Asymptotically exact fit for linear mixed model

Yongtao Guan and Daniel Levy
National Heart, Lung, and Blood Institute

March 25, 2024

Abstract

The linear mixed model (LMM) has become a standard in genetic association studies to account for population stratification and relatedness in the samples to reduce false positives. Much recent progresses in LMM focused on approximate computations. Exact methods remained computationally demanding and without theoretical assurance. The computation is particularly challenging for multiomics studies where tens of thousands of phenotypes are tested for association with millions of genetic markers. We present IDUL and IDUL† that use iterative dispersion updates to fit LMMs, where IDUL† is a modified version of IDUL that guarantees likelihood increase between updates. Practically, IDUL and IDUL† produced identical results, both are markedly more efficient than the state-of-the-art Newton-Raphson method, and in particular, both are highly efficient for additional phenotypes, making them ideal to study genetic determinants of multiomics phenotypes. Theoretically, the LMM likelihood is asymptotically uni-modal, and therefore the gradient ascent algorithm IDUL† is an asymptotically exact method. A software package implementing IDUL and IDUL† for genetic association studies is freely available at `https://github.com/haplotype/IDUL`.

1 Introduction

Genome-wide association studies (GWAS) play a pivotal role in identifying genetic variants associated with diverse traits and diseases. A key challenge in these studies is controlling for population stratification and relatedness in the sample, confounding factors, which, if unaddressed, can lead to false-positive associations. To tackle this issue, the linear mixed model (LMM) has emerged as the standard analytical approach. Earlier work focused on feasibility, as exemplified by TASSEL (Yu et al., 2006) and EMMA (Kang et al., 2008). Later works

focused on improving efficiency, as exemplified by EMMAX (Kang et al., 2010), P3D (Zhang et al., 2010), FaST-LMM (Lippert et al., 2011), and GEMMA (Zhou and Stephens, 2012). More recent work, such as BOLT-LMM (Loh et al., 2015) and fastGWA (Jiang et al., 2019), aimed to make the computation feasible for large biobank datasets. One particular setting that requires high efficiency in fitting the linear mixed model is multiomics analysis, where tens of thousands of phenotypes are tested for association with millions of genetic markers.

There are two ways to improve efficiency, one is approximate computation. EMMAX and P3D fit LMM under the null and used parameters estimated from the null model for all SNPs without fitting LMM for each SNP. Svishcheva et al. (2012) approximated a SNP-specific weight with a so-called GRAMMAR-Gamma factor that is shared between SNPs, which effectively performed genomic control internally. Since the factor can be computed efficiently, this approach reduces the computation of score test from quadratic to linear (in sample sizes). BOLT-LMM framed the standard LMM in a Bayesian whole genome regression, and used a variational approximation to fit Bayesian linear regressions with Gaussian mixture priors (Loh et al., 2015). FastGWA (Jiang et al., 2019) combined three approximations: one involves fitting the LMM once under the null and using it for all SNPs; the second adopts the GRAMMAR-Gamma approach in computing score test statistics; and the third uses hard thresholding to make the kinship matrix sparse, which allows fast evaluation of the likelihood function that paves the way for a grid-search method to fit the LMM. All these approximations produced different ranking of test statistics compared to the exact computation and hence a potential power loss (more details below).

Another way to improve efficiency is through algorithmic innovation while maintaining exact computation. (Here *exact* means without aforementioned approximations; it is also short for asymptotically exact, to be discussed below.) Exact computation removes the need to consider which approximate computation works best for a given dataset (Zhou and Stephens, 2012). As an exact method, FaST-LMM first rotates the genotypes and phenotypes according to eigenvectors of the genetic relatedness matrix so that the rotated data become uncorrelated, and then optimizes a single parameter in the variance component using Brent's method. The rotation reduces computation complexity in optimization. Another exact method GEMMA also rotates genotypes and phenotypes during optimization, but it does so implicitly. The innovation of GEMMA is its ability to evaluate the second derivatives, so that the Newton-Raphson method can be used for optimization, which converges faster than Brent's method. The comparison between FaST-LMM and GEMMA can be found in Zhou and Stephens (2012). The Newton-Raphson method suffers from inconsistency when the initial values are distant from the optimum, to overcome this inconsistency, GEMMA starts its optimization iterations with Brent's method and then switches to the Newton-Raphson method.

In this paper, we present IDUL and IDUL† that use an iterative dispersion update to fit LMMs in genetic association studies, where IDUL† is designed to be a gradient ascent algorithm by insisting on a likelihood increase between IDUL updates. We demonstrate that IDUL and IDUL† are consistent and much more efficient than the state-of-the-art Newton-Raphson method in fitting LMMs, and that both are highly efficient for additional phenotypes, and thus well suited to study genetic determinants of multiomics phenotypes. Most importantly, we show that the LMM likelihood is asymptotically unimodal, and consequently IDUL†, a gradient ascent algorithm by design, is asymptotically exact.

2 Results

2.1 The model and the rotation

Consider a standard linear mixed model

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

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, $\mathbf{u}$ is the random effect with $\mathbf{Z}$ as its loading matrix and kinship $\mathbf{K}$ as its covariance (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. Denote $\mathbf{X} = (\mathbf{W}, \mathbf{x})$ and $\mathbf{b} = (\mathbf{a}, \beta)$, then $\mathbf{X}\mathbf{b}$ is the fixed effect, and we assume $\mathbf{X}$ has a full rank c. 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. We aim to find the maximum likelihood estimate (MLE) of η, which is the ratio between two dispersion terms (random effect $\mathbf{u}$ and random noise $\mathbf{e}$), and conditioning on $\hat{\eta}$ we can test the null hypothesis $\beta = 0$.

