Received: 1 March 2021 | Revised: 11 June 2021 | Accepted: 4 October 2021

DOI: 10.1002/gepi.22432

RESEARCH ARTICLE

Genetic Epidemiology

OFFICIAL JOURNAL INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY www.geneticepi.org

WILEY

REHE: Fast variance components estimation for linear mixed models

Kun Yue[1] | **Jing Ma**[2] | **Timothy Thornton**[1] | **Ali Shojaie**[1]

[1]Department of Biostatistics, University of Washington, Seattle, Washington, USA

[2]Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, Washington, USA

Correspondence
Ali Shojaie, 334 Hans Rosling Center for Population Health, Department of Biostatistics, University of Washington, Seattle, WA 98195, USA.
Email: ashojaie@uw.edu

Funding information
National Institutes of Health, Grant/Award Numbers: R01-GM114029, R01-GM133848, R01-HL141989

Abstract

Linear mixed models are widely used in ecological and biological applications, especially in genetic studies. Reliable estimation of variance components is crucial for using linear mixed models. However, standard methods, such as the restricted maximum likelihood (REML), are computationally inefficient in large samples and may be unstable with small samples. Other commonly used methods, such as the Haseman–Elston (HE) regression, may yield negative estimates of variances. Utilizing regularized estimation strategies, we propose the restricted Haseman–Elston (REHE) regression and REHE with resampling (reREHE) estimators, along with an inference framework for REHE, as fast and robust alternatives that provide nonnegative estimates with comparable accuracy to REML. The merits of REHE are illustrated using real data and benchmark simulation studies.

KEYWORDS

genome-wide association study, heritability study, linear mixed model, restricted Haseman–Elston regression, variance component

1 | INTRODUCTION

Linear mixed model are a convenient and powerful tool for analyzing correlated data, with a wide range of applications in scientific research. It is especially useful for genetic studies of complex traits, including heritability estimation (Sofer, 2017), genome-wide association studies (GWAS) (Aulchenko et al., 2007), and network-based pathway enrichment analysis (NetGSA) (Shojaie & Michailidis, 2009). Variance components estimation is an essential step when applying linear mixed models and the restricted maximum likelihood (REML) approach is considered to be the gold-standard for this task (Patterson & Thompson, 1971). REML works by iteratively maximizing the residual likelihood with respect to the variance component parameters. During each iteration, REML computes the inverse of two $n \times n$ matrices, where n is the sample size of the data set. As a result, the computation for REML quickly becomes prohibitive for large sample sizes, especially when the correlations among observations are nonsparse—such as between-subject correlations due to genetic relatedness (Kang et al., 2010). Despite efforts to improve the computational efficiency of REML, such as average information REML (Gilmour et al., 1995), Monte Carlo REML (Matilainen et al., 2013), and REML based on grid search (Jiang et al., 2019), the scalability of REML to very large data sets is often limited in many applications. On the other hand, consistency and asymptotic normality of REML estimates are large sample properties. In this paper we illustrate that REML can be numerically unstable and can also provide unreliable estimates and/or confidence intervals of variance components in some settings (Section 4, Supplementary Note 2.3 and Note 2.4). Given these shortcomings, new approaches for fast and reliable estimators of variance components for linear mixed model are needed.

When computational efficiency is a primary concern, moment estimators of variance components have frequently been used as alternatives to REML. These include analysis of variance (ANOVA), minimum norm quadratic unbiased estimation (MINQUE), and the Haseman–Elston (HE) regression estimator (Haseman & Elston, 1972; Rao, 1970; Rasch & Masata, 2006; Sofer, 2017). These methods bypass the most time-consuming step in REML—the inversion of $n \times n$ matrices. An essential component of of these alternative methods to REML is to set up estimating equations by equating the mean squared errors to its expectation, the error variance. ANOVA, originated from ideas by R. A. Fisher in 1920s, has been well established for estimating variance components. The resulting estimators are minimum variance quadratic unbiased (Graybill & Hultquist, 1961), and minimum variance unbiased under normality assumptions on the random effects and the errors (Graybill, 1954; Graybill & Wortham, 1956). MINQUE (Rao, 1970), which can be viewed as an extension of the ANOVA method, is equivalent to the first iteration of REML (Searle, 1995). It relaxes the assumption of normality using estimating equations that rely on initial values for the variance components. The HE estimator, first introduced in Haseman and Elston (1972), has been recently used for linear mixed model variance component estimation in genetic studies (Sofer, 2017; Zhou, 2017). Its simple implementation and fast computation make it favorable when working with large and densely correlated data sets. A key limitation of these moment estimators, however, is that they do not guarantee nonnegative estimates for the variance components. This leads to difficulties in both interpretation and downstream analyses.

To address the shortcomings of existing approaches in variance component estimation, we propose a new estimation method based on restricted Haseman–Elston (REHE) regression. REHE is computationally efficient, and ensures nonnegative estimates of variance components. We demonstrate that the REHE estimates are comparably accurate to the REML estimates when REML performs well, and are robust in the settings where REML does not provide good estimates. To accommodate the need for a strictly positive variance estimates in some applications, we also propose REHE with resampling (reREHE), which provides positive variance components estimates with high probability. Furthermore, to facilitate inference, we propose bootstrap confidence intervals for REHE estimates. We demonstrate that the REHE boostrap confidence intervals are more robust than their REML counterparts. Finally, we also show that REHE in combination with correlation matrix sparsification (Gogarten et al., 2019), when applicable, can result in a substantial increase in computational speed for variance component inference. REHE is both computationally efficient and flexible, and can be used across a broad range of study designs. Here we demonstrate the utility of REHE in three different contexts: heritability estimation, GWAS and NetGSA. We benchmark the proposed methods' performance with simulation studies, and illustrate their advantages using data from the Hispanic Community Health Study/Study of Latinos (HCHS/SOL) (Conomos et al., 2016; Sorlie et al., 2010), where there is extensive nonzero genetic correlations among the 12,803 study subjects with GWAS data available, and The Cancer Genome Atlas (TCGA) breast cancer data set (TCGA, 2012).

The rest of the paper is organized as follows. In Section 2, we introduce the REHE estimator, discuss its properties and propose a bootstrap inference framework. We also introduce the reREHE estimator as an alternative to the REHE estimator. We demonstrate the performance of the REHE and reREHE estimators with real data applications in Section 3. In Section 4, we benchmark their performance with extensive simulation studies. Section 5 concludes the paper with discussions on the results and potential improvements. Additional results on REHE, reREHE, and matrix sparsification are provided in the Supplementary Notes.

2 | METHODS

Consider a generic linear mixed model for an outcome vector Y of length n:

$$Y = X\boldsymbol{\beta} + \sum_{k=1}^{K} \sigma_k \boldsymbol{\gamma}_{\mathbf{k}} + \sigma_0 \boldsymbol{\epsilon}. \tag{1}$$

Here, X is an $n \times p$ design matrix for p covariates, and $\boldsymbol{\beta}$ is a p-dimensional fixed effect coefficient vector. For $k = 1,2, ..., K$, $\boldsymbol{\gamma}_{\mathbf{k}}$ is a length n vector of random effects following $N_n(0, D_k)$, where each $n \times n$ matrix D_k defines one source of correlation among the observations, and is assumed to be known. The noise $\boldsymbol{\epsilon}$ is a length n vector following $N_n(0, \mathbf{I}_n)$. The parameters σ_k's $(k = 0,1, ..., K)$ are the variance components. For $k = 1, ..., K$, $h_k = \sigma_k^2 / \sum_{l=0}^{K} \sigma_l^2$ estimates the proportion of variation explained by D_k. In the context of genetic studies, a genetic relatedness matrix (GRM) is often used for D_1 and h_1 is viewed as a measure of heritability, that is, the proportion of the total trait variation that is due to genetic variation.

Our main objective is to estimate the variance components $\sigma_k^2 (k = 0,1, ..., K)$. For expositional clarity, we assume the model has no fixed effect, and the outcome vector Y is centered. We also assume $K = 1$ such that the correlations among the observations can be modeled by a single random effect $\boldsymbol{\gamma}_1$. However, our methods can be easily extended to models with fixed effects (Section 2.2.2), or with more than one random effects. Denoting $D_0 = \mathbf{I}_n$ and $\boldsymbol{\gamma}_0 = \boldsymbol{\epsilon}$, the simplified form of model [1] becomes

$$Y = \sigma_0 \boldsymbol{\gamma_0} + \sigma_1 \boldsymbol{\gamma_1}. \quad (2)$$

2.1 | The HE regression

The HE regression approach (Haseman & Elston, 1972; Sofer, 2017; Zhou, 2017) estimates the variance components via the method of moments. Specifically, since model [2] implies that $Var(Y) = \sigma_0^2 D_0 + \sigma_1^2 D_1$, n^2 estimating equations are constructed:

$$E[Y_i Y_j] = \sigma_0^2 D_0^{ij} + \sigma_1^2 D_1^{ij}, i, j = 1, 2, \ldots n,$$

where D_k^{ij} denotes the (i, j) entry of matrix $D_k (k = 0,1)$. Estimation of the variance components can thus be recast as a linear regression problem. Let $Y = \mathrm{vec}(YY^\top)$ denote the vectorization of the $n \times n$ matrix $YY^\top$ by stacking its columns, $X = (vec(D_0), vec(D_1))$ and $\boldsymbol{\sigma^2} = (\sigma_0^2, \sigma_1^2)^\top$. We then have $E(Y) = X\boldsymbol{\sigma^2}$. HE solves for variance components $\boldsymbol{\sigma^2}$ by linear regression, specifically, by minimizing the residual sum of squares

$$(\tilde{Y} - \tilde{X}\boldsymbol{\sigma}^2)^T (\tilde{Y} - \tilde{X}\boldsymbol{\sigma}^2). \quad (3)$$

The resulting estimator has a closed form expression $\boldsymbol{\sigma}^2 = (X^\top X)^{-1} X^\top Y$. A nice property of the HE estimator is unbiasedness, even if we only use a subset of the n^2 estimating equations to compute the estimates; see Appendix A for details.

