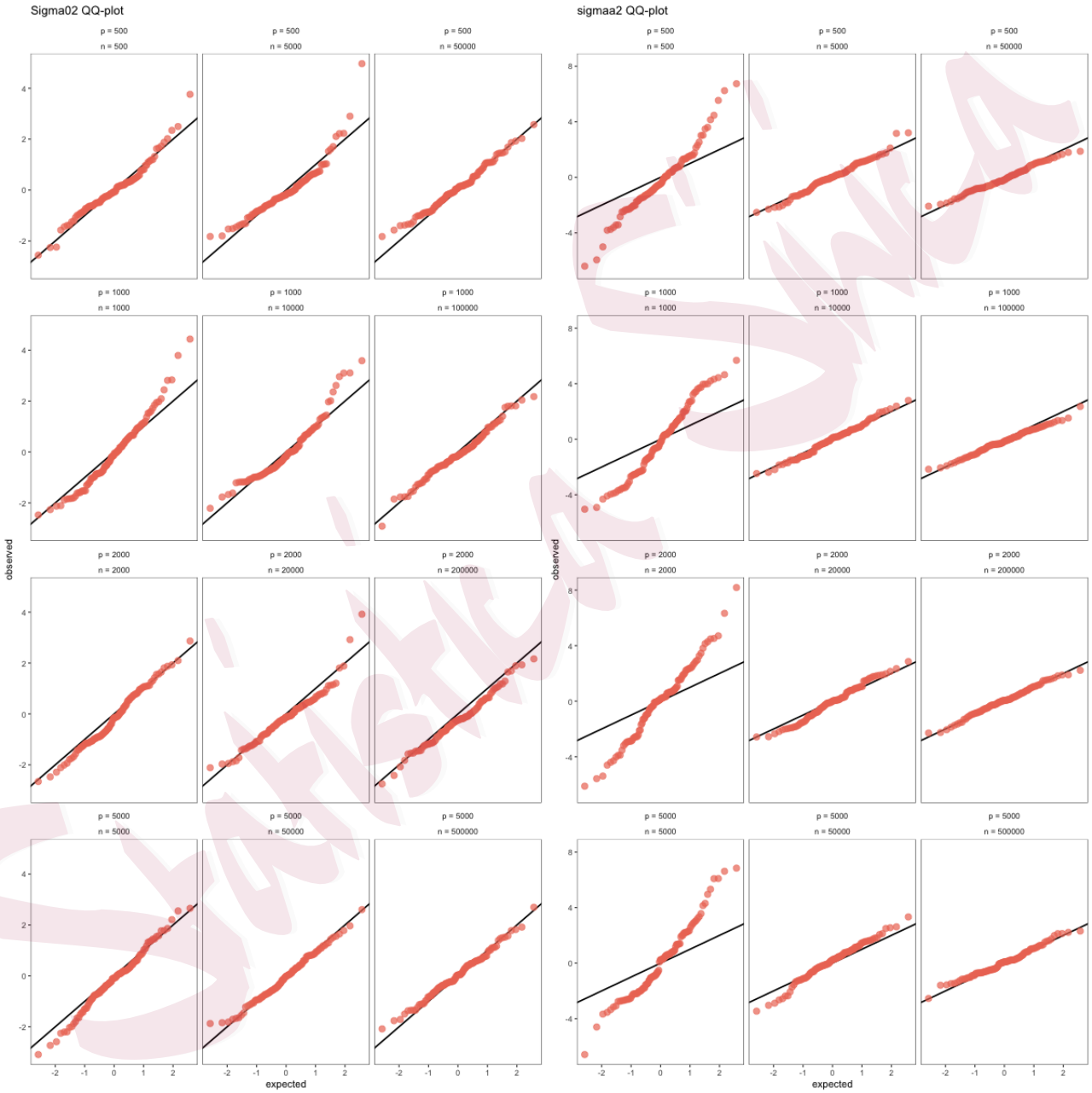
was compared to the standard Gaussian distribution, as depicted in the QQ-Plot (Figure 1). We can see that as n/p increased, or when n/p was maintained constant while n and p increased, the distribution of the standardized MoL estimator converged towards the standard normal distribution. Additionally, by fixing $p = 500$, the effect of increasing n (with values $n = 500, 5000, 50000$) shows that larger n leads the distribution of the standardized estimators to converge more closely towards the standard normal distribution for the same p. Similarly, by fixing $n = 5000$ and varying p (with values $p = 500, 5000$), it is evident that increasing p makes the convergence towards the standard normal distribution more challenging for the same n. These conclusions are consistent across different values of p and n.