Denote $\mathbf{G} = \mathbf{Z}\mathbf{K}\mathbf{Z}^t$ and its eigen decomposition $\mathbf{Q}\mathbf{D}\mathbf{Q}^t$ (such that $\mathbf{Q}\mathbf{Q}^t = \mathbf{I}_n$) where j-th column of $\mathbf{Q}$ is an eigenvector whose corresponding eigenvalue is the j-th diagonal element of the diagonal matrix $\mathbf{D}$. Rotate both sides of (1) by multiplying $\mathbf{Q}^t$ to get

$$y_{\mathbf{Q}} \sim \mathrm{MVN}_n(\mathbf{X}_{\mathbf{Q}}\mathbf{b}, \tau^{-1}\mathbf{H}) \tag{2}$$

where $\mathbf{X}_{\mathbf{Q}} = \mathbf{Q}^t\mathbf{X}$, $\mathbf{y}_{\mathbf{Q}} = \mathbf{Q}^t\mathbf{y}$, and $\mathbf{H} = \eta\mathbf{D} + \mathbf{I}_n$ is a diagonal matrix.

2.2 The IDUL algorithm

The iterative dispersion update for linear mixed model (IDUL) algorithm follows:

S0 Initialize η and specify a desired precision threshold ϵ.

S1 For a given η compute $\mathbf{H} = \eta\mathbf{D} + \mathbf{I}_n$, fit (2) using weighted least squares with weight $\mathbf{H}^{-1}$ to obtain residual $\mathbf{r}$, and compute $\hat{t} = \frac{1}{n}\sum_j \mathbf{r}_j^2/\text{Diag}(\mathbf{H})_j$.

S2 Fit $\mathbf{r}^2 \sim \text{MVN}_n(\mu + \text{Diag}(\mathbf{D})\gamma, \tau^{-1}\mathbf{H}^2)$ using weighted least squares with weight $\mathbf{H}^{-2}$ to obtain $\hat{\gamma}$ and $\hat{\mu}$, and compute $\eta^\dagger = \hat{\gamma}/\hat{t} + (1 - \hat{\mu}/\hat{t}\,)\eta$.

S3 If $|\eta^\dagger - \eta| < \epsilon$, goto S4. Otherwise, update $\eta \leftarrow \eta^\dagger$, and goto S1.

S4 Finish and output η.

Intuitively, the update is mostly informed by the different level of dispersion of the residual $\mathbf{r}$, and hence the name of the algorithm. IDUL is easy to implement; both rotation and iterative updates require only several lines of code in R (Appendices).

2.3 Analytic update of the IDUL

IDUL is equivalent to the following analytical update:

$$\eta^\dagger = \eta + \frac{2\eta^2}{nV} f'_{ml}(\eta), \tag{3}$$

where $V = \text{tr}(\mathbf{H}^{-2})/n - \text{tr}(\mathbf{H}^{-1})^2/n^2$ is a function of η and $0 < V < 1$, and $f'_{ml}(\eta)$ is the first derivative of the log-likelihood and can be computed analytically (Data and Methods). Note Equation (3) is derived here to study the analytic properties of the IDUL algorithm, not meant to replace step S2 in the algorithm. We make the following observations on update (3). First, when $f'_{ml}(\eta) > 0$ IDUL increases η, and when $f'_{ml}(\eta) < 0$ IDUL decreases η, until it converges to $f'_{ml}(\eta) = 0$, which is a local optimum. Second, Taylor expansion of the log-likelihood at η to get

$$f_{ml}(\eta^\dagger) - f_{ml}(\eta) = \frac{2\eta^2}{nV} f'_{ml}(\eta)^2 \left(1 + \frac{1}{2} f''_{ml}(\xi)\frac{2\eta^2}{nV}\right).$$

Although the factor $\left(1 + \frac{1}{2} f''_{ml}(\xi)\frac{2\eta^2}{nV}\right)$ is likely to be positive (Data and Methods), there is no guarantee.

2.4 The IDUL† algorithm

We therefore modify the algorithm and only update $\eta \leftarrow \eta^\dagger$ when the likelihood increases, and if likelihood decreases, we successively halve the step size until the likelihood increases. This

technique, successive over-relaxation, is often used in iterative methods (c.f. Zhou and Guan, 2019).

R0 Initialize η and specify a desired precision threshold ϵ.

R1 With input η, compute $\mathbf{H} = \eta^\dagger \mathbf{D} + \mathbf{I}$, fit (2) using weighted least squares with weight $\mathbf{H}^{-1}$ to obtain residual $\mathbf{r}$, compute $\hat{t} = \frac{1}{n}\sum_j \mathbf{r}_j^2/\mathrm{Diag}(\mathbf{H})_j$ and $l(\eta) = -\sum_j \log(\mathrm{Diag}(\mathbf{H})_j) - n\log\hat{t}$.

R2 Fit $\mathbf{r}^2 \sim \mathrm{MVN}_n(\mu + \mathrm{Diag}(\mathbf{D})\gamma, \tau^{-1}\mathbf{H}^2)$ using weighted least squares with weight $\mathbf{H}^{-2}$ to obtain $\hat{\gamma}$ and $\hat{\mu}$, and compute $\eta^\dagger = \hat{\gamma}/\hat{t} + (1 - \hat{\mu}/\hat{t}\,)\eta$.

R3 If $|\eta^\dagger - \eta| < \epsilon$ goto R4; otherwise, do R1 with input $\eta^\dagger$ and obtain $l(\eta^\dagger)$, and if $l(\eta^\dagger) > l(\eta)$, update $\eta \leftarrow \eta^\dagger$, goto R2; otherwise update $\eta^\dagger \leftarrow \frac{1}{2}(\eta^\dagger + \eta)$, goto R3.