The computational complexity of HE is $O(Kn^2)$ compared with $O(n^3)$ for REML. When the sample size n is large and the number of variance components K is small, as is typically the case in practice, HE offers substantial improvement in computation over REML. However, the ordinary least squares solution for variance components by HE is not guaranteed to be nonnegative, leading to difficulties in downstream analyses and interpretation. In practice, negative estimates from HE are often truncated at zero; yet such naive truncation does not minimize the residual sum of squares in [3] within the parameter space $(\sigma_k^2 \geq 0, k = 0,1)$. In addition, as will be illustrated in simulation studies in Section 4 and the Supplementary Note 2.3, naive truncation-based HE estimates generally have larger mean square error than estimates by both REML and our proposed REHE method, which we introduce in the next subsection.

2.2 | The REHE regression

To prevent negative estimation of variance components by HE, while still preserving its computational efficiency, we propose a new variance components estimation method, termed the REHE regression. Similar to HE, REHE is a moment estimator which regresses the empirical covariance of the observations on pre-specified correlation matrices that encode sample relatedness. However, instead of the ordinary least squares estimate by HE, REHE finds the nonnegative minimizer of the residual sum of squares, ensuring sensible estimation of variance components (Figure S10). Following [3], the REHE estimates of the variance components are expressed as

$$\left(\sigma_0^2, \sigma_1^2\right) = \arg \min_{\left(\sigma_0^2 \geq 0, \sigma_1^2 \geq 0\right)} \sum_{l=1}^{n^2} \left(Y_l - X_{l1}\sigma_0^2 - X_{l2}\sigma_1^2\right)^2. \quad (4)$$

There is a closed form solution to [4] with only two variance components (Supplementary Note 1.1). With more than two variance components, iterative algorithms for nonnegative least squares (NNLS) can easily solve [4] (Franc et al., 2005; Goldfarb & Idnani, 1983; Kim et al., 2006; Lawson & Hanson, 1995). The convexity of [4] guarantees the numerical solutions of different solvers converge to the same global minimizer. Using the R package quadprog (v1.5-7, Berwin & Turlach 2019), REHE estimation has approximately the same computational cost as HE, and is thus substantially faster than REML. In addition, the REHE estimator is consistent under mild conditions, and asymptotically normal when the correlation matrix for the random effect is sparse and block-diagonal, such as a sparse empirical kinship matrix; see Appendix B for details.

2.2.1 | REHE with resampling (reREHE)

Estimates obtained from REHE are nonnegative. However, in some applications, a zero estimate for the variance component may still make the interpretation and/or subsequent analyses challenging. To address this issue, we equip REHE with a resampling procedure that provides strictly positive variance component estimates with high probability. The resampling procedure utilizes repeated subsamples, which can further improve the computational efficiency of REHE. The resulting approach is termed reREHE.

The idea of subsampling for REHE is simple: instead of estimating the variance components based on all the observations at the same time, we only use a small subsample of the data. The full-sample-based estimates and inference are usually well approximated by statistics based on subsamples (Politis et al., 1999). Similar subsampling techniques are also extensively used in stochastic gradient descent methods (Metel, 2017). Recently, subsampling has also been

used with HE-based estimating equations (Zhou, 2017). Our proposed reREHE procedure described in Algorithm [1] is unique in that it subsamples repeatedly to obtain the estimates. Although at a cost of reduced computational efficiency compared with using a single subsample, this resampling offers considerable advantages.

By averaging the estimates from repeated subsamples, the reREHE estimates have much higher accuracy compared with estimates based on a single subsample. At the same time, we obtain strictly positive estimates, unless in extremely rare cases when all subsamples yield zero estimates. Other summaries, such as median, can also be used to summarize estimates from repeated subsamples. When the sampling rate r_s and the number of subsamples B satisfy $r_s^2 B < 1$, reREHE achieves higher computational efficiency than REHE. On the other hand, choosing larger r_s and B results in more stable results. In the simulation studies and data applications, we chose $B = 50$ and varied r_s within (0.05,0.1) to achieve a balance between accuracy and computational efficiency.

Algorithm 1 reREHE Approach Estimation.

for $b = 1$ to B **do**

(a) Sample with replacement from Y to obtain a length $[nr_s]$ vector $Y^{(b)}$ with sampling rate r_s, where $[x]$ is a function to round x to the nearest integer; subset the correlation matrices accordingly as $D_k^{(b)}$, for $k = 0,1$.

(b) Compute the variance component estimates $(\sigma_{0,re}^{2(b)}, \sigma_{1,re}^{2(b)})$ based on $Y^{(b)}$ and $D_k^{(b)}$'s using REHE [4].

end for

Estimate the variance components as $\sigma_{k,re}^2 = \frac{1}{B}\sum_{b=1}^{B} \sigma_{k,re}^{2(b)}$, $k = 0,1$.

2.2.2 | REHE with fixed effects

The REHE estimation procedure can be easily modified to accommodate fixed effects. Consider the full model [1] where X is the design matrix with p covariates and $\boldsymbol{\beta}$ is the fixed effect coefficient vector. Let $P_X^{\perp} = I_n - X(X^\top X)^{-1}X^\top$ denote the projection matrix onto the orthogonal complement of the column space of X. We project the outcome Y and the random effects $\boldsymbol{\gamma_k}$'s (including the noise term $\boldsymbol{\gamma_0}$) as

$$Y^\dagger = P_X^{\perp} Y, \boldsymbol{\gamma_k}^\dagger = P_X^{\perp}\boldsymbol{\gamma_k}.$$

Recall that each random effect $\boldsymbol{\gamma_k}$ follows a normal distribution with zero mean and covariance D_k. Writing $D_k^\dagger = P_X^{\perp} D_k P_X^{\perp}$, model [1] becomes

$$Y^\dagger = \sum_{k=0}^{K} \sigma_k \boldsymbol{\gamma_k}^\dagger, \boldsymbol{\gamma_k}^\dagger \sim N_n\left(0, D_k^\dagger\right), k = 0, ..., K. \quad (5)$$

With model [5], we can directly apply the REHE approach as introduced in Section 2.2 to estimate the variance components. When the sample size n is large, computing the projected correlation matrices $D_k^\dagger$ is time-consuming. In genetic and genomics applications, the number of fixed effect covariates p is much smaller than the sample size n. We also have balanced design in many of these applications. In those settings we are able to obtain good estimates of the fixed effects, and we can directly use the original correlation matrices D_k instead of projected matrices $D_k^\dagger$. In such cases, as in our data applications and simulation studies, we expect results based on $D_k^\dagger$ to be very close to those based on D_k. We thus suggest using D_k for computational efficiency when estimating model [5] via REHE or reREHE.

If the fixed effect coefficients $\boldsymbol{\beta}$ are of interest, they can be estimated using ordinary least squares as $\boldsymbol{\beta} = (X^\top X)^{-1}X^\top Y$, or weighted least squares as, $\hat{\boldsymbol{\beta}} = (X^T\hat{\Sigma}^{-1}X)^{-1}X^T\hat{\Sigma}^{-1}Y$ where $\Sigma = \sum_{k=0}^{K}\sigma_k^2 D_k$ is based on previously estimated variance components σ_k^{2}'s. While the resulting $\boldsymbol{\beta}$ is consistent for $\boldsymbol{\beta}$, one can iteratively update σ_k^{2}'s and $\boldsymbol{\beta}$ as in (Sofer, 2017).

2.2.3 | Constructing confidence intervals with REHE

To obtain confidence intervals for variance component estimates, we use a parametric bootstrap procedure as summarized in Algorithm [2].

Algorithm 2 Parametric Bootstrap Confidence Interval Construction for REHE.

