

TECHNICAL NOTE

Bayes factor for linear mixed model in genetic association studies

Yongtao Guan | Daniel Levy

Framingham Heart Study, National Heart, Lung, and Blood Institute, Framingham, Massachusetts, USA

Correspondence
 Yongtao Guan.
 Email: yguan@bimsa.cn
Present address

Yongtao Guan, Beijing Institute of Mathematical Sciences and Applications, Beijing, China

Abstract

Bayes factor has advantages over p -value as test statistics for association, particularly when comparing multiple non-nested alternative models. An efficient method to compute Bayes factor for linear mixed model (LMM) in the context of genetic association studies is lacking. In this study, we transform the standard LMM to a Bayesian linear regression by substituting the random effect with fixed effects, where the covariates of the fixed effects are eigenvectors of the genetic relatedness matrix and their respective prior effect sizes are proportional to the corresponding eigenvalues. Using conjugate normal inverse gamma priors on regression parameters, Bayes factors can be computed in a closed form. We demonstrate numerically the known relationship between Bayes factors and p -values for the LMM. We then show that predictions based on the transformed Bayesian linear regression are identical to those of the best linear unbiased prediction (BLUP) of the standard LMM. Our results provided a new perspective and derivation to a known connection between BLUP and Bayesian estimates. Methods described in this note are implemented in the software IDUL as two new functionalities: computing Bayes factors and residuals for the LMM.

KEYWORDS

Bayes factor, BLUP, genetic association studies, IDUL, linear mixed model

1 | INTRODUCTION

The linear mixed model (LMM) has emerged as an effective tool in controlling for population structure and relatedness in a study sample to reduce false positives in genetic association studies [1]. Consider a standard LMM in genetic association literature

$$\begin{aligned}
 \mathbf{y} &= \mathbf{W}\boldsymbol{\alpha} + \mathbf{x}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e} \\
 \mathbf{u} &\sim \text{MVN}_n(0, \tau^{-1} \boldsymbol{\eta} \mathbf{K}) \\
 \mathbf{e} &\sim \text{MVN}_n(0, \tau^{-1} \mathbf{I}_n),
 \end{aligned} \tag{1}$$

where $\mathbf{W}$ contains conventional covariates such as age and sex, including a column of 1; $\mathbf{x}$ contains genetic variant(s) to be tested for association; and $\mathbf{u}$ is the random effect with $\mathbf{Z}$ as its loading matrix and $\mathbf{K}$, whose (i, j) -th entry is twice the kinship between samples i and j , as its covariance matrix (both $\mathbf{Z}$ and $\mathbf{K}$ are known). MVN_n denotes an n -dimensional multivariate normal distribution, $\mathbf{I}_n$ is n -dimensional identity matrix. In genetic association studies, the random effect $\mathbf{Z}\mathbf{u}$ is a nuisance term that absorbs part of the phenotype $\mathbf{y}$ that is attributable to population stratification and relatedness. Here $\boldsymbol{\eta}$ is a hyperparameter, measuring the ratio between two dispersion terms (random effect $\mathbf{u}$ and random noise $\mathbf{e}$).

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

Published 2026. This article is a U.S. Government work and is in the public domain in the USA. *Quantitative Biology* published by John Wiley & Sons Australia, Ltd on behalf of Higher Education Press.

In classical literature, η is known [2]. In Genome-Wide Association Study applications, however, η is an important parameter to be estimated, either using maximum likelihood (ML) or restricted maximum likelihood (REML) approaches [3]. Conditioning on η , estimates of $\hat{\mathbf{a}}$ and $\hat{\mathbf{b}}$ can be obtained and either the likelihood ratio test or Wald test can be used to compute a p -value to test against the following null model

$$\begin{aligned} \mathbf{y} &= \mathbf{W}\mathbf{a} + \mathbf{Z}\mathbf{u} + \mathbf{e} \\ \mathbf{u} &\sim \text{MVN}_n(0, \tau^{-1} \eta \mathbf{K}) \\ \mathbf{e} &\sim \text{MVN}_n(0, \tau^{-1} \mathbf{I}_n), \end{aligned} \quad (2)$$

where the hyperparameter η was estimated from the alternative Model (1). IDUL algorithm is an efficient and asymptotically exact algorithm to estimate η [3], and IDUL software can compute likelihood ratio test p -value using ML estimate of η or Wald test p -value using REML estimate of η .