R4 Finish and output η.

The IDUL† algorithm is a gradient ascent algorithm by design. Since the likelihood is bounded and the sequence of the likelihood is non-decreasing, by the standard Monotone Convergence Theorem, the IDUL† algorithm must converge to a local optimum η^* such that $f'_{ml}(\eta^*) = 0$. IDUL† is also easy to implement in R (Appendices).

2.5 Asymptotically uni-modal

At an optimum such that $f'_{ml}(\eta) = 0$, the second derivative can be simplified to the following form

$$f''_{ml}(\eta) = \frac{n}{2\eta^2}\left[-V + \epsilon_n\right], \tag{4}$$

where V is defined in (3) and both mean and variance of ϵ_n varnish linearly (proportional to $1/n$) as n increases (details in Data and Methods). In other words, at the local optimum, $f''_{ml}(\eta) < 0$ asymptotically almost sure, or with probability 1. This asymptotically local concaveness implies that with a sufficiently large sample size, the log-likelihood $f_{ml}(\eta)$ attains its unique global maximum at $f'_{ml}(\eta) = 0$. If to the contrary there are at least two local maxima, then owing to smoothness of the likelihood function Equation (5) and its derivative Equation (7), there must exist a minimum η^* such that $f'_{ml}(\eta^*) = 0$ but with $f''_{ml}(\eta^*) > 0$, which produces a contradiction. (Intuitively, there must be a valley between two locally concave peaks, and the valley violates local concaveness.) Therefore, the likelihood function is unimodal asymptotically almost sure (or with probability 1).

2.6 Asymptotically exact

The notion that "an iterative method cannot be exact" (a much appreciated feedback from a reader) is false. For example, the Euclidean algorithm, used to find greatest common divisor between two integers, is an iterative algorithm, and it is exact. Another example is the Banach fixed-point-theorem, which guarantees that, when certain conditions are satisfied, fixed-point iterations always converge to a fixed point, no matter where the iteration starts. A convergence sequence is precise to an arbitrary precision and thus exact. The IDUL† is a gradient ascent algorithm, and if the likelihood is unimodal, then IDUL† updates produce an convergence sequence, and therefore IDUL† is exact. Since the likelihood is asymptotically almost sure unimodal, so IDUL† is asymptotically exact.

2.7 Connection with Newton-Raphson

With a sufficiently large number of samples and η near the optimum, ϵ_n in Equation (4) can be safely ignored. The analytic update of the IDUL then becomes $\eta^\dagger = \eta - f'_{ml}(\eta)/f''_{ml}(\eta)$, which is the Newton-Raphson method. And the IDUL† becomes Newton-Raphson with successive over-relaxation. But IDUL and IDUL† require no computation of the second derivative, which are expensive, and outside the neighborhood of the optimum, where the Newton-Raphson method is known to be numerically unstable (Burden and Faires, 2010), IDUL and IDUL† are stable and consistent (below).

2.8 Consistency of IDUL

Since IDUL† is asymptotically exact, it is consistent over different starting points. We numerically study IDUL's consistency and compare it with that of the Newton-Raphson method. The genotype and phenotype datasets we used for comparison are described in Data and Methods. For each phenotype, we fitted LMM using IDUL and Newton-Raphson with the same sets of initial values. To generate initial values, we chose four non-overlapping segments from the unit interval, namely, $V_1 = (0.01, 0.25), V_2 = (0.25, 0.5), V_3 = (0.5, 0.75)$, and $V_4 = (0.75, 0.99)$, and for each phenotype we drew h_0 uniformly from each segments to produce initial value $\eta_0 = h_0/(1 - h_0)$.

It takes IDUL on average 7.3 iterations to converge for each combination of phenotype and initial value, compared to 12.5 for Newton-Raphson. We compare the consistency of fitted $\hat{\eta}$ when initial values η_0 were drawn from different V segments. IDUL had perfect consistency among four sets of initial values (Figure 1 lower triangle); Newton-Raphson did not (Figure 1 upper triangle). Taking IDUL estimates as the truth, Newton-Raphson made one error when initial values were generated from V_1, 19 errors from V_2, 58 errors from V_3, and