Compute REHE estimates σ_0^2, σ_1^2 from [4], based on Y, D_0, D_1;

for $b = 1$ to B **do**

(a) Generate outcome vector $Y^{*(b)}$ from $N_n(0, \sigma_0^2 D_0 + \sigma_1^2 D_1)$;

(b) Compute REHE estimates ${\sigma_0}^{2(b)}, {\sigma_1}^{2(b)}$ from [4], based on $Y^{*(b)}$, D_0, D_1;

end for

Using the bootstrap samples of REHE estimates, $(\sigma_k^{2(b)})_{b=1}^{B}$, for $k = 0,1$, we can construct Wald-type confidence intervals as

$$\left[\sigma_k^2 - z_{\alpha/2} \times \mathrm{SD}\left(\sigma_k^{2(b)}\right), \sigma_k^2 + z_{\alpha/2} \times \mathrm{SD}\left(\sigma_k^{2(b)}\right)\right],$$

$$k = 0, 1,$$

where $z_{\alpha/2}$ is the $(1-\alpha/2)\times 100$th percentile of the standard normal distribution, and SD$(\cdot)$ denotes the sample standard deviation over bootstrap samples. Wald-type confidence intervals are valid, provided that the estimates are normally distributed. We can also construct quantile bootstrap confidence intervals as

$$\left[\sigma_k^2-\left(\sigma_k^{2(b)}-\sigma_k^2\right)_{1-\alpha},\ \sigma_k^2-\left(\sigma_k^{2(b)}-\sigma_k^2\right)_{\alpha}\right],\ k=0,1,$$

where $(\sigma_k^{2(b)}-\sigma_k^2)_\alpha$ is the $(1-\alpha)\times 100$ th empirical quantile of $(\sigma_k^{2(b)}-\sigma_k^2)_{b=1}^B$.

For small sample sizes, the quantile confidence intervals are expected to be more robust than their Wald-type counterparts. When the REHE estimator is close to be normally distributed under large sample size, Wald-type confidence interval might have higher accuracy based on the same number of bootstrap samples. In simulation studies and real data applications, we chose the number of bootstrap samples $B=50$ to balance computational time and confidence interval accuracy. Confidence intervals for functions of variance components, such as heritability, can be similarly obtained by transforming the bootstrap samples accordingly.

When the correlation matrix D_1 is dense and the sample size n is large, it can be computationally prohibitive to compute a matrix decomposition (through Cholesky or singular value decomposition) of D_1, which is required for the sampling step (a) in the Algorithm [2]. To speed up the inference procedure in this case, we propose using correlation matrix sparsification (Gogarten et al., 2019). The sparsification approximates the dense correlation matrix D_1 by a sparse block-diagonal matrix D_1^*, and largely accelerates matrix decomposition. Such a sparse block-diagonal structure is often approximately satisfied in genetic studies: for example, subjects are highly genetically correlated within the same family, and are remotely correlated across families (see Figure S12 for an example of the kinship matrix). Note that a standard GRM calculated from genome-screen data may not be sparse. However, an empirical kinship matrix that reflects close familial relationship (and that can be sparsified) along with principal components for population structure has been proposed as an alternative to using a single GRM in linear mixed models (Gogarten et al., 2019; Hu et al., 2021). This approach has been used in various genetic studies with relatedness and population structure, including the HCHS/SOL (Conomos et al., 2016), the Population Architecture using Genomics and Epidemiology (PAGE) study (Wojcik et al., 2019), and a recent analysis for the Trans-Omics for Precision Medicine (TOPMed) program (Hu et al., 2021). Gogarten et al. (2019) provide a detailed comparison of using a linear mixed model with a GRM, a dense kinship matrix (with PCs), and a sparse kinship matrix (with PCs). The genetic association testing results were shown to be very similar for all three approaches, but with a substantial increase in computational efficiency by using a sparse empirical kinship matrix. In the examples presented in Section 3, we use a dense empirical kinship matrix and PCs to model the correlation structure in the sample. We investigate the performance of sparsification in the additional simulation studies in Supplementary Note 2.1 and Note 2.3. Our application of matrix sparsification for inference of variance components in genetic studies is novel, and is discussed in detail in Supplementary Note 1.2.

3 | APPLICATIONS

3.1 | GWAS and heritability studies with HCHS/SOL data

To evaluate the performance of REHE and reREHE in genetics applications, we conducted a genome-wide association analysis as well as a heritability analysis using a publicly available data set from the HCHS/SOL (Conomos et al., 2016; Sorlie et al. 2010). The HCHS/SOL sample survey design consisted of a two-stage probability sample of households at four recruitment centers. For each center, census block groups were selected in defined communities, and households were sampled within census block groups (Conomos et al., 2016; LaVange et al., 2010). Among a total of 16,415 subjects enrolled in the study at baseline, 12,803 subjects were genotyped on a genome-wide single-nucleotide polymorphisms (SNP) array containing 4,100,028 SNPs.

To perform a genome-wide association analysis, we tested the association between each SNP and the red blood cell count using a linear mixed model (Conomos et al., 2016). We first fit a null model, which was a linear mixed model without any genotype effect (Aulchenko et al., 2007). We included fixed effects covariates age, gender, cigarette use, field center indicator, genetic subgroup indicator, the first five principal components for population stratification effect, and individual sampling weights (Conomos et al., 2016). We removed subjects that have missing values for the above covariates, and included 12,502 subjects in the analysis. Correlations among subjects was modeled by three random effects: genetic relatedness represented by estimated kinship, membership of household, and membership of community group (Conomos et al., 2016).

We separately applied REHE, reREHE and REML to estimate the null model with household membership matrix, community membership matrix, and estimated dense kinship matrix. For reREHE, we chose sampling rate $r_s = 0.1$ and used mean summary function based on 50 repeated subsamples. With n =12,502 subjects to be analyzed, REML took 23.9 min to estimate the null model, while REHE took only 2.4 min, a 10-fold improvement compared with REML. reREHE was similarly fast as REHE. Based on each estimated null model, we applied score tests for the association between each SNP and the red blood cell count (Conomos et al., 2016), and compared the resulting p-values. We focus here on the 164 SNPs with p-values no larger than the family-wise error rate (FWER) threshold of 5×10^{-8} by at least one approach (Dudbridge & Gusnanto, 2008), and presented p-values for all SNPs in Figure S11. As shown in Figure 1a,b, results based on REHE and reREHE have negligible differences from those based on REML. The correlations are all higher than 0.99 for the p values on the $-\log_{10}$ scale between REHE and REML, as well as between reREHE and REML. This concordance among REML, reREHE, and REHE is not surprising as the estimated variance components are similar (Figure 1c).

For the heritability analysis, we used the same data set and fit the same linear mixed null model as in the above genome-wide association analysis. The model was estimated based on REML and REHE separately. We obtained point estimates and confidence intervals for heritability (corresponding to the estimated dense kinship correlation matrix), proportions of variance explained by household membership, community block membership, and noise. REHE took 18.2 min to conduct the inference, compared with 23.9 min by REML. Heritability and variance proportions estimates, as well as the confidence intervals obtained by the REHE approach are all very similar compared with those obtained by REML (Figure 1c).

We used the R package GENESIS (v2.14.3, Conomos et al., 2019) for REML estimation and for conducting the genome-wide association analysis. All analyses were conducted on a computer with 2 × 6-core Intel Xeon CPU E5-2620 @ 2.00 GHz 128GB RAM.

3.2 | Network-based pathway enrichment analysis with breast cancer data

To further demonstrate that REHE and reREHE facilitate downstream analysis with fast variance component estimation, we performed a network-based pathway enrichment analysis, with a breast cancer data set from The Cancer Genome Atlas (TCGA) (TCGA, 2012), preprocessed by Ma et al. (2019). The data set contains RNA-seq measurements for 2598 genes from 100 genetic pathways, with 403 subjects from the ER positive subtype and 117 from the ER negative subtype.

Network-based pathway enrichment analysis tests for differential gene pathways associated with particular phenotypes under different conditions (Ma et al., 2016). It assumes a linear mixed model for the relationship between gene expressions and the phenotype (see Supplementary Note 1.3 and Ma et al., 2016 for details). Here, we compared the activities of 100 genetic pathways between the two ER groups. The ER group-specific gene networks—more specifically, the adjacency and influence matrices supplied to the linear mixed model—were estimated according to Ma et al. (2016). We estimated the variance components using REML, REHE and reREHE. For reREHE, we chose the sampling rate $r_s = 0.1$ for sampling the subjects, and additionally sampled gene entries within each subject with sampling rate 0.5 (see Supplementary Note 1.3 for details). The reREHE estimate was based on the mean of 50 repeated subsamples. After obtaining the variance components estimates, we tested for differences in the activity of each of the 100 genetic pathways (Shojaie & Michailidis, 2009).

We observed substantial improvement in computational efficiency of reREHE and REHE compared with REML: reREHE- and REHE-based analyses both took less than 2 min, whereas analysis with REML took over 1 h. Comparing the resulting p values, REHE and reREHE produce slightly more conservative p values than REML (Figure 2). Moreover, REHE yields a zero estimate for the noise variance component, the reREHE estimate is 0.0120, while the REML estimate is 0.266. The corresponding network variance estimates are also quite different: 0.273 by REML, 0.534 by reREHE and 0.610 by REHE. This may be an evidence that the variation explained by the network is much larger than the variation from noise in the true model. As illustrated in our additional simulation studies in Supplementary Note 2.4, REML may yield unreliable estimates under similar settings. We should thus take extra caution when interpreting REML-based estimates and test results in this application.

We conducted analyses for NetGSA using the R package netgsa (v3.1.0, Ma et al., 2016) on a computer with 2 × 6-core Intel Xeon X5650 @ 2.67 GHz, 96GB RAM.