Bayes factors have advantages over p -values, particularly for comparing multiple non-nested alternative models [4]. An efficient method to compute Bayes factors for LMM in the context of genetic association studies is missing. Our main objective here is to develop an efficient method to evaluate Bayes factors for the LMM. In order to do that, we first transform the LMM Model (1) to a Bayesian linear regression with a specific prior, then use standard normal inverse gamma priors on regression parameters to express Bayes factors in a closed form, which can be evaluated efficiently. The transformed Bayesian linear regression can be used for prediction, and we show that the Bayesian prediction is identical to that of best linear unbiased prediction (BLUP) [5].

2 | RESULTS

2.1 | Bayesian linear regression

With reference to Model (1), the genetic relatedness matrix can be computed as $\mathbf{G} = \mathbf{Z}\mathbf{K}\mathbf{Z}^t$. Let the eigen decomposition of $\mathbf{G}$ as $\mathbf{G} = \mathbf{Q}\mathbf{D}\mathbf{Q}^t$, where $\mathbf{D} = \text{diag}(d_1, \dots, d_n)$ with $d_1 \geq d_2 \geq \dots \geq d_n$ and $\mathbf{Q}\mathbf{Q}^t = \mathbf{I}_n$. Model (1) can be rewritten as

$$\begin{aligned} \mathbf{y} &= \mathbf{W}\mathbf{a} + \mathbf{x}\boldsymbol{\beta} + \mathbf{Q}\mathbf{b} + \mathbf{e} \\ \mathbf{b} &\sim \text{MVN}_n(0, \tau^{-1} \eta \mathbf{D}) \\ \mathbf{e} &\sim \text{MVN}_n(0, \tau^{-1} \mathbf{I}_n), \end{aligned} \quad (3)$$

where the prior on $\mathbf{b}$ has been specified. Model (3) is equivalent to Model (1). To see this, $E(\mathbf{Z}\mathbf{u}\mathbf{u}^t\mathbf{Z}^t) = \eta\mathbf{Z}\mathbf{K}\mathbf{Z}^t = \eta\mathbf{G} = \eta\mathbf{Q}\mathbf{D}\mathbf{Q}^t = E(\mathbf{Q}\boldsymbol{\gamma}\boldsymbol{\gamma}^t\mathbf{Q}^t)$. Similarly, Model (2) can be rewritten as

$$\begin{aligned} \mathbf{y} &= \mathbf{W}\mathbf{a} + \mathbf{Q}\mathbf{b} + \mathbf{e} \\ \mathbf{b} &\sim \text{MVN}_n(0, \tau^{-1} \eta \mathbf{D}) \\ \mathbf{e} &\sim \text{MVN}_n(0, \tau^{-1} \mathbf{I}_n). \end{aligned} \quad (4)$$

Note the priors on $\mathbf{b}$ are the same between the null and the alternative models. The hyperparameter η can be estimated and plugged in. Its ML estimate can be efficiently computed [3].

2.2 | Priors and Bayes factors

Let $\mathbf{X} = (\mathbf{Q}, \mathbf{W})$. By specifying a Gamma prior on τ , we have the standard normal-inverse-gamma on a linear model [6, 7].

$$\begin{aligned} \mathbf{a} &\sim \text{MVN}(0, \tau^{-1} V_0) \\ \boldsymbol{\beta} &\sim N(0, \tau^{-1} \sigma^2) \\ \tau &\sim \Gamma(\kappa_1/2, \kappa_2/2) \end{aligned} \quad (5)$$

For Model (4), after integrating out $\mathbf{a}, \mathbf{b}$ and τ and letting $\kappa_1, \kappa_2 \rightarrow 0$, we have

$$\begin{aligned} p(\mathbf{y}|\mathbf{M}_0) &= \frac{(2\pi)^{-n/2} \Gamma(n/2)}{|\mathbf{X}^t\mathbf{X} + V_{\mathbf{x}}^{-1}|^{1/2} \cdot |V_{\mathbf{x}}|^{1/2}} \\ &\cdot \left(\frac{\mathbf{y}^t\mathbf{y} - \mathbf{y}^t\mathbf{X}(\mathbf{X}^t\mathbf{X} + V_{\mathbf{x}}^{-1})^{-1}\mathbf{X}^t\mathbf{y}}{2} \right)^{-n/2}, \end{aligned} \quad (6)$$