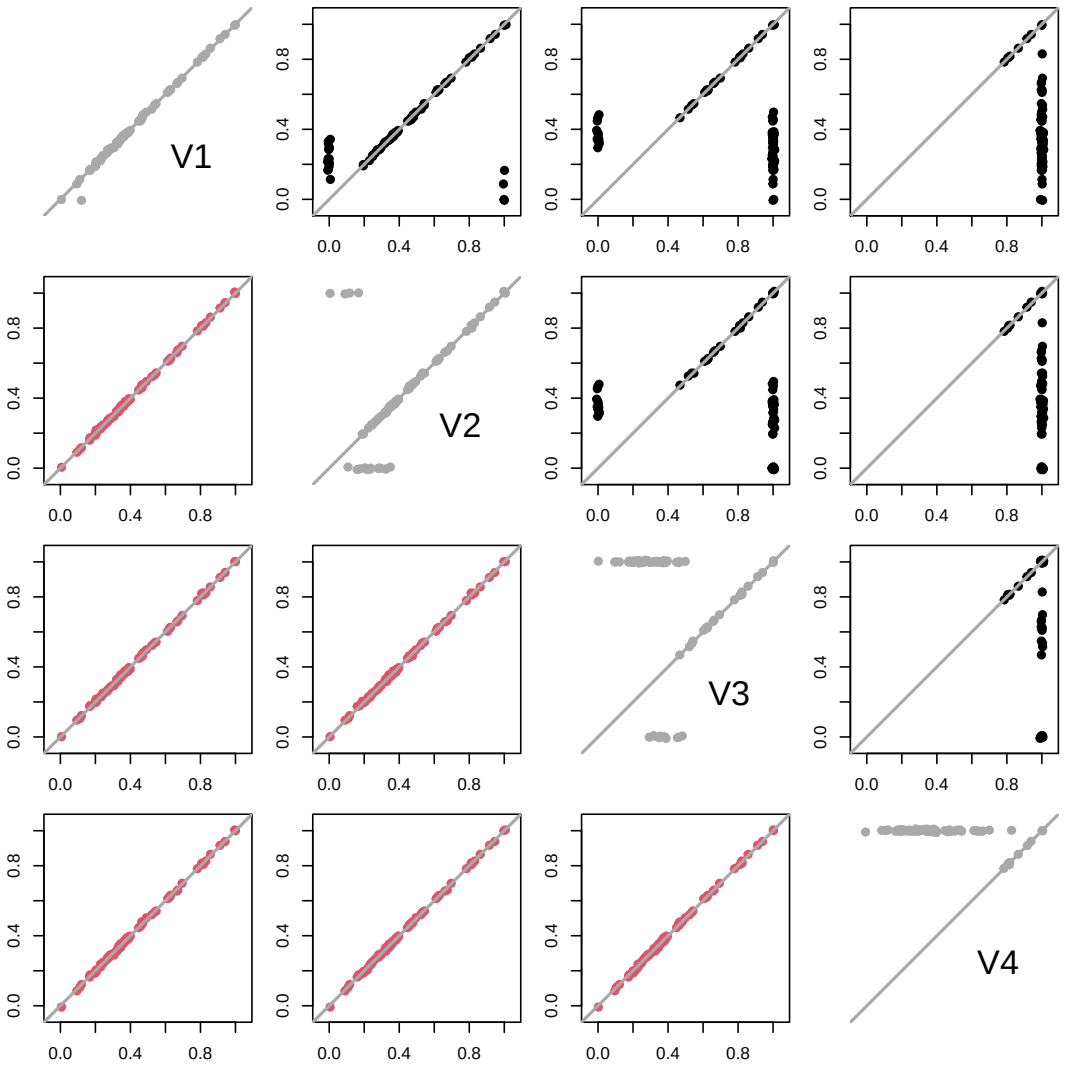

Figure 1: Consistency of IDUL and Newton-Raphson w.r.t different initial values. IDUL and Newton-Raphson were run on the same sets of initial values. For each phenotype, four initial values were generated with seeds randomly selected from four segments. So both IDUL and Newton-Raphson produced 4 columns of estimates each with 79 rows. The pairwise plots of the four columns of IDUL estimates were in the lower-triangle and colored in red. Those of Newton-Raphson estimates were in the upper-triangle and in black. The four diagonal plots (in gray) showed consistency or lack of it between IDUL and Newton-Raphson for different set of initial values. Estimates $\hat{\eta}$ were transformed to $\hat{h} = \hat{\eta}/(1+\hat{\eta})$ for plotting so that different panels are on the same scale. Points are jittered slightly by adding random noises for clarity.

70 errors from V_4 (Figure 1 Diagonal). The erroneous estimates were either 0 or 1 and the proportion of 1 increases as the initial values increase (Figure 1 Diagonal). These observations are consistent with Newton-Raphson's dependence on good initial values and that a long near flat likelihood tend to fail Newton-Raphson. Similar patterns of inconsistency of Newton-Raphson were also observed with phenotypes simulated from diverse populations in the 1000 Genomes project (Supplementary Figure S1). To overcome the inconsistency, implementing the Newton-Raphson method to fit LMM requires multiple runs from different starting points. As a comparison, IDUL and IDUL† only need to run once for each model fitting, each run takes fewer iterations, and each iteration requires less computation.

2.9 Efficiency of IDUL and IDUL†

We implemented IDUL and IDUL† into a software package IDUL that fits the LMMs and computes test statistics, and compared with results from GEMMA, a software package that fits LMMs using the Newton-Raphson method primed by the Brent's method. The datasets we used for comparison were described in Data and Methods. We compared the likelihood ratio test (LRT) and the Wald test p-values between different methods, where the LRT requires maximum likelihood estimates of η and the Wald test requires REML estimates of η (Data and Methods). For both the LRT and the Wald test p-values, IDUL, IDUL†, and GEMMA reached almost perfect agreement (Supplementary Figures S2 and S3). Since IDUL† is asymptotically exact, these results suggested that both IDUL and GEMMA are practically exact methods.

Both IDUL and IDUL† are much more efficient than GEMMA (Table 1). For a single phenotype, GEMMA with maximum 112 threads used about 76.5 minutes for the LRT, compared with 13.6 to 15.3 minutes of IDUL and IDUL† with either 32 and 64 threads. So for LRT, IDUL and IDUL† are at least five times as efficient as GEMMA. GEMMA with maximum 112 threads used 98.9 minutes for Wald test, compared with 13.4 to 15.7 minutes for IDUL and IDUL† with either 32 or 64 threads. So for Wald test, IDUL and IDUL† is six or seven times as efficient as GEMMA.