4 | SIMULATION STUDIES

4.1 | Simulation settings

To benchmark the improvement of REHE and reREHE over HE and REML for variance components and heritability estimation, we generated synthetic data based on

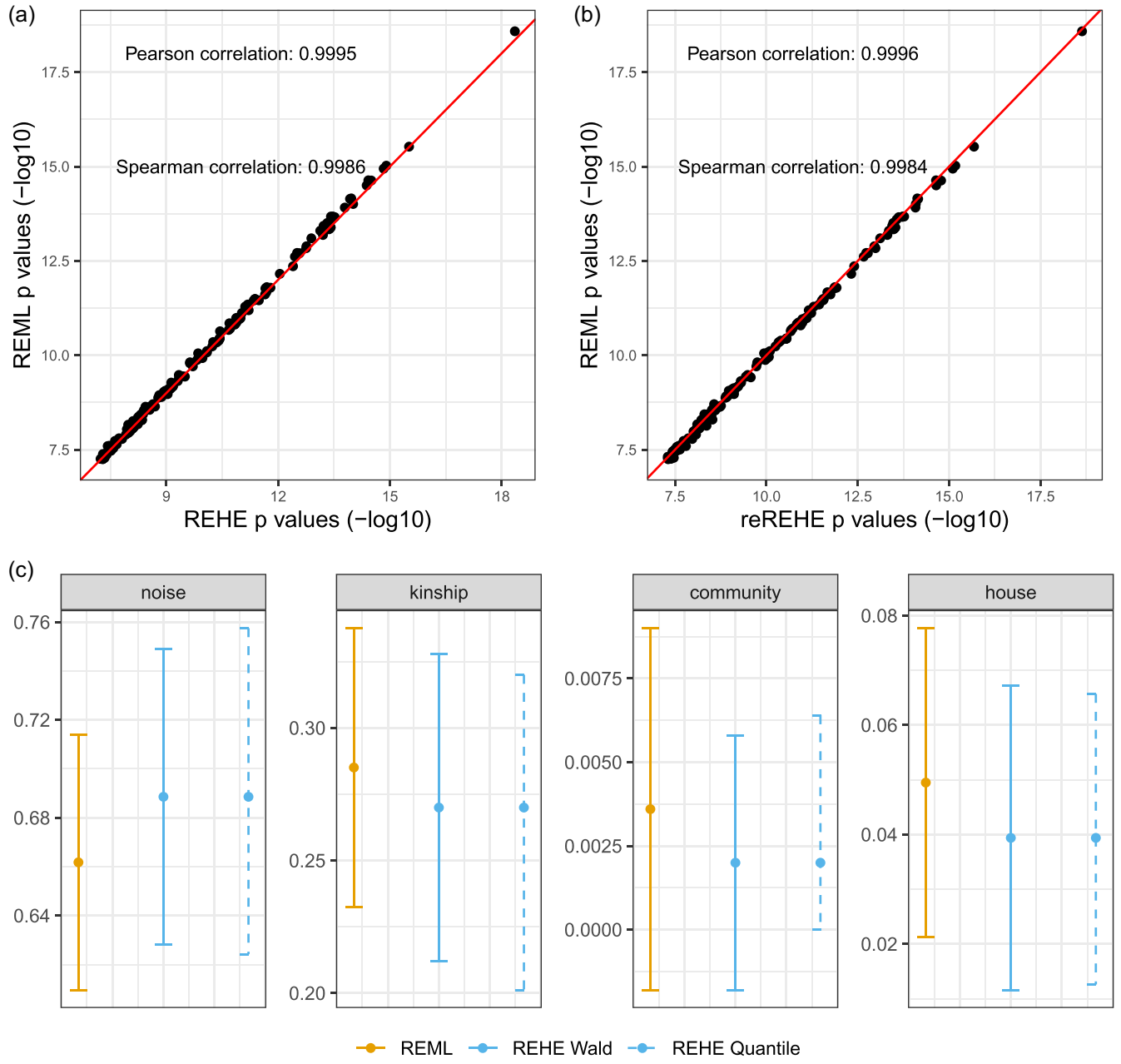

FIGURE 1 Results of genome-wide association testing analysis and heritability analysis with a HCHS/SOL data set. Association score test p-values ($-\log_{10}$ scale) based on: (a) REML against REHE estimated null models; (b) REML against reREHE estimated null models. Only 164 SNPs with resulted p values no larger than 5×10^{-8} by at least one approach were presented. (c) Dots represent point estimates of proportion of variance attributed to noise, kinship (heritability), community membership and household membership; bars represent corresponding confidence intervals. Results by REHE and REML are displayed. Two types of REHE-based confidence intervals are presented: Wald-type confidence intervals (REHE Wald), and quantile-type confidence intervals (REHE Quantile). HCHS/SOL, Hispanic Community Health Study/Study of Latinos; REHE, restricted Haseman–Elston; REML, restricted maximum likelihood

the HCHS/SOL design (Conomos et al., 2016; Sorlie et al. 2010). HE and reREHE approaches were implemented only for point estimation comparison. We truncated negative HE estimates at zero. For reREHE, we used $B = 50$ repeated subsamples, and chose sampling rates $r_s = 0.05$ (reREHE 0.05) and $r_s = 0.1$ (reREHE 0.1). Point estimates were evaluated in terms of the root mean squared error (RMSE). We constructed Wald-type (REHE-Wald) and quantile-type (REHE-quantile) confidence intervals at 95% level for REHE estimates, and compared their performances with REML-based confidence intervals in terms of coverage and interval width.

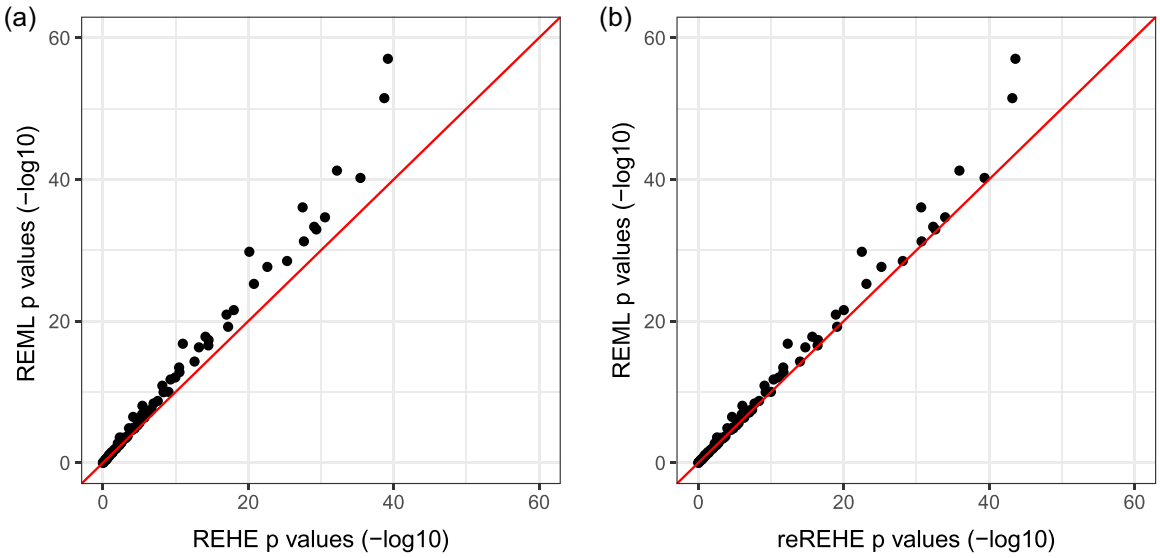

FIGURE 2 Results of network-based pathway enrichment analysis based on a breast cancer data set. p values ($-\log_{10}$ scale) of t tests for group difference of each gene pathway: (a) compare REHE results against REML results; (b) compare reREHE results against REML results. Two out-of-range data points are omitted from the plots, which correspond to: Glycosphingolipid biosynthesis—lacto and neolacto series pathway, with p value 1.09×10^{-307} by REML, 5.67×10^{-276} by REHE, 3.74×10^{-304} by reREHE; and Caffeine metabolism pathway with p value 3.97×10^{-201} by REML, 2.60×10^{-179} by REHE, and 2.13×10^{-198} by reREHE. reREHE, regression and REHE with resampling; REHE, restricted Haseman–Elston; REML, restricted maximum likelihood

We simulated data based on the linear mixed model [2]. We used sample size $n \in \{3000, 6000, 9000, 12{,}000\}$. For each sample size, we set the true values of the variance components to $(\sigma_0^2, \sigma_1^2) \in \{(0.1,0.1), (0.04,0.1), (0.1,0.04), (0.01,0.1), (0.1,0.01)\}$. These values were chosen based on previous simulation study settings (Sofer, 2017) and estimates from real data applications (Fenger, 2007). For each sample size n selected, we generated the correlation matrix D_1 as a random sub-matrix of the kinship correlation matrix from the HCHS/SOL data set. For example, for $n = 3000$, we subsampled 3000 out of 12,803 subjects without replacement, and used the corresponding (subsample) kinship correlation matrix as D_1. Under each scenario, we ran 200 replicates. Computation was performed on a computer with 2 6-core Intel Xeon CPU E5-2620 @ 2.00 GHz 128GB RAM.

We also conducted additional simulation studies to compare the approaches under different correlation structures; details of these experiments can be found in Supplementary Note 2.3.