3.3 Heritability estimation: Comparing MoL and PQLseq

The efficacy of the MoL and PQLseq methods were evaluated and compared via three performance metrics: 'var_ratio', which assesses the variability of the MoL estimators relative to the PQL estimators; 'time_ratio', which evaluates the computational expense of the MoL estimators compared to the PQL estimators; and RASE, defined via (1.5), which combines the statistical and computational efficiencies into a single measure; see discussion about the RASE metric in the introduction. If var_ratio or time_ratio fall below 1, MoL is considered outper-

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0092

Figure 1: **Asymptotic Distribution of MoL Estimators for** $\sigma_0^2, \sigma_\alpha^2$



Statistica Sinica: Preprint
doi:10.5705/ss.202024.0092

forming PQLseq under the corresponding metric; if RASE is above 1, MoL is considered outperforming PQLseq overall. The results are shown in Table 3.

Table 3: Comparison with PQLseq: Variance/Time Ratios (MoL/PQLseq) and RASE

p	PQL_n (PQL_n / p)	MoL_n (MoL_n / p)	var_ratio	time_ratio	RASE
500	500 (1)	500 (1)	5.272	0.013	14.488
1000	1000 (1)	1000 (1)	7.656	0.018	7.245
2000	2000 (1)	2000 (1)	7.832	0.012	10.262
5000	5000 (1)	5000 (1)	9.070	0.005	24.089
500	500 (1)	5000 (10)	0.194	0.103	50.008
1000	1000 (1)	10000 (10)	0.221	0.171	26.451
2000	2000 (1)	20000 (10)	0.248	0.119	33.978
5000	5000 (1)	50000 (10)	0.291	0.003	1131.952
500	500 (1)	50000 (100)	0.041	0.054	456.694
1000	1000 (1)	100000 (100)	0.044	0.104	217.962
2000	2000 (1)	200000 (100)	0.055	0.098	187.166
5000	5000 (1)	500000 (100)	0.066	0.047	322.916

As presented in Table 3, the variance of PQLseq is smaller than that of MoL when the sample size are identical. However, when the sample size for MoL is increased while maintaining PQLseq's sample size due to the computational constraints, MoL exhibits a smaller variance than PQLseq. Moreover, MoL demonstrates a significant computational advantage over PQLseq even when the sample

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0092

size for MoL is 100 times of that for PQLseq. MoL also consistently displays a higher RASE than PQLseq, suggesting that MoL integrates the statistical and computational efficiencies more effectively than PQLseq. These findings highlight the advantages of MoL under a notion of modern data science, in which statistical performance and computational efficiency are considered jointly. The computational superiority of MoL not only accelerates the estimation process but also accommodates a larger sample size when (much) more data are available, thereby improving the estimation accuracy.

4. Real data analysis

In this study, we applied our proposed MoL method to the UK Biobank dataset (Sudlow *et al.* (2015)), a large cohort study aimed at understanding the causes of complex traits. We focus on estimating heritability using both MoL and PQLseq. The traits of interest include weekly champagne and weekly red wine consumption habits, with the unit of measure being glasses. After excluding individuals with null, negative and zero phenotype values and non-European ancestries, there were 111,351 and 133,610 individuals with both genotypes and phenotypes, respectively. Our initial step involved conducting a GWAS to derive marginal association P-values for each SNP. SNPs were then filtered with a threshold of 0.001 to obtain significant ones. After that, independent SNPs

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0092

were randomly selected through a pruning process with a window size of 100kb, resulting in 132 SNPs in weekly champagne and 165 SNPs in weekly red wine. The column n_SNPs in Table 4 represents the number of SNPs used for the heritability estimation. The MoL estimation results, obtained in a minute for each trait, are presented in Table 4. These results suggest potential genetic influences on the behavior of weekly champagne and red wine consumptions.