where $V_{\mathbf{x}} = \text{diag}(\hat{\eta}\mathbf{D}, V_0)$. Similarly, for Model (3), we have

$$\begin{aligned} p(\mathbf{y}|\mathbf{M}_1) &= \frac{(2\pi)^{-n/2} \Gamma(n/2)}{|(\mathbf{X}, \mathbf{x})^t(\mathbf{X}, \mathbf{x}) + V_{(\mathbf{x}, \mathbf{x})}^{-1}|^{1/2} \cdot |V_{(\mathbf{x}, \mathbf{x})}|^{1/2}} \\ &\cdot \left(\frac{\mathbf{y}^t\mathbf{y} - \mathbf{y}^t(\mathbf{X}, \mathbf{x})((\mathbf{X}, \mathbf{x})^t(\mathbf{X}, \mathbf{x}) + V_{(\mathbf{x}, \mathbf{x})}^{-1})^{-1}(\mathbf{X}, \mathbf{x})^t\mathbf{y}}{2} \right)^{-n/2}, \end{aligned} \quad (7)$$

where $V_{(\mathbf{x}, \mathbf{x})} = \text{diag}(\hat{\eta}\mathbf{D}, V_0, \sigma^2)$. The Bayes factor $p(\mathbf{y}|M_1)/p(\mathbf{y}|M_0)$, after letting $V_0^{-1} \rightarrow \mathbf{0}$, is

$$\text{BF}(\sigma) = \frac{|\mathbf{X}^t\mathbf{X} + V_{\mathbf{x}}^{-1}|^{1/2}}{[(\mathbf{X}, \mathbf{x})^t(\mathbf{X}, \mathbf{x}) + V_{(\mathbf{x}, \mathbf{x})}^{-1}]^{1/2}} \frac{1}{\sigma} \left(\frac{\mathbf{y}^t\mathbf{y} - \mathbf{y}^t(\mathbf{X}, \mathbf{x})((\mathbf{X}, \mathbf{x})^t(\mathbf{X}, \mathbf{x}) + V_{(\mathbf{x}, \mathbf{x})}^{-1})^{-1}(\mathbf{X}, \mathbf{x})^t\mathbf{y}}{\mathbf{y}^t\mathbf{y} - \mathbf{y}^t\mathbf{X}(\mathbf{X}^t\mathbf{X} + V_{\mathbf{x}}^{-1})^{-1}\mathbf{X}^t\mathbf{y}} \right)^{-n/2}. \quad (8)$$

Note that, by letting $V_0^{-1} \rightarrow \mathbf{0}$, we used improper priors, but both the posterior and Bayes factors are proper [6]. The Bayes factor can be evaluated efficiently (Section 4).

In practice, we need to specify σ to compute Bayes factors. Following prior works [6], we used $\sigma = 0.5$. With this choice, the prior $N(0, \sigma^2)$ puts a probability of 0.34 for effect size to be in interval $(-0.5, 0.5)$, a probability of 0.63 for effect size to be in interval $(-1, 1)$, and a probability of 0.93 for effect size to be in interval $(-2, 2)$. One can also learn prior effect sizes from a histogram estimates of past study and use it as prior for the current study.

2.3 | Numerical examples

Figure 1 shows comparisons between Bayes factors computed using Equation (8) and p -values from the likelihood ratio test calculated using IDUL software

[3], stratified according to minor allele frequencies (MAF) of the SNPs. The data, consisting 5757 whole genome sequenced samples with high level of relatedness and 79 protein assay phenotypes, were described previously [8]. We make the following observations: (1) In general, large Bayes factors correspond to small p -values. (2) Different Bayes factors might have the same p -values, and different p -values might have the same Bayes factors. (3) The left panel shows that, for SNPs with the same p -values, those having larger MAF tend to have larger Bayes factors than those having smaller MAF. This is because Bayes factors penalize SNPs of small MAF for their lower information contents through the prior effect size [9]. (4) Previously, it has been demonstrated that asymptotically $2\log_e \text{BF} = \lambda\chi^2 + \log_e(1 - \lambda)$, where $\lambda \in (0, 1)$ increases with the SNP MAF and is also a function of prior effect size [7]. The right panel reflects the mathematical insight between Bayes factors and their corresponding χ^2 -values.