4.2 | Simulation results

Simulation results clearly demonstrate the improvement in computational efficiency by REHE compared with REML. For point estimation, REHE was over 50 times faster than REML (Figure 3a). At the same time, REHE does not compromise estimation accuracy. Figure 3b,c show that REHE estimates of both the variance components and heritability are very close to those obtained by REML. Another advantage of REHE is that it corrects negative variance estimates from HE. To quantify this difference, We calculated the proportion of simulation replicates resulting in negative HE estimates (before zero-thresholding). This proportion reaches 23% with $n = 3000$, $(\sigma_0^2 = 0.01, \sigma_1^2 = 0.1)$, but reduces to 1.5% at $n = 12{,}000$. As pointed out before, REHE automatically corrects the issue of negative estimates without hard-thresholding. Besides, REHE has lower RMSE for point estimates when HE estimates are likely being negative (Figure 3c).

The simulation results confirm that reREHE can provide strictly positive estimates with high probability. By providing a positive variance estimate where REHE gives a zero estimate (up to 23% of the simulation repetitions), reREHE is helpful for interpretation and downstream analysis, especially under small sample sizes. As shown in Figure 3b, reREHE based estimates have smaller RMSE than all other methods at sample size $n = 3000$. With larger samples, the RMSE of reREHE is comparable to other methods under some settings (Figure 3b), but is much larger in other settings (Figure 3c, Supplementary Note 2.1 and Note 2.3). Setting a higher subsampling rate (0.1 compared with 0.05) reduces RMSE (Figure 3b,c), but comes at the cost of reduced computational efficiency—reduction in computation time compared with REHE diminishes from 67% to 10% (Figure 3a).

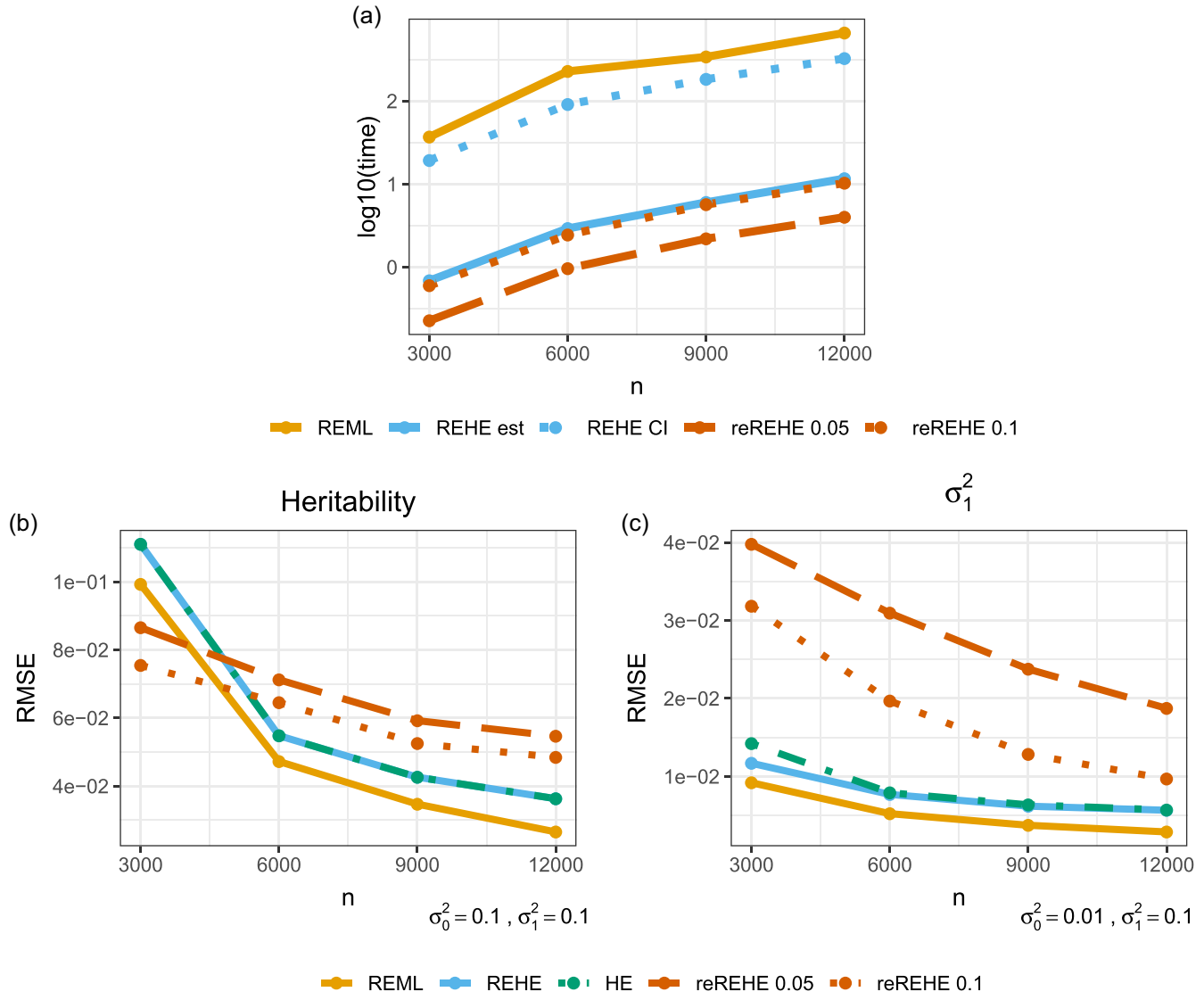

FIGURE 3 Simulation results for REHE and reREHE compared with REML. (a) CPU time in seconds ($\log_{10}$ scale): time is presented separately for fitting the model using REML (REML), for only computing point estimates by REHE (REHE est), for constructing confidence interval for variance components with REHE (REHE CI), for only computing point estimates by reREHE with subsampling rate $r_s = 0.05$ (reREHE 0.05) and reREHE with subsampling rate $r_s = 0.1$ (reREHE 0.1). (b) Root mean squared error (RMSE) for heritability estimation; simulation was based on true values $\sigma_0^2 = 0.1$, $\sigma_1^2 = 0.1$, where σ_0^2 is the variance for noise, and σ_1^2 is the variance for the random effect. (c) RMSE for σ_1^2 estimation; simulation was based on true values $\sigma_0^2 = 0.01$, $\sigma_1^2 = 0.1$. reREHE, regression and REHE with resampling; REHE, restricted Haseman–Elston; REML, restricted maximum likelihood; RMSE, root mean squared error

Turning to inference of variance components and heritability, both REHE based quantile-type and Wald-type confidence intervals provide reasonably good coverage with comparable interval width to REML confidence intervals (Figure 4). The empirical coverage is close to nominal level under most cases (Figure 4a,b), considering a Monte Carlo error of 0.03 based on 200 simulation replicates. REHE quantile-type intervals generally have better coverage than Wald-type intervals when the true variance components are very different (Figure 4b). In terms of confidence interval width, quantile-type REHE confidence intervals are generally narrower than Wald-type, and both are comparable to REML-based intervals (Figure 4c,d). Inference based on REHE is more time-consuming than REHE-based point estimation; however, it still achieves 50% reduction of computation time compared with REML (Figure 3a).

Finally, in some simulation settings, we noticed that REML confidence intervals may suffer from under-coverage. For instance, with n =3000 samples, when one variance component is substantially smaller, the

coverage of REML confidence intervals drop below 87% (Figure 4b). In other settings, REML confidence intervals even have coverage below 60%, and have little improvement with increasing sample size (Supplementary Note 2.3). Another concern is the numerical stability of REML, which fails to provide a confidence interval if the estimate of any variance component becomes zero during the iterative updates. We noticed frequent occurrence of this issue when the true variance components are unbalanced and the sample size is small—for n =3000 and (σ_0^2, σ_1^2) = (0.1,0.01), REML is unable to provide a confidence interval in 17.5% of the replicates. This proportion increases to 30.5% in other settings (Supplementary Note 2.3). We view these two issues as a warning sign for REML-based inference in real applications, especially when the underlying variance components are very