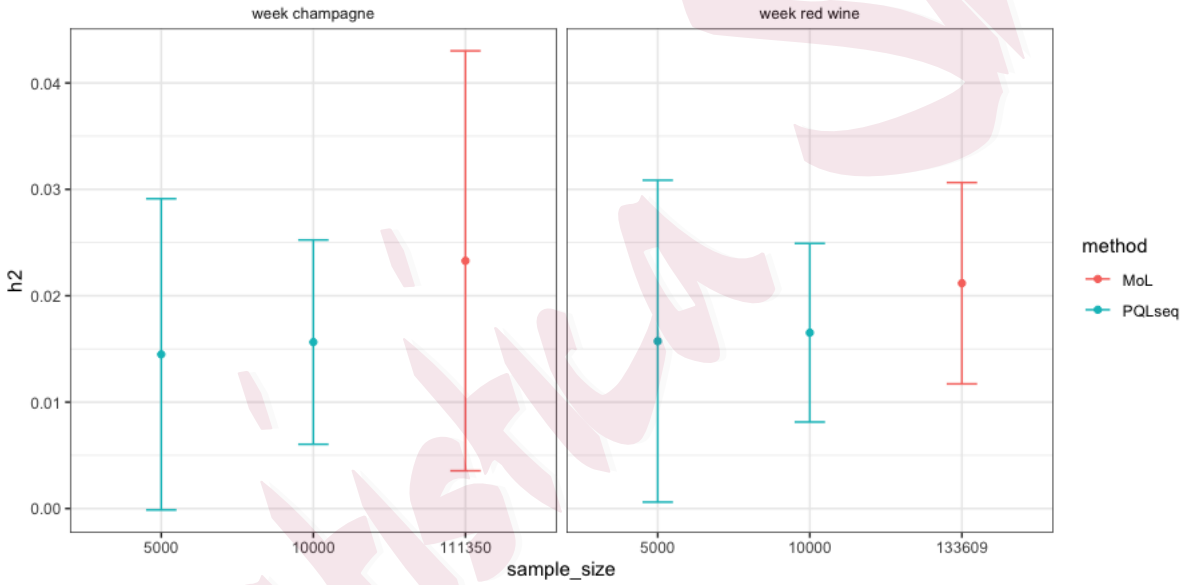
Table 4: MoL Real Data Results

trait	n_SNPs	σ_0^2	σ_α^2	$\mathrm{var}(\sigma_0^2)$	$\mathrm{var}(\sigma_\alpha^2)$	h^2	$\mathrm{var}(h^2)$	ω
week champagne	132	0.700	0.017	9.56e-05	5.10e-05	0.023	1.01e-04	0.122
week red wine	165	0.558	0.012	2.97e-05	5.63e-06	0.021	2.33e-05	0.471

Heritability estimations for these two traits were also generated using the PQLseq method. Due to the computational limitations, PQLseq could not utilize the entire dataset for estimation. Consequently, subsets of 5,000 and 10,000 samples were randomly drawn from the whole dataset 100 times to estimate heritability for each sub-dataset, requiring an average of 5,240 and 37,253 seconds respectively. The means and 1.96 times the standard errors for the 5,000 and 10,000 samples are plotted in Figure 2 based on all sub-dataset estimations from PQLseq, and these were compared to the 95% confidence interval derived from MoL, as depicted in Figure 2.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0092

Figure 2: **Real Data Analysis for MoL and PQLseq**



Statistica Sinica: Preprint
doi:10.5705/ss.202024.0092

As presented in Figure 2, the PQLseq and MoL estimation results are relatively comparable, although MoL estimates of the heritability are higher than those of PQLseq. The increase in the values of PQLseq estimates with rising sample size might account for the observed discrepancy between the two methods. Namely, the observed trends seem to suggest that, if it were possible to compute the PQLseq estimates using the full data, one would obtain something even closer to the MoL estimates. Given that MoL utilizes a sample size 10 times larger than that for PQLseq, it is plausible that the estimator based on averaging the smaller sub-dataset results may be biased compared to the estimator based on the entire dataset, if computation of the latter were possible.