Most remarkably, IDUL and IDUL† are highly efficient for additional phenotypes. Taking LRT and 64 threads as an example, IDUL and IDUL† only spent about one extra minute (about 8% of time spent for the first phenotype) to compute like ratio test for additional 9 phenotypes, and less than six extra minutes (about 40% of time spent for the first phenotype) for additional 78 phenotypes. Similar high efficiency for additional phenotypes was also observed with the Wald test and with 32 threads (Table 1). Also note that doubling the number of threads used by IDUL/IDUL† resulted in a small improvement in speed for a single phenotype, but a larger improvement with additional phenotypes. The high efficiency for extra phenotypes makes IDUL/IDUL† ideal to study genetic determinants of multiomics phenotypes.

LRT			Methods		Wald		
$p = 1$	10	79	Algo	Threads	$p = 1$	10	79
76.5	-	-	GEMMA	112	98.9	-	-
13.6	14.9	19.3	IDUL	64	13.4	16.8	24.5
13.9	14.7	19.2	IDUL†	64	13.9	16.5	24.6
15.3	16.4	23.8	IDUL	32	15.7	18.8	31.9
14.9	15.9	25.2	IDUL†	32	15.1	19.7	33.0

Table 1: Times (minutes) used for IDUL/IDUL† (version 0.81) and GEMMA (version 0.98.5) to process $1, 10$, and 79 phenotypes. GEMMA used maximum 112 threads and IDUL/IDUL† used 64 and 32. LRT: likelihood ratio test, which requires maximum likelihood estimates of η. Wald: Wald test, which requires REML estimates of η. Taking multiple phenotypes to analyze one by one was not implemented in GEMMA, and its wall time for 10 and 79 phenotypes are missing.

2.10 Exact vs approximation

The IDUL† is an asymptotically exact method. Zhou and Stephens (2012) classified methods into approximate methods and (practically) exact methods and demonstrated that 1) among exact methods available at the time, GEMMA is most efficient, outperforming FaST-LMM, which in turn outperforms EMMA by an order of magnitude; and 2) approximate method such as EMMAX, which uses the parameter estimated under the null to compute test statistics for all SNPs, evidently biased test statistics in some dataset. A recent approximate method, fastGWA by (Jiang et al., 2019), in addition to other approximations, applied hard thresholding on kinship matrix K in Equation (1) to make it sparse and exploited the sparsity in fitting the linear mixed model and computing test statistics. The approximation makes fastGWA capable of analyzing large datasets such as UK Biobank. For multiomics datasets such as that of the Framingham Heart Study, however, the hard thresholding approach appears less satisfactory, presumably due to closer relatedness between samples and larger effect sizes in multiomics dataset. Figure 2 shows a comparison of test statistics of four protein phenotypes between exact computation and hard thresholding approximation. The inconsistency can be rather pronounced for some phenotypes, suggesting potential difficulty with approximate methods in multiomics data for closely related samples such as in the Framingham Heart Study.

3 Discussion

In this paper we documented two novel methods IDUL and IDUL† to fit linear mixed model in the context of genetic association studies. IDUL†, a modification of IDUL, is a gradient

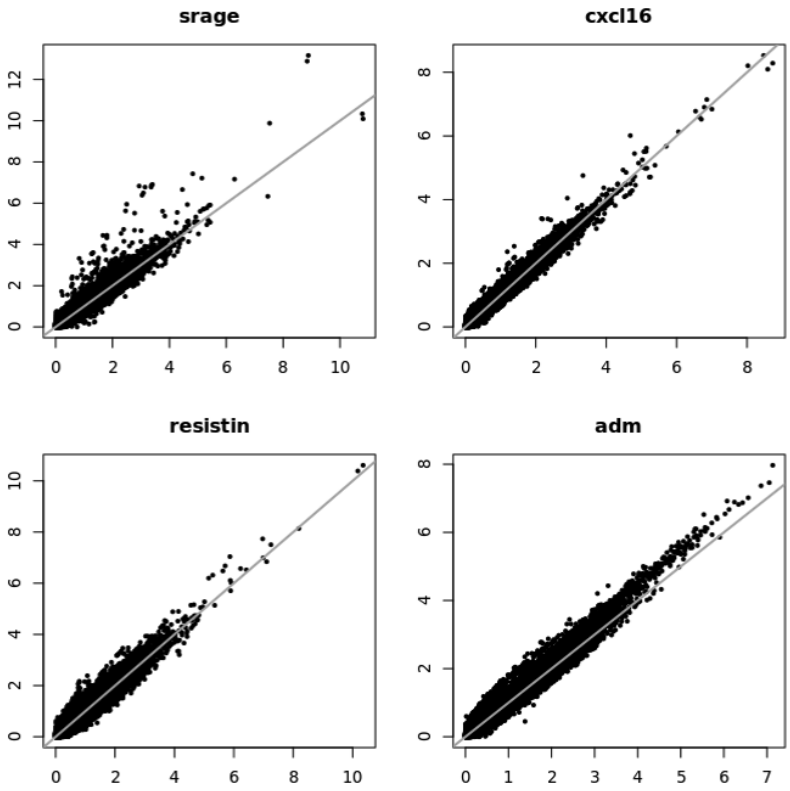

Figure 2: Comparison between exact and approximate method. The test statistics are $-\log_{10}$ p-values, with the exact statistics on x-axis, and the approximate statistics on y-axis. The genomic control values for the exact methods are $0.998, 0.997, 1.013, 1.045$ vs $1.036, 1.007, 1.043, 1.066$ for the approximate method.

ascent algorithm by design. Both IDUL and IDUL† are much more efficiency than the state-of-the-art Newton-Raphson method. The fundamental contribution of the paper, however, is that the log-likelihood of the linear mixed model is asymptotically locally concave at the optimum (we hypothesized that the same is true for REML likelihood, see Data and Methods). Consequently, the likelihood of the standard linear mixed model is asymptotically uni-modal. Therefore, IDUL†, a gradient ascent algorithm , is an asymptotically exact method.