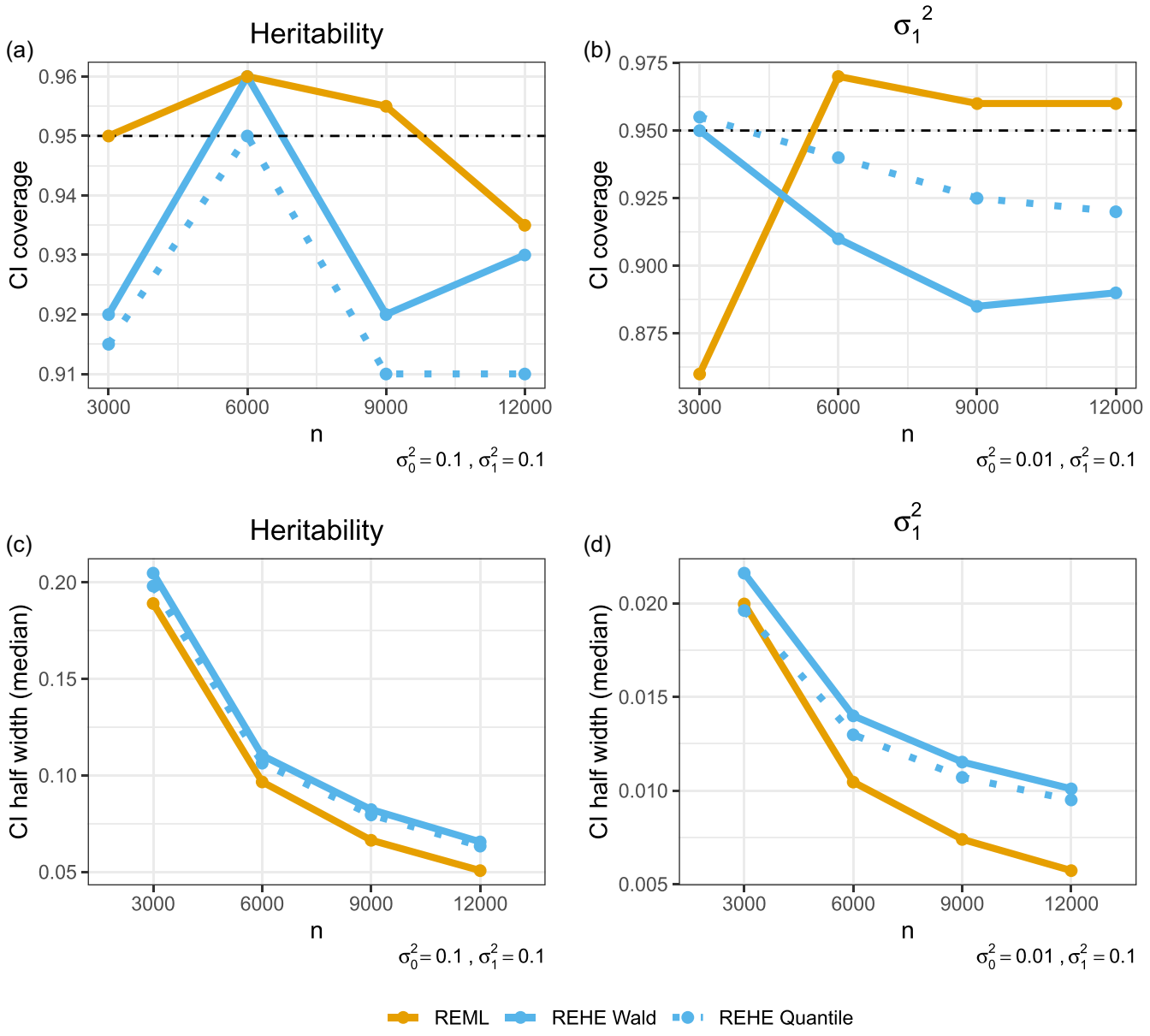

FIGURE 4 Simulation results for confidence interval performance in terms of coverage and width. σ_0^2 is the variance for noise, and σ_1^2 is the variance for the random effect. (a) Coverage for heritability confidence interval (CI); simulation was based on true values $\sigma_0^2 = 0.1$, $\sigma_1^2 = 0.1$. Monte Carlo error of 0.03 is expected for 200 simulation replications. (b) Coverage for σ_1^2 CI; simulation was based on true values $\sigma_0^2 = 0.01$, $\sigma_1^2 = 0.1$. Monte Carlo error of 0.03 is expected for 200 simulation replicates. (c) Line charts of median half width of heritability CI with increasing sample sizes; simulation was based on true values $\sigma_0^2 = 0.1$, $\sigma_1^2 = 0.1$. (d) Line charts of median half width of σ_1^2 CI with increasing sample sizes; simulation was based on true values $\sigma_0^2 = 0.01$, $\sigma_1^2 = 0.1$. REHE, restricted Haseman–Elston; REML, restricted maximum likelihood

different and the sample size is small. In contrast, REHE-based inference is robust across different settings with valid confidence intervals and acceptable empirical coverage.

The above simulation results are supported by evidence from additional simulation studies in Supplementary Note 2.3.

5 | DISCUSSION

We proposed REHE for fast estimation of variance components in linear mixed models. This new approach is motivated by large-scale genetic studies, including HCHS/SOL, but is applicable more broadly to a wide range of study designs. Through simulation studies and data applications, we demonstrated its substantial gain in computational efficiency over REML, with little compromise in point estimation accuracy. We also showed in several simulation settings that REHE can be more robust than REML for estimating the variance components (Supplementary Note 2.4). Compared with HE, REHE corrects the issue of negative estimates, and offers potentially large gains in estimation accuracy. Therefore, REHE can be superior compared with HE and be a good alternative to REML for point estimation of variance components in linear mixed models.

We also proposed reREHE based on the resampling technique to produce strictly positive variance component estimates with high probability. Strictly positive estimates are more interpretable and may be more appealing for downstream analyses. Though reREHE estimates may have lower accuracy than REHE, the magnitude of the increase in RMSE was small in our experiments. We have also seen in the real data application that based on a subsampling rate of 0.1 and 50 subsamples, reREHE-based downstream analysis results are close to REML-based results. With suitably chosen subsampling rate and number of subsampling replicates, reREHE can achieve higher computational efficiency than REHE.

As mentioned previously, one can also use the median of the subsample results as the reREHE estimate. We explored this choice in Supplementary Note 2.2 and Note 2.3. When the underlying variance components are very different, median-based reREHE estimates generally have smaller RMSE; otherwise mean-based reREHE performs better. However, median-based reREHE is more likely to yield zero estimates. A posthoc selection of the summary function can be made after observing the distribution of the subsample estimates based on reREHE.

As illustrated in the genome-wide association and pathway enrichment analysis examples in Section 3, in many applications, only variance component estimates are needed for downstream analyses. The computational burden of REML-based estimation prohibits these analyses on large data sets. Restricting the analysis to subsets of data reduces reliability and may yield contradictory conclusions. Given the fast and reliable estimates by REHE and reREHE in large data sets, we see great potential for their application in areas that only require point estimation of variance components.

When confidence intervals are also of interest, REHE remains a competitive alternative to REML for its robustness and numerical stability. As illustrated in our simulation studies, when the sample size is small and the true variance components are unbalanced, REML based inference may suffer from numerical instability and/or poor coverage. REHE consistently provides valid inference across all settings. Therefore, when the true variance components are expected to be very different and the sample size is not large, we recommend REHE over REML if inference on variance components is needed, as REML-based inference results may be unreliable.

Constructing confidence intervals for variance components when sample size is large (e.g. $n > 10{,}000$) is inherently computationally challenging. REHE only offers marginal improvements in computational efficiency over REML when it comes to inference. However, we can improve the computational efficiency for REHE confidence interval by parallelizing the bootstrap procedure. An alternative acceleration approach is to use correlation matrix sparsification (Gogarten et al., 2019, Supplementary Note 1.2). In Supplementary Note 2.1 and Note 2.3, we explored the application of sparsification for constructing confidence intervals. Our conclusion is that sparsification improves computational efficiency in large sample settings ($n > 12{,}000$); however, it may result in less robust confidence intervals for both REHE and REML. We did not explore application of sparsification to linear mixed models with more than one random effect. We expect a much larger sample size beyond which sparsification would show improvement in computational efficiency.

We did not explore confidence interval construction for reREHE estimates in this paper. Due to the repeated subsampling procedure of reREHE, an analytical expression for the confidence intervals is not trivial. The parametric bootstrap procedure for REHE confidence interval construction is readily extendable to reREHE, which we expect to have similar performance as REHE confidence intervals. However, the computational burden will also be similar to those of REHE confidence

intervals. Future research should explore fast inference procedure for REHE and reREHE estimates.

ACKNOWLEDGMENTS

This study was partially funded by grants R01-GM114029, R01-HL141989 and R01-GM133848 from the National Institutes of Health.

CONFLICT OF INTERESTS

The authors declare that there are no conflict of interests.

DATA AVAILABILITY STATEMENT

The HCHS/SOL data are available under accession numbers dbGaP: phs000880.v1.p1 and phs000810.v1.p1. The TCGA breast cancer data set is publicly available at https://github.com/drjingma/NetGSAreview. The proposed methods are implemented with codes written in the R language, which are available at https://github.com/yuek9/REHE.

ORCID

Kun Yue http://orcid.org/0000-0001-8850-2758

REFERENCES

Aulchenko, Y. S., De Koning, D. J., & Haley, C. (2007). Genomewide rapid association using mixed model and regression: A fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. *Genetics*, *177*(1), 577–585.

Berwin, A., & Turlach, A. W. (2019). quadprog: Functions to solve quadratic programming problems. R package version 1.5-7.

Conomos, M. P., Gogarten, S. M., Brown, L., Chen, H., Rice, K., Sofer, T., Thornton, T., & Yu, C. (2019). GENESIS: GENetic EStimation and Inference in Structured samples (GENESIS): Statistical methods for analyzing genetic data from samples with population structure and/or relatedness. R package version 2.14.3.

Conomos, M. P., Laurie, C. A., Stilp, A. M., Gogarten, S. M., McHugh, C. P., Nelson, S. C., Sofer, T., Fernández-Rhodes, L., Justice, A. E., Graff, M., Young, K. L., Seyerle, A. A., Avery, C. L., Taylor, K. D., Rotter, J. I., Talavera, G. A., Daviglus, M. L., Wassertheil-Smoller, S., Schneiderman, N., ... Laurie, C. C. (2016). Genetic diversity and association studies in US Hispanic/Latino populations: Applications in the Hispanic Community Health Study/Study of Latinos. *The American Journal of Human Genetics*, *98*(1), 165–184.