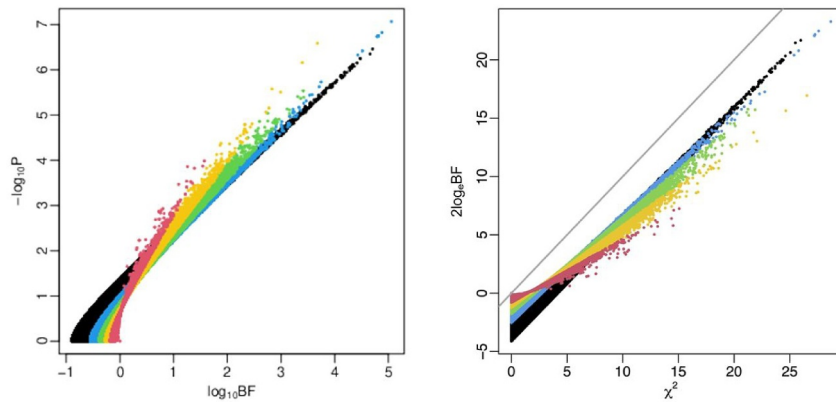


FIGURE 1 Comparison between p -values and Bayes factors for linear mixed model. The left panel shows $\log_{10} \text{BF}$ versus $-\log_{10} P$. The right panel shows χ^2 versus $2\log_{10} \text{BF}$, where χ^2 values are quantiles of the p -values. The gray line is $x = y$. The color is based on MAF f , with black for $f \in (0.05, 0.5]$, blue for $f \in (0.02, 0.05]$, green for $f \in (0.01, 0.02]$, orange for $f \in (0.005, 0.01]$, and red for $f \in (0, 0.005]$. Note each color contains a range of allele frequencies instead of a single allele frequency. A single allele frequency corresponding to a single λ in the relationship of $2\log_e \text{BF} = \lambda\chi^2 + \log_e(1 - \lambda)$.

2.4 | Bayesian prediction and BLUP

From Model (4) we estimate parameters of $\hat{\mathbf{a}}$ and $\hat{\mathbf{b}}$ by $(\mathbf{X}^t\mathbf{X} + V_{\mathbf{X}}^{-1})^{-1}\mathbf{X}^t\mathbf{y}$ (Section 4). Note both are column vectors. For a new sample, we have a row vector $\mathbf{w}$ for covariates and a row vector $\mathbf{g}$ for genetic relatedness between the new sample with n existing samples. Then the Bayesian prediction for the new sample is

$$\hat{\mathbf{y}} = \mathbf{w}\hat{\mathbf{a}} + \mathbf{q}\hat{\mathbf{b}}, \quad (9)$$

where $\mathbf{q} = \mathbf{g}\mathbf{G}^{-1}\mathbf{Q}$ is the least square prediction (based on genetic relatedness) of the new sample onto $\mathbf{Q}$ (Section 4). For an existing sample j , $\mathbf{q}$ is the j -th row of the matrix $\mathbf{Q}$.

We now show that the BLUP is identical to Bayesian prediction. From the frequentist point of view, $\mathbf{u}$ is a random vector instead of a fixed parameter, it can only be *predicted* instead of *estimated*. The Bayesian interpretation of BLUP is well known [5]. Here we offer a new derivation that is perhaps more natural from the perspective of genetic association studies.

For BLUP one first estimates $\tilde{\mathbf{a}} = (\mathbf{W}^t\mathbf{H}^{-1}\mathbf{W})^{-1}\mathbf{W}^t\mathbf{H}^{-1}\mathbf{y}$, where $\mathbf{H} = \hat{\eta}\mathbf{G} + \mathbf{I}_n$, and then estimate $\tilde{\mathbf{u}} = \hat{\eta}\mathbf{Z}\mathbf{K}\mathbf{Z}^t\mathbf{H}^{-1}(\mathbf{y} - \mathbf{W}\tilde{\mathbf{a}})$. Note both are column vectors. Thus for the new sample, BLUP prediction is

$$\tilde{\mathbf{y}} = \mathbf{w}\tilde{\mathbf{a}} + \mathbf{z}\tilde{\mathbf{u}}, \quad (10)$$

where $\mathbf{z} = \mathbf{g}\mathbf{G}^{-1}\mathbf{Z}$ is the least square prediction (based on genetic relatedness) of the new sample onto $\mathbf{Z}$ (Section 4). For an existing sample j , $\mathbf{z}$ is the j -th row of the matrix $\mathbf{Z}$.