5. Discussion

The increasing number of large datasets necessitates the development of computationally feasible statistical methods. In this study, we presented the method of limits (MoL) as an appealing alternative to the traditional likelihood-based estimation techniques, especially in the context of GWAS with count data. Unlike the traditional approaches that often treat count data as continuous for heritability estimation, we employ the Poisson model, ensuring a more accurate representation of count data structures and subsequently enhancing statistical efficiency.

Our primary theoretical contribution is the establishment of consistency and

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0092

asymptotic normality for the MoL estimators when estimating the heritability of count phenotypes. Unlike the commonly used PQLseq method which requires extensive iterations for convergence, the MoL approach overcomes computational challenges in large datasets by offering closed-form solutions. These solutions not only simplify the estimation process but also pave the way for statistical inference regarding the heritability. Through extensive simulations, we validated the theoretical properties of MoL in estimating the heritability. Based on real data analysis, we conclude that computational efficiency not only accelerates the estimation process, but also facilitates the utilization of the entire large genetic datasets during the estimation, thus enhancing statistical efficiency.

Furthermore, our work provides a consistent estimator for the proportion of causal SNPs, an essential component in understanding the genetic structure that has not been addressed in previous studies. Both simulation and empirical data analysis validate the estimator. This finding extends the understanding of GWAS and may foster further exploration in genetic modeling and estimation.

We have also introduced new evaluation metrics, namely the average statistical efficiency (ASE) and relative ASE (RASE). These metrics incorporate a novel concept of combining two efficiency measures simultaneously, offering a comprehensive evaluation, and comparison, of different methods.

However, our study has limitations. The method of limits assumes the in-

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0092

dependence of genetic variants, limiting the number of genetic variants that one can utilize. Future studies are needed to extend the method by taking correlated genetic variants into consideration. The method also requires analytic skills to (correctly) derive the limits of certain base statistics, which in some cases could be challenging to a practitioner.

In conclusion, our paper introduced the MoL, established its theoretical properties and demonstrated its real data applications. Our findings highlight its advantages in computational efficiency and how this can lead to improved statistical efficiency in large datasets. Moreover, the introduction of ASE and RASE evaluation metrics enables a unified approach to assess both computational and statistical efficiency and compare different methods.

Supplementary Materials

The Supplementary Material contains proofs of the main theoretical results.

The code for simulations and real data analysis is available at `https://github.com/LeqiXu/MoL_analysis`.

Acknowledgements

The research of Jiming Jiang is partially supported by the NSF grants DMS-1713120, DMS-1914465 and DMS-2210569. The research of Hongyu Zhao is partially supported by DMS 1713120 and NIH R01 GM134005. The research was conducted using the UKBB resource under approved data requests (access

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0092

ref: 29900).

References

Booth, J. G. and Hobert, J. P. (1999), Maximum generalized linear mixed model likelihood with an automated Monte Carlo EM algorithm, *J. Roy. Statist. Soc. B* 61, 265–285.

Breslow, N. E. and Clayton, D. G. (1993), Approximate inference in generalized linear mixed models, *J. Amer. Statist. Assoc.* 88, 9–25.

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.

Dao, C., Jiang, J., Paul, D., and Zhao, H. (2021), Variance estimation and confidence intervals from high-dimensional genome-wide association studies through misspecified mixed model analysis, *J. Stat. Plan. Inference* 220, 15–23.

Golan, D., Lander, E. S., and Rosset, S. (2014), Measuring missing heritability: Inferring the contribution of common variants, *PNAS* 111, E5272–E5281.

Jiang, J. (1998), Consistent estimators in generalized linear mixed models, *J. Amer. Statist. Assoc.* 93, 720–729.

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

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

Jiang, J., Li, C., Paul, D., Yang, C., and Zhao, H. (2016), On high-dimensional misspecified mixed model

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0092

analysis in genome-wide association study, *Ann. Statist.* 44, 2127–2160.

Little, R. J. A. and Rubin, D. B. (2002), *Statistical Analysis with Missing Data*, 2nd ed., Wiley, New York.

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.

Sun, S., Zhu, J., Mozaffari, S., Ober, C., Chen, M., and Zhou, X. (2019), Heritability estimation and differential analysis of count data with generalized linear mixed models in genomic sequencing studies, *Bioinformatics* 35, 487–496.

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.