Dudbridge, F., & Gusnanto, A. (2008). Estimation of significance thresholds for genome wide association scans. *Genetic Epidemiology*, *32*(3), 227–234.

Fenger, M. (2007). Heritability and genetics of lipid metabolism. *Future Lipidology*, *2*(4), 433–444.

Franc, V., Hlavác, V., & Navara, M. (2005). Sequential coordinate-wise algorithm for the nonnegative least squares problem. In International Conference on Computer Analysis of Images and Patterns, Springer, pp. 407–414.

Gilmour, A. R., Thompson, R., & Cullis, B. R. (1995). Average information REML: An efficient algorithm for variance parameter estimation in linear mixed models. *Biometrics*, *55*, 1440–1450.

Gogarten, S. M., Sofer, T., Chen, H., Yu, C., Brody, J. A., Thornton, T. A., Rice, K. M., & Conomos, M. P. (2019). Genetic association testing using the genesis r/bioconductor package. *Bioinformatics*, *35*(24), 5346–5348.

Goldfarb, D., & Idnani, A. (1983). A numerically stable dual method for solving strictly convex quadratic programs. *Mathematical Programming*, *27*(1), 1–33.

Graybill, F. A. (1954). On quadratic estimates of variance components. *The Annals of Mathematical Statistics*, *25*(2), 367–372.

Graybill, F. A., & Hultquist, R. A. (1961). Theorems concerning Eisenhart's model II. *The Annals of Mathematical Statistics*, *32*(1), 261–269.

Graybill, F. A., & Wortham, A. (1956). A note on uniformly best unbiased estimators for variance components. *Journal of the American Statistical Association*, *51*(274), 266–268.

Haseman, J., & Elston, R. (1972). The investigation of linkage between a quantitative trait and a marker locus. *Behavior Genetics*, *2*(1), 3–19.

Hu, Y., Stilp, A. M., McHugh, C. P., Rao, S., Jain, D., Zheng, X., Lane, J., Méric de Bellefon, S., Raffield, L. M., Chen, M.-H., Yanek, L. R., Wheeler, M., Yao, Y., Ren, C., Broome, J., Moon, J. Y., de Vries, P. S., Hobbs, B. D., Sun, Q., ... NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium. (2021). Whole-genome sequencing association analysis of quantitative red blood cell phenotypes: The NHLBI topmed program. *The American Journal of Human Genetics*, *108*(5), 874–893.

Jiang, L., Zheng, Z., Qi, T., Kemper, K. E., Wray, N. R., Visscher, P. M., & Yang, J. (2019). A resource-efficient tool for mixed model association analysis of large-scale data. *Nature Genetics*, *51*(12), 1749–1755.

Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S.-Y., Freimer, N. B., Sabatti, C., Eskin, E., Zaitlen, N. A., Kong, S. Y., Freimer, N. B., Sabatti, C., & Eskin, E. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics*, *42*(4), 348–354.

Kim, D., Sra, S., & Dhillon, I. S. (2006). *A new projected quasi-newton approach for the nonnegative least squares problem* (Technical Report TR-06-54). Computer Science Department, University of Texas at Austin Austin.

LaVange, L. M., Kalsbeek, W. D., Sorlie, P. D., Avilés-Santa, L. M., Kaplan, R. C., Barnhart, J., Liu, K., Giachello, A., Lee, D. J., Ryan, J., Criqui, M. H., & Elder, J. P. (2010). Sample design and cohort selection in the Hispanic community health study/study of Latinos. *Annals of Epidemiology*, *20*(8), 642–649.

Lawson, C. L., & Hanson, R. J. (1995). *Solving least squares problems* (Vol. 15). Siam.

Ma, J., Shojaie, A., & Michailidis, G. (2016). Network-based pathway enrichment analysis with incomplete network information. *Bioinformatics*, *32*(20), 3165–3174.

Ma, J., Shojaie, A., & Michailidis, G. (2019). A comparative study of topology-based pathway enrichment analysis methods. *BMC Bioinformatics*, *20*(1), 546.

Matilainen, K., Mäntysaari, E. A., Lidauer, M. H., Strandén, I., & Thompson, R. (2013). Employing a Monte Carlo algorithm in newton-type methods for restricted maximum likelihood estimation of genetic parameters. *PLoS One*, *8*(12), e80821.

Metel, M. R. (2017). Mini-batch stochastic gradient descent with dynamic sample sizes. arXiv preprint arXiv:1708.00555.

Patterson, H. D., & Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, *58*(3), 545–554.

Politis, D. N., Romano, J. P., & Wolf, M. (1999). *Subsampling*. Springer Science & Business Media.

Rao, C. R. (1970). Estimation of heteroscedastic variances in linear models. *Journal of the American Statistical Association*, *65*(329), 161–172.

Rasch, D., & Masata, O. (2006). Methods of variance component estimation. *Czech Journal of Animal Science*, *51*(6), 227–235.

Searle, S. R. (1995). An overview of variance component estimation. *Metrika*, *42*(1), 215–230.

Shojaie, A., & Michailidis, G. (2009). Analysis of gene sets based on the underlying regulatory network. *Journal of Computational Biology*, *16*(3), 407–426.

Sofer, T. (2017). Confidence intervals for heritability via Haseman-Elston regression. *Statistical Applications in Genetics and Molecular Biology*, *16*(4), 259–273.

Sorlie, P. D., Avilés-Santa, L. M., Wassertheil-Smoller, S., Kaplan, R. C., Daviglus, M. L., Giachello, A. L., Schneiderman, N., Raij, L., Talavera, G., Allison, M., Lavange, L., Chambless, L. E., & Heiss, G. (2010). Design and implementation of the Hispanic community health study/study of Latinos. *Annals of Epidemiology*, *20*(8), 629–641.

TCGA. (2012). Comprehensive molecular portraits of human breast tumours. *Nature*, *490*(7418), 61–70.

Wojcik, G. L., Graff, M., Nishimura, K. K., Tao, R., Haessler, J., Gignoux, C. R., Highland, H. M., Patel, Y. M., Sorokin, E. P., Avery, C. L., Belbin, G. M., Bien, S. A., Cheng, I., Cullina, S., Hodonsky, C. J., Hu, Y., Huckins, L. M., Jeff, J., Justice, A. E., ... Carlson, C. S. (2019). Genetic analyses of diverse populations improves discovery for complex traits. *Nature*, *570*(7762), 514–518.

Xie, M., & Yang, Y. (2003). Asymptotics for generalized estimating equations with large cluster sizes. *The Annals of Statistics*, *31*(1), 310–347.

Zhou, X. (2017). A unified framework for variance component estimation with summary statistics in genome-wide association studies. *The Annals of Applied Statistics*, *11*(4), 2027–2051.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Yue, K., Ma, J., Thornton, T., & Shojaie, A. (2021). REHE: Fast variance components estimation for linear mixed models. *Genetic Epidemiology*, 45, 891–905. https://doi.org/10.1002/gepi.22432

APPENDIX A: HE UNBIASEDNESS

The HE estimator is unbiased. The unbiasedness holds even if we use only a subset of the n^2 estimating equations, that is, ignoring some entries of the relatedness matrices D_k's.

Recall n^2 estimating equations are constructed:

$$\mathrm{E}[Y_i Y_j] = \sigma_0^2 D_0^{ij} + \sigma_1^2 D_1^{ij},\ i, j = 1, 2, \ldots n.$$

Suppose we use S out of the n^2 estimating equations without duplicates to compute the HE estimates, $S \leq n^2$. Let $(i_1, j_1), \ldots, (i_S, j_S)$ denote the selected indices. Similar to Section 2.1, we recast the problem of solving the S estimating equations as the following linear regression problem. Let $Y_S = (Y_{i_1} Y_{j_1}, \ldots, Y_{i_S} Y_{j_S})$ denote the linear regression outcome vector, and let X_S denote the design matrix by stacking the tuples $(D_0^{i_s j_s}, D_1^{i_s j_s})$, $s = 1, \ldots, S$. Note that we have

$$\mathrm{E}[Y_S] = X_S \boldsymbol{\sigma}^2,$$

where the equality holds regardless of the higher moments of Y or the positions of the S indices. The closed form expression of the HE estimates in this case is $\boldsymbol{\sigma}^2 = (X_S^\top X_S)^{-1} X_S^\top Y_S$. We thus have:

$$\mathrm{E}(\boldsymbol{\sigma}^2) = \left(X_S^\top X_S\right)^{-1} X_S^\top \mathrm{E}(Y_S) = \left(X_S^\top X_S\right)^{-1} X_S^\top X_S \boldsymbol{\sigma}^2 = \boldsymbol{\sigma}^2.$$

Therefore, the HE estimates is unbiased, even if a subset of the n^2 estimating equations is used.

APPENDIX B: REHE CONSISTENCY AND ASYMPTOTIC NORMALITY

The REHE estimator is consistent under mild conditions, and is asymptotically normal when the random effect correlation matrix is sparse and block-diagonal.