To show Bayesian prediction $\hat{\mathbf{y}}$ in Equation (9) is identical to BLUP $\tilde{\mathbf{y}}$ in Equation (10), it is sufficient to show that $\hat{\mathbf{a}} = \tilde{\mathbf{a}}$ and $\mathbf{Q}\hat{\mathbf{b}} = \mathbf{Z}\tilde{\mathbf{u}}$, which we defer to Section 4. Therefore, BLUP is identical to Bayesian prediction under a specific prior. For the existing samples, we can compute residuals using either Bayesian estimates or BLUP as following:

$$\begin{aligned} \mathbf{r} &= \mathbf{y} - \mathbf{W}\hat{\mathbf{a}} - \mathbf{Q}\hat{\mathbf{b}} \\ \mathbf{r} &= \mathbf{H}^{-1}(\mathbf{y} - \mathbf{W}\hat{\mathbf{a}}). \end{aligned} \quad (11)$$

Residuals are useful in removing spurious correlations due to population structure and (cryptic) relatedness for subsequent analysis, such as studying gene-gene interaction and gene networks using gene expression values.

3 | DISCUSSION

It is a common practice to control for top principal components (PCs, the columns of the $\mathbf{Q}$ matrix) in genetic association study to control for population stratification. It is self-evident that if a putative causal SNP is correlated with both phenotype and top PCs, this practice of controlling for top PCs will reduce power. Our derivation suggests that when controlling for top PCs, one should put prior on their effect sizes $\hat{\eta}\mathbf{D}^*$, where $\mathbf{D}^*$ is a submatrix of $\mathbf{D}$ corresponding to the top PCs. Intuitively, the prior will keep some variations of PCs in the phenotypes for the putative causal SNPs, and a priori, the eigenvalues are the right amount of variation to be controlled for.

In classical literature dealing with LMM [2, 5], the covariance structure of the error term is a more general $\mathbf{R}$ instead of our diagonal $\mathbf{I}_n$. We can left multiplying $\mathbf{R}^{-\frac{1}{2}}$ (where $\mathbf{R}^{-}$ is a generalized inverse) to convert the classical form to the form in Model (1). In fact, we can rotate the system further and compute the (empirical) Bayes factor based on a rotated simple linear regression. Multiplying both side of Model (4) by $\mathbf{Q}^t$ we get $\mathbf{y}_{\mathbf{Q}} \sim \text{MVN}(\mathbf{W}_{\mathbf{Q}}\hat{\eta}\mathbf{D} + \mathbf{I}_n)$, where $\mathbf{y}_{\mathbf{Q}} = \mathbf{Q}^t\mathbf{y}$ and $\mathbf{W}_{\mathbf{Q}} = \mathbf{Q}^t\mathbf{W}$. Let $\Lambda = \eta\mathbf{D} + \mathbf{I}_n$ and scale $\mathbf{y}_{\mathbf{Q}}$ and $\mathbf{W}_{\mathbf{Q}}$ by left multiplying $\Lambda^{-\frac{1}{2}}$, we get

$$\begin{aligned} \mathbf{y}_{\Lambda} &= \mathbf{W}_{\Lambda} \cdot \mathbf{a} + \mathbf{e} \\ \mathbf{e} &\sim \text{MVN}_n(0, \tau^{-1} \mathbf{I}_n), \end{aligned} \quad (12)$$

where $\mathbf{y}_{\Lambda} = \Lambda^{-\frac{1}{2}}\mathbf{Q}^t\mathbf{y}$ and $\mathbf{W}_{\Lambda} = \Lambda^{-\frac{1}{2}}\mathbf{Q}^t\mathbf{W}$. Similarly Model (3) becomes

$$\begin{aligned} \mathbf{y}_{\Lambda} &= \mathbf{W}_{\Lambda} \cdot \mathbf{a} + \mathbf{x}_{\Lambda} \cdot \beta + \mathbf{e} \\ \mathbf{e} &\sim \text{MVN}_n(0, \tau^{-1} \mathbf{I}_n), \end{aligned} \quad (13)$$

where $\mathbf{x}_{\Lambda} = \Lambda^{-\frac{1}{2}}\mathbf{Q}^t\mathbf{x}$. Then using priors specified in Equation (5), we can compute Bayes factors [6]. This rotation approach was briefly mentioned by Wen [10]. Bayes factors from this approach is equivalent to Equation (8), but this approach obscures insights on Bayesian prediction and its equivalence to BLUP.