We demonstrated that IDUL and IDUL† are much more efficient than the Newton-Raphson method in fitting the standard LMM. We would like to point out that IDUL and IDUL† are specialized algorithms that take advantage of a specific dispersion structure only available in limited settings such as the LMMs. Newton-Raphson, on the hand, is a general method that can be applied in many settings. In addition, our theoretical analysis relies on the assumption of normality of the phenotypes. Although the assumption can perhaps be relaxed to having finite first and second moments, such as binary phenotypes, it is prudent to examine phenotypes and perform quantile normalization when they show severe departure from normality.

When there are population structures among the samples, such as in the simulation studies using 1000 Genomes datasets, IDUL and IDUL† updates oscillate like a damping pendulum. The algorithms still converge, but the oscillation increases the number of iterations from several to several dozen. This oscillation can be resolved by controlling for leading eigenvectors (such as top three PCs). It can also be resolved by making IDUL and IDUL† lazy. Specifically, the update $\eta^\dagger = \hat{\gamma}/\hat{t}$ has a step size that is a fraction (specifically $\mathrm{tr}(\mathbf{H}^{-1})/n$) of the step size of that IDUL update $\eta^\dagger = \hat{\gamma}/\hat{t} + (1 - \hat{\mu}/\hat{t}\,)\eta$. We can take an average of the two updates to get $\eta^\dagger = \hat{\gamma}/\hat{t} + (1 - \hat{\mu}/\hat{t}\,)\eta/2$ (more details in Appendices), and this lazy update brings oscillation to a quick stop (Supplementary Figure S4).

IDUL is designed with genetic association with multiomics data in mind, where the same set of genotypes are tested against thousands or tens of thousands phenotypes for association, where rotation of genotype vectors (left multiplying Q^t) only needs to be done once for all phenotypes. Suppose in a study we have n sample, m SNPs, p phenotypes, and c covariates, then the total complexity is $O(n^3 + (m + p + c)n^2 + tmp(nc^2 + c^3))$, where $O(n^3)$ is for eigen decomposition, $O((m + p + c)n^2)$ is for rotation, and $O(tmp(nc^2 + c^3))$ is for model fitting of all SNP-phenotype pairs, where $O(nc^2 + c^3)$ is the complexity of linear regression, and we assume IDUL converges on average in t iterations. For a typical study where $m > p > n \gg c$, the dominate term in total complexity is $O(tmpnc^2)$, which is linear in the size of the study, namely, n, m, and p. IDUL† has the same complexity as IDUL, with a slightly larger equivalent t for extra computation to evaluate likelihood. The Newton-Raphson method also has the same complexity, but it tends to have a much larger equivalent t than IDUL and IDUL† because essentially it needs to run multiple times to compensate for its lack of consistency,

each run takes more iterations to converge, and each iteration takes more computation because it requires second derivatives. Table 1 in fact confirmed this intuition.

This strategy of reusing intermediate computation of each genetic variant for multiple phenotypes was also a feature in Regenie (Mbatchou et al., 2021), a whole genome regression (WGR) approach to the linear mixed model that is comparable to BOLT-LMM. Compare to WGR, the standard linear mixed model is more flexible in its applications. For example, in testing for parental origin effect and/or controlling for local ancestry, the standard model only needs to add extra variates and covariates, while the rest of the computation remains the same. But WGR has to change model and priors, and perhaps the details on computation, to make it happen. Strictly speaking, WGR is not a standard linear mixed model. For example, the standard linear mixed model can incorporate different estimates of genetic relatedness matrix such as the one estimated by Kindred (Guan and Levy, 2024), while WGR is stuck with sample correlation as its equivalent genetic relatedness matrix.

In our application, the random effect was treated as a nuisance parameter and our goal is testing fixed effect. Under this context, the MLE is preferred over the REML estimate, because ML estimate of the fixed effects are unbiased (West et al., 2014). REML estimates, however, is preferred when the interest is the variance component, such as in estimating trait heritability, because it produces unbiased estimate of the variance component (i.e, η). IDUL can be adapted to obtain REML estimate based on its analytical update (Data and Methods). The standard software package to estimate heritability is GCTA (Yang et al., 2011), which employs the Average Information REML for model fitting (Gilmour et al., 1995). The software has trouble dealing with modestly small sample sizes. Using an Australian height dataset (Yang et al., 2010), we performed down-sampling at $90\%, 70\%$, and 50% of total 3925 samples 100 times each, obtain REML estimates using IDUL for different estimates of kinship matrices without any issue (Supplementary Figure S5). As a comparison, the Average Information REML implemented in GCTA had difficulty even with 90% down-sampling and produced untenable results.

4 Data and Methods

4.1 Data sets

We used datasets from the Framingham Heart Study (FHS) to conduct numerical comparisons between IDUL and the Newton-Raphson method. Funded by the National Heart, Lung, and Blood Institute (NHLBI), the FHS includes many independent three generational pedigrees, nuclear families, trios, duos, and singletons (Kannel et al., 1979). We used 5757 samples

with whole genome sequencing data through NHLBI's TOPMed program (Taliun et al., 2021) and who also have protein immunoassays obtained through the NHLBI's Systems Approach to Biomarker Research in Cardiovascular Disease (SABRe CVD) Initiative (Ho et al., 2018). With 79 phenotypes, this dataset represents a mini example of multiomics data.