For expositional clarity, we assume the diagonal entries of the kinship matrix D_1 are scaled to 1. We proceed with first establishing consistency of the HE estimator. By [3], the HE estimate is the ordinary least squares solution to a linear regression problem. We write out the model for this linear regression as

$$Y = X\boldsymbol{\sigma}^2 + \epsilon,$$

for the length n^2 outcome vector $Y = \mathrm{vec}(YY^\top)$, the $n^2 \times 2$ fixed effect design matrix $X = (\mathrm{vec}(D_0), \mathrm{vec}(D_1))$ and the length n^2 error vector ϵ with covariance matrix R. Note that $R^{ij} = \mathrm{Cov}(Y_i, Y_j)$, which is related to the forth moment of Y. Thus R is not a diagonal matrix. Under this

framework, the HE estimator is a special case of the generalized estimating equations estimator, with I_n as the *working correlation matrix*, the total number of clusters bounded (equals to 1 in this case), and the cluster size n^2 going to infinity. Such as estimator has been studied in Xie and Yang (2003), see Example 2.1 and Example 5.1. The consistency of HE estimates holds under the condition (I_ω) or (I_ω^*) in Xie and Yang (2003), which in our case corresponds to

$$(I_\omega): \lambda_{\min}\left(H_n M_n^{-1} H_n\right) \xrightarrow{n\to\infty} \infty,$$

$$(I_\omega^*): \lambda_{\min}(H\_n)/\lambda_{\max}(R) \xrightarrow{n\to\infty} \infty,$$

where

$$H_n = X^\top X$$

$$M_n = X^\top R X,$$

and $\lambda_{\min}(X)(\lambda_{\max}(X))$ denotes the smallest (largest) eigenvalue of the symmetric matrix X. As discussed in Xie and Yang (2003), for (I_ω) or (I_ω^*) to hold it suffices to bound the correlation of ϵ by $|R^{ij}/\sqrt{R^{ii}R^{jj}}| \le \rho_h$, where $h = |i-j|$ and $\{\rho_h\}_{h=1}^\infty$ is a sequence of values with $\lim_{h\to\infty} \rho_h = 0$. One can then verify that Assumption [1] is sufficient to guarantee that the correlations of ϵ are properly bounded, based on $\mathrm{Cov}(Y_i Y_j, Y_{i'} Y_{j'}) = V_Y^{ii'} V_Y^{jj'} + V_Y^{ij'} V_Y^{ji'}$, for $V_Y = \mathrm{E}(YY^\top) = \sigma_0^2 D_0 + \sigma_1^2 D_1$.

Assumption 1 (Weak Condition). *The covariance matrix* D_1 satisfies $|D_1^{ij}| \le \rho_h^*$, $h = |i-j|$, for some sequence of values $\{\rho_h^*\}_{h=1}^\infty$ with $\lim_{h\to\infty} \rho_h^* = 0$.

Assumption [1] is natural in many genetic studies as it is expected that off-diagonal correlation entries would decrease to zero after properly rearranging the rows and columns of the correlation matrix.

The asymptotic normality of the HE estimator requires an stronger assumption on the structure of D_1. Note that we do not need this strong assumption to show the consistency of REHE or to construct bootstrap confidence intervals for the REHE estimator. Suppose Assumption [2] holds for a sparse and block-diagonal kinship matrix D_1.

Assumption 2 (Strong Condition). D_1 is sparse and block-diagonal such that

$$D_1 = \begin{pmatrix} D_1^{(1)} & 0 & 0 & \dots \\ 0 & D_1^{(2)} & 0 & \dots \\ \vdots & & \ddots & \\ 0 & 0 & 0 & D_1^{(M)} \end{pmatrix},$$

where $D_1^{(m)} (m = 1, \dots, M)$ are square blocks along the diagonal of D_1, and M is the total number of blocks.

Such correlation structures are often approximately satisfied in genetic studies: for example, subjects are highly genetically correlated within the same family, and are remotely correlated across families. Let s_m denote the number of rows/columns in $D_1^{(m)}$, and $Y^{(m)}$ denote the $s_m \times 1$ subvector of Y corresponding to the m^{th} block. For example, $Y^{(1)}$ is the subvector corresponding to the first s_1 elements of Y. By construction, $Y^{(m)}$'s are independent and normally distributed with zero mean and covariance $\sigma_0^2 I_{s_m} + \sigma_1^2 D_1^{(m)}$, for $m = 1, \dots, M$. For sparse block-diagonal D_1, we simplify Y and X in [3] by discarding elements corresponding to zero entries of D_1:

$$Y = \begin{pmatrix} Y^{(1)} \\ Y^{(2)} \\ \vdots \\ Y^{(M)} \end{pmatrix}, \quad X = \begin{pmatrix} X^{(1)} \\ X^{(2)} \\ \vdots \\ X^{(M)} \end{pmatrix},$$

where we denote $\tilde{Y}^{(m)} = \mathrm{vec}(Y^{(m)} Y^{(m)\top})$ and $\tilde{X}^{(m)} = (\mathrm{vec}(I_n^{(m)}), \mathrm{vec}(D_1^{(m)}))$, for $m = 1, \dots, M$. The independence of $Y^{(m)}$'s implies independence of $\tilde{Y}^{(m)}$'s. As the number of blocks $M \to \infty$, the HE estimates $\boldsymbol{\sigma}_M^2 = (\sigma_{0,M}^2, \sigma_{1,M}^2)$ converge in probability to the true variance components $\boldsymbol{\sigma}^2 = (\sigma_0^2, \sigma_1^2)$ at a rate of $\sqrt{M}$ or faster (Xie & Yang, 2003). Moreover, when the maximum block size $s = \max_{1\le m\le M}(s_m)$ does not go to infinity too fast as $M \to \infty$, the HE estimates are asymptotically normal, with a rate of $\sqrt{M}$ (Xie & Yang, 2003):

$$W_M^{-1/2} H_M \left(\hat{\boldsymbol{\sigma}}_M^2 - \boldsymbol{\sigma}^2\right) \xrightarrow{d} N_2(0, I_2), \tag{6}$$

where

$$W_M = \sum_{m=1}^M (X^{(m)})^\top R^{(m)} X^{(m)}, R^{(m)} = \mathrm{Cor}(Y^{(m)}),$$

$$H_M = \sum_{m=1}^M (X^{(m)})^\top X^{(m)}.$$

In genetic studies, it is typical that subjects belong to small unrelated groups so that s is bounded, and the number of groups M increases with increasing sample sizes. These settings satisfy the conditions for the HE estimator's consistency and asymptotic normality.

We are now ready to establish the asymptotic properties of the REHE estimator. As implied by [4], the REHE estimates $\boldsymbol{\sigma}^2$ are different from the HE estimates $\boldsymbol{\sigma}^2$ only when HE yields negative estimates for some variance components. Let $\mathbb{P}(\sigma_0^2 \geq 0, \sigma_1^2 \geq 0)$ denote the probability of the HE estimates being nonnegative, which equals $\mathbb{P}(\boldsymbol{\sigma}^2 = \boldsymbol{\sigma}^2)$. Based on the consistency of the HE estimator, we have:

$$\mathbb{P}(\boldsymbol{\sigma}^2 = \boldsymbol{\sigma}^2) = \mathbb{P}(\sigma_0^2 \geq 0, \sigma_1^2 \geq 0) \xrightarrow{n\to\infty} 1,$$

indicating asymptotic equivalence of the HE and REHE estimators. This implies the consistency of the REHE estimator, which only requires Assumption [1] on D_1. Note that despite their asymptotic equivalence, our simulation results in Section 4 clearly show the advantages of REHE over HE in finite samples.

Next, we show that under the stronger Assumption [2], $W_M^{-1/2} H_M(\boldsymbol{\sigma}^2 - \boldsymbol{\sigma}^2)$ is $o_p(1)$:

$$\mathbb{P}\left(W_M^{-1/2} H_M(\tilde{\boldsymbol{\sigma}}^2 - \hat{\boldsymbol{\sigma}}^2) = 0\right) \geq \mathbb{P}(\tilde{\boldsymbol{\sigma}}^2 - \hat{\boldsymbol{\sigma}}^2 = 0) \xrightarrow{M\to\infty} 1.$$

We then have

$$\begin{aligned} W_M^{-\frac{1}{2}} H_M(\tilde{\boldsymbol{\sigma}}^2 - \boldsymbol{\sigma}^2) &= W_M^{-\frac{1}{2}} H_M(\tilde{\boldsymbol{\sigma}}^2 - \hat{\boldsymbol{\sigma}}^2) \\ &\quad + W_M^{-\frac{1}{2}} H_M(\hat{\boldsymbol{\sigma}}^2 - \boldsymbol{\sigma}^2) \\ &= o_p(1) + W_M^{-\frac{1}{2}} H_M(\hat{\boldsymbol{\sigma}}^2 - \boldsymbol{\sigma}^2). \end{aligned}$$

Therefore, we have established that under Assumption [2], the REHE estimator $\boldsymbol{\sigma}^2$ satisfies:

$$W_M^{-1/2} H_M(\tilde{\boldsymbol{\sigma}}^2 - \boldsymbol{\sigma}^2) \xrightarrow{d} N_2(0, I_2). \tag{7}$$

Here we note that the REHE estimator is slightly biased in finite samples due to the correction of negative estimates. However, as we have observed in the simulation studies, the variance and mean squared error of the REHE estimator were smaller than those of the HE estimator, indicating the REHE a better estimator.