We took an empirical Bayesian approach and plugged in ML estimate of the hyperparameter $\hat{\eta}$. Of course, we can take the full Bayesian approach by coming up with a prior for η . One sensible choice is to define $h = \eta/(1 + \eta)$ and assign a uniform prior $\text{Unif}(0, 1)$ on h . Then η has a probability density function $f(\eta) = 1/(1 + \eta)^2$, which is heavy tailed [9]. This full Bayesian approach is advantageous in principle, as it accounts for variation of η , but the computation is difficult, as this prior cannot be integrated out analytically, and

either sampling or numerical integration are required for computation. For example, Wen [10] took a full Bayesian approach assigning priors on η and τ , and elected to compute approximate Bayes factors for LMM. On the other hand, by plugging in ML estimate $\hat{\eta}$, the subsequent computation can be done analytically. And this empirical Bayes approach is considered approximation to the full Bayesian treatment.

Finally, Bayes factor is more desirable than p -values in applications that require comparing multiple non-nested alternative models. One such example is comparing genotype effect, paternal effect, and maternal effect, and paternal-maternal joint effects of the same SNP. To compare two alternative models such as paternal effect and maternal effect, one can simply divide two Bayes factors to obtain a valid Bayes factor between two models [4], while it is difficult to derive a valid p -value (from two p -values) to compare the two models. More details can be found in our manuscript on parent-of-origin effect eQTL [11]. Methods developed here are implemented in a software package IDUL [3].

4 | METHODS

4.1 | Computation

Bayes factor Equation (8) can be evaluated efficiently. This computation was documented previously [8], and we reproduced it here for context

clarity. As $\mathbf{X}^t\mathbf{X} + V_{\mathbf{X}}^{-1} = \begin{pmatrix} \mathbf{I}_n + \frac{1}{\eta\mathbf{D}} & \mathbf{Q}^t\mathbf{W} \\ \mathbf{W}^t\mathbf{Q} & \mathbf{W}^t\mathbf{W} \end{pmatrix}$ and denote

$\mathbf{F} = \left(\mathbf{I}_n + \frac{1}{\eta\mathbf{D}}\right)^{-1}$, $\mathbf{M} = \left(\mathbf{W}^t\mathbf{W} - \mathbf{W}^t\mathbf{Q}\mathbf{F}\mathbf{Q}^t\mathbf{W}\right)^{-1}$, we compute its determinant using the identity

$$\begin{vmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{vmatrix} = |\mathbf{A}| \cdot |\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B}| \text{ to get}$$

$$|\mathbf{X}^t\mathbf{X} + V_{\mathbf{X}}^{-1}| = \left|\mathbf{I}_n + \frac{1}{\eta\mathbf{D}}\right| \cdot |\mathbf{W}^t\mathbf{W} - \mathbf{W}^t\mathbf{Q}\mathbf{F}\mathbf{Q}^t\mathbf{W}|;$$

Using the identities

we get

$$\left(\mathbf{X}^t\mathbf{X} + V_{\mathbf{X}}^{-1}\right)^{-1} = \begin{pmatrix} \mathbf{F} + \mathbf{FQ}^t\mathbf{W}\mathbf{M}\mathbf{W}^t\mathbf{Q}\mathbf{F}, & -\mathbf{FQ}^t\mathbf{W}\mathbf{M} \\ -\mathbf{M}\mathbf{W}^t\mathbf{Q}\mathbf{F}, & \mathbf{M} \end{pmatrix}. \quad (14)$$