We also used genotype data from the 1000 Genomes project (Auton et al., 2015) with simulated phenotypes to demonstrate the effectiveness of our method for diverse populations. Finally, Australia height data from Queensland Institute of Medical Research was used for down-sampling study to demonstrate the robustness of IDUL with small sample sizes.

4.2 Likelihood and the derivatives

Following notations in Equations (1) and (2), define projections $\mathbf{P}_0 = \mathbf{X_Q}(\mathbf{X_Q^t}\mathbf{H}^{-1}\mathbf{X_Q})^{-1}\mathbf{X_Q^t}\mathbf{H}^{-1}$ and $\mathbf{P} = \mathbf{I}_n - \mathbf{P}_0$. Denote $\mathbf{P}_x = \mathbf{H}^{-1}\mathbf{P}$. The marginal log-likelihood function for η for model 2 is

$$f_{ml}(\eta) = \frac{n}{2}\log(\frac{n}{2\pi}) - \frac{n}{2} - \frac{1}{2}\log|\mathbf{H}| - \frac{n}{2}\log\left(\mathbf{y_Q^t}\mathbf{P}_x\mathbf{y_Q}\right) \tag{5}$$

Because $\mathbf{H}$ is diagonal and Equation 17, $f_{ml}(\eta)$ can be evaluated efficiently. For log-restricted likelihood is

$$\begin{aligned} f_{re}(\eta) = &\frac{n-c}{2}\log(\frac{n-c}{2\pi}) - \frac{n-c}{2} - \frac{1}{2}\log|\mathbf{H}| \\ &- \frac{n-c}{2}\log\left(\mathbf{y_Q^t}\mathbf{P}_x\mathbf{y_Q}\right) + \frac{1}{2}\log|\mathbf{X_Q^t}\mathbf{X_Q}| \\ &- \frac{1}{2}\log|\mathbf{X_Q^t}\mathbf{H}^{-1}\mathbf{X_Q}|. \end{aligned} \tag{6}$$

The first and second derivatives of the log-likelihood function are

$$f'_{ml}(\eta) = -\frac{1}{2}\mathrm{tr}(\mathbf{H}^{-1}\mathbf{D}) + \frac{n}{2}\frac{\mathbf{y_Q^t}\mathbf{P}_x\mathbf{D}\mathbf{P}_x\mathbf{y_Q}}{\mathbf{y_Q^t}\mathbf{P}_x\mathbf{y_Q}}, \tag{7}$$

$$\begin{aligned} f''_{ml}(\eta) = &\frac{1}{2}\mathrm{tr}(\mathbf{H}^{-1}\mathbf{D}\mathbf{H}^{-1}\mathbf{D}) \\ &- \frac{n}{2}\frac{(2\mathbf{y_Q^t}\mathbf{P}_x\mathbf{D}\mathbf{P}_x\mathbf{D}\mathbf{P}_x\mathbf{y_Q})(\mathbf{y}_Q^t\mathbf{P}_x\mathbf{y_Q}) - \left(\mathbf{y_Q^t}\mathbf{P}_x\mathbf{D}\mathbf{P}_x\mathbf{y_Q}\right)^2}{(\mathbf{y_Q^t}\mathbf{P}_x\mathbf{y_Q})^2}. \end{aligned} \tag{8}$$

Since we work with the rotated system (2), the only matrix calculus identity needed to derive these is $\frac{dA^{-1}}{dx} = -A^{-1}\frac{dA}{dx}A^{-1}$ where matrix A is a function of a scalar x. For log-restricted likelihood we have

$$f'_{re}(\eta) = -\frac{1}{2}\mathrm{tr}(\mathbf{P}_x\mathbf{D}) + \frac{n-c}{2}\frac{\mathbf{y_Q^t}\mathbf{P}_x\mathbf{D}\mathbf{P}_x\mathbf{y_Q}}{\mathbf{y_Q^t}\mathbf{P}_x\mathbf{y_Q}}, \tag{9}$$

$$f''_{re}(\eta) = \frac{1}{2}\mathrm{tr}(\mathbf{P}_x\mathbf{D}\mathbf{P}_x\mathbf{D}) - \frac{n-c}{2}\frac{(2\mathbf{y}_\mathbf{Q}^t\mathbf{P}_x\mathbf{D}\mathbf{P}_x\mathbf{D}\mathbf{P}_x\mathbf{y}_\mathbf{Q})(\mathbf{y}_Q^t\mathbf{P}_x\mathbf{y}_\mathbf{Q}) - \left(\mathbf{y}_\mathbf{Q}^t\mathbf{P}_x\mathbf{D}\mathbf{P}_x\mathbf{y}_\mathbf{Q}\right)^2}{(\mathbf{y}_\mathbf{Q}^t\mathbf{P}_x\mathbf{y}_\mathbf{Q})^2}. \tag{10}$$

These likelihood and derivatives are in the same form as those in (Zhou and Stephens, 2012).

4.3 Evaluation of the first derivatives

We simplify the directive of likelihood functions using residuals from S1 of the IDUL.