Note that for large matrices the computation only involves an inexpensive matrix multiplying vector, without expensive matrix multiplication and matrix inversion. Specifically, in computing $\mathbf{FQ}^t\mathbf{W}$, $\mathbf{Q}^t$ is n by n , $\mathbf{W}$ is n by p with $p \ll n$. So that computing $T = \mathbf{Q}^t\mathbf{W}$ consists of p matrix-vector multiplication. Since $\mathbf{F}$ is a diagonal, computing $\mathbf{F}T$ is just linearly scaling the rows of T . For small matrices such $\mathbf{W}^t\mathbf{W}$ and $\mathbf{M}$, which are p by p , the multiplication and inversion are inexpensive. The only expensive calculation is the eigen decomposition to obtain $\mathbf{Q}$ and $\mathbf{D}$, which only needs to be done once. Similarly, both $|(\mathbf{X}, \mathbf{x})^t(\mathbf{X}, \mathbf{x}) + V_{(\mathbf{X}, \mathbf{x})}^{-1}|$ and $\left((\mathbf{X}, \mathbf{x})^t(\mathbf{X}, \mathbf{x}) + V_{(\mathbf{X}, \mathbf{x})}^{-1}\right)^{-1}$ can be efficiently computed. Finally,

$$\begin{aligned} \begin{pmatrix} \mathbf{b} \\ \mathbf{a} \end{pmatrix} &= \left(\mathbf{X}^t\mathbf{X} + V_{\mathbf{X}}^{-1}\right)^{-1} \cdot \begin{pmatrix} \mathbf{Q}^t\mathbf{y} \\ \mathbf{W}^t\mathbf{y} \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{F} + \mathbf{FQ}^t\mathbf{W}\mathbf{M}\mathbf{W}^t\mathbf{Q}\mathbf{F}, & -\mathbf{FQ}^t\mathbf{W}\mathbf{M} \\ -\mathbf{M}\mathbf{W}^t\mathbf{Q}\mathbf{F}, & \mathbf{M} \end{pmatrix} \cdot \begin{pmatrix} \mathbf{Q}^t\mathbf{y} \\ \mathbf{W}^t\mathbf{y} \end{pmatrix} \quad (15) \\ &= \begin{pmatrix} \mathbf{FQ}^t\mathbf{y} + \mathbf{FQ}^t\mathbf{W}\mathbf{M}\mathbf{W}^t\mathbf{Q}\mathbf{FQ}^t\mathbf{y} & -\mathbf{FQ}^t\mathbf{W}\mathbf{M}\mathbf{W}^t\mathbf{y} \\ -\mathbf{M}\mathbf{W}^t\mathbf{Q}\mathbf{FQ}^t\mathbf{y} & +\mathbf{M}\mathbf{W}^t\mathbf{y} \end{pmatrix}. \end{aligned}$$

4.2 | Least square projection

Let $\mathbf{Q}_j$ be the j -th column of $\mathbf{Q}$, and consider model $\mathbf{Q}_j = \mathbf{G}\beta_j + \epsilon_j$. The least square estimates of $\beta_j = \mathbf{G}^{-1}\mathbf{Q}_j$. For a new observation $\mathbf{g}$, the prediction on the $\mathbf{Q}_j$ is $\mathbf{g}\mathbf{G}^{-1}\mathbf{Q}_j$. Since this is true for each j , we have the least square prediction onto $\mathbf{Q}$ based on genetic relatedness is $\mathbf{g}\mathbf{G}^{-1}\mathbf{Q}$. The same arguments apply to $\mathbf{Z}$ as well.

$$\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{B}(\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{C}\mathbf{A}^{-1}, & -\mathbf{A}^{-1}\mathbf{B}(\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1} \\ -(\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{C}\mathbf{A}^{-1}, & (\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1} \end{pmatrix},$$

4.3 | Equivalence

With the notations defined in the main text, we aim to show $\hat{\mathbf{a}} = \tilde{\mathbf{a}}$ and $\mathbf{Q}\hat{\mathbf{b}} = \mathbf{Z}\tilde{\mathbf{u}}$. With reference to Equation (15) and the fact that $\mathbf{QFQ}^t = \mathbf{I}_n - \mathbf{H}^{-1}$, we have

$$\begin{aligned}\hat{\mathbf{a}} &= -\mathbf{MW}^t\mathbf{QFQ}^t\mathbf{y} + \mathbf{MW}^t\mathbf{y} \\ &= -\mathbf{MW}^t(\mathbf{I}_n - \mathbf{H}^{-1})\mathbf{y} + \mathbf{MW}^t\mathbf{y} \\ &= \mathbf{MW}^t\mathbf{H}^{-1}\mathbf{y} = (\mathbf{W}^t\mathbf{H}^{-1}\mathbf{W})^{-1}\mathbf{W}^t\mathbf{H}^{-1}\mathbf{y} = \tilde{\mathbf{a}},\end{aligned}\quad (16)$$

and

$$\begin{aligned}\mathbf{Q}\hat{\mathbf{b}} &= \mathbf{QFQ}^t\mathbf{y} + \mathbf{QFQ}^t\mathbf{W}\mathbf{M}\mathbf{W}^t\mathbf{QFQ}^t\mathbf{y} - \mathbf{QFQ}^t\mathbf{W}\mathbf{M}\mathbf{W}^t\mathbf{y} \\ &= \mathbf{QFQ}^t\mathbf{y} + \mathbf{QFQ}^t\mathbf{W}\mathbf{M}\mathbf{W}^t(\mathbf{I}_n - \mathbf{H}^{-1})\mathbf{y} \\ &\quad - \mathbf{QFQ}^t\mathbf{W}\mathbf{M}\mathbf{W}^t\mathbf{y} \\ &= \mathbf{QFQ}^t\mathbf{y} - \mathbf{QFQ}^t\mathbf{W}\mathbf{M}\mathbf{W}^t\mathbf{H}^{-1}\mathbf{y} \\ &= (\mathbf{I}_n - \mathbf{H}^{-1})(\mathbf{y} - \mathbf{W}\mathbf{M}\mathbf{W}^t\mathbf{H}^{-1}\mathbf{y}) \\ &= (\mathbf{H} - \mathbf{I}_n)\mathbf{H}^{-1}(\mathbf{y} - \mathbf{W}\tilde{\mathbf{a}}) \\ &= \eta\mathbf{ZKZ}^t\mathbf{H}^{-1}(\mathbf{y} - \mathbf{W}\tilde{\mathbf{a}}) = \mathbf{Z}\tilde{\mathbf{u}}.\end{aligned}\quad (17)$$

AUTHOR CONTRIBUTIONS

Yongtao Guan: Conceptualization, formal analysis, investigation, methodology, software, writing—original draft. **Daniel Levy:** Resources, supervision, writing—review and editing.

ACKNOWLEDGMENTS

This research is supported by the Division of Intramural Research of the National Heart, Lung, and Blood Institute, Bethesda, MD (Daniel Levy is the Principal Investigator).

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

This study analyzed existing datasets from the public domain and simulated datasets. Data from the 1000 Genomes project can be found at IGSR website (internationalgenome.org), and datasets from Framingham Heart Study can be obtained via dbGaP.

REFERENCES

- [1] Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. *Nat Genet.* 2012;44(7):821–4.
- [2] Henderson CR. Best linear unbiased estimation and prediction under a selection model. *Biometrics.* 1975;31(2):423–47.
- [3] Guan Y, Levy D. Asymptotically exact fit for linear mixed model in genetic association studies. *Genetics.* 2024;228(2):iyae143.
- [4] Kass RE, Raftery AE. Bayes factors. *J Am Stat Soc.* 1995; 90(430):773–95.
- [5] Robinson G. That BLUP is a good thing: the estimation of random effects (with discussions). *Stat Sci.* 1991;6(1):15–51.
- [6] Servin B, Stephens M. Imputation-based analysis of association studies: candidate regions and quantitative traits. *PLoS Genet.* 2007;3(7):1–13.
- [7] Zhou Q, Guan Y. On the null distribution of Bayes factors in linear regression. *J Am Stat Soc.* 2018;113(523):1362–71.
- [8] Guan Y, Levy D. Estimation of inbreeding and kinship coefficients via latent identity-by-descent states. *Bioinformatics.* 2024;40(2):btae082.
- [9] Guan Y, Stephens M. Bayesian variable selection regression for genome-wide association studies and other large-scale problems. *Ann Appl Stat.* 2011;5(3):1780–815.
- [10] Wen X. Bayesian model comparison in genetic association analysis: linear mixed modeling and SNP set testing. *Biostatistics.* 2015;16(4):701–12.
- [11] Guan Y, Huan T, Levy D. Abundant parent-of-origin effect EQTL: the Framingham Heart Study. *eLife.* 2025;14:RP107920.

How to cite this article: Guan Y, Levy D. Bayes factor for linear mixed model in genetic association studies. *Quantitative Biology.* 2026; e70036. <https://doi.org/10.1002/qub2.70036>