Proposition 1. *Let* $\mathbf{r} = \mathbf{P}\mathbf{y}_\mathbf{Q}$, *the following hold:*

$$\begin{aligned} f'_{ml}(\eta) &= \frac{1}{2\eta}\left(tr(\mathbf{H}^{-1}) - n\frac{\mathbf{r}^t\mathbf{H}^{-2}\mathbf{r}}{\mathbf{r}^t\mathbf{H}^{-1}\mathbf{r}}\right), \\ f'_{re}(\eta) &= \frac{1}{2\eta}\left(tr(\mathbf{H}^{-1}\mathbf{P}) - (n-c)\frac{\mathbf{r}^t\mathbf{H}^{-2}\mathbf{r}}{\mathbf{r}^t\mathbf{H}^{-1}\mathbf{r}}\right). \end{aligned} \tag{11}$$

Proof of Proposition 1 is deferred to Appdendices. Because $\mathbf{H}$ is diagonal, evaluation of f'_{ml} has a complexity of $O(n)$. To evaluate f'_{re} note $\mathrm{tr}(\mathbf{H}^{-1}\mathbf{P}) = \mathrm{tr}(\mathbf{H}) - \mathrm{tr}(\mathbf{H}\mathbf{P}_0)$, while $\mathrm{tr}(\mathbf{H}\mathbf{P}_0) = \mathrm{tr}(\mathbf{H}^{-1}\mathbf{X}(\mathbf{X}^t\mathbf{H}^{-1}\mathbf{X})^{-1}\mathbf{X}^t\mathbf{H}^{-1}) = \mathrm{tr}(\frac{\mathbf{X}^t\mathbf{H}^{-2}\mathbf{X}}{\mathbf{X}^t\mathbf{H}^{-1}\mathbf{X}})$. Since $\mathbf{X}^t\mathbf{H}^{-1}\mathbf{X}$ is $c \times c$, its inverse has complexity $O(c^3)$. Therefore the complexity of evaluating f'_{re} is $O(n + nc + c^3)$.

4.4 IDUL update in an analytical form

To study the theoretical property of IDUL, we derive its analytic form by computing $\eta^\dagger = \hat{\gamma}/\hat{t} + (1 - \hat{\mu}/\hat{t}\,)\eta$ in S2 of the IDUL.

Lemma 2. *The update of IDUL for maximum likelihood estimate is*

$$\eta^\dagger = \eta + \frac{2\eta^2}{nV}f'_{ml}(\eta), \tag{12}$$

where $V = tr(\mathbf{H}^{-2})/n - tr(\mathbf{H}^{-1})^2/n^2 > 0$ *and is bounded, and* $f'_{ml}(\eta)$ *is the derivative of the log-likelihood evaluated at* η.

Proof of Lemma 2 is deferred to Appendices. The Taylor expansion of f_{ml} at η is $f_{ml}(\eta + x) = f_{ml}(\eta) + f'_{ml}(\eta)x + \frac{1}{2}f''_{ml}(\xi)x^2$, for some $\xi \in (\eta - x, \eta + x)$. Let $x = \eta^\dagger - \eta$, we get $f_{ml}(\eta^\dagger) - f_{ml}(\eta) = \frac{2\eta^2}{nV}{f'_{ml}}^2(\eta) + \frac{1}{2}f''_{ml}(\xi)|\eta^\dagger - \eta|^2$. Substitute $\eta^\dagger - \eta = \frac{2\eta^2}{nV}f'_{ml}(\eta)$ back in, we have

$$f_{ml}(\eta^\dagger) - f_{ml}(\eta) = \frac{2\eta^2}{nV}{f'_{ml}}^2(\eta)\left(1 + \frac{1}{2}f''_{ml}(\xi)\frac{2\eta^2}{nV}\right). \tag{13}$$

When η is near the optimal, we have $f_{ml}(\eta^\dagger) - f_{ml}(\eta) > 0$ for large n by virtue of Equation 15. But when η is not near optimal, there is no guarantee that the likelihood always increase. Thus we modify IDUL by evaluating and comparing likelihood to guarantee the likelihood increase between updates.

4.5 From MLE to REML

With maximum likelihood update (12) at hand, we can obtain REML estimate by substituting S2 in IDUL and R2 in IDUL† with

$$\eta^\dagger = \eta + \frac{2\eta^2}{nV}\, f'_{re}(\eta). \tag{14}$$

Of course, for IDUL† the likelihood in R1 needs to be revised to REML likelihood.

4.6 Asymptotically locally concave at optimum

Finally, we quantify the second derivative of the log-likelihood function to show it is asymptotically locally concave at the local optimum.

Theorem 3. *Let η be an optimum of log-likelihood function such that $f'_{ml}(\eta) = 0$, then*

$$f''_{ml}(\eta) = \frac{n}{2\eta^2}\left[-V + \epsilon_n\right], \tag{15}$$

where $V = tr(\mathbf{H}^{-2})/n - tr(\mathbf{H}^{-1})^2/n^2 > 0$ and is bounded, and both mean and variance of ϵ_n decreases linearly $O(\frac{1}{n})$. Thus, at a local optimum such that $f'_{ml}(\eta) = 0$, we have $f''_{ml}(\eta) < 0$ asymptotically almost sure.

Proofs of Theorem 3 is deferred to Appendices. Owing to the similarity of expression between $f''_{re}(\eta)$ and $f''_{ml}(\eta)$, we believe the following conjecture can be proved by exploring the connections between eigenvalues of $\mathbf{H}^{-1}\mathbf{P}$ and $\mathbf{H}^{-1}$, evidenced by that $\mathbf{H}^{-1}\mathbf{P}\mathbf{P}\mathbf{v} = \lambda\mathbf{P}\mathbf{v}$ implies $\mathbf{H}^{-1}\mathbf{P}\mathbf{v} = \lambda\mathbf{P}\mathbf{v}$.

Corollary 4. *Let η be an optimum of REML-likelihood function such that $f'_{re}(\eta) = 0$, then*

$$f''_{re}(\eta) = \frac{n-c}{2\eta^2}\left[-V + \epsilon_n\right], \tag{16}$$

where $V = tr(\mathbf{H}^{-1}\mathbf{P}\mathbf{H}^{-1}\mathbf{P})/(n-c) - tr(\mathbf{H}^{-1}\mathbf{P})^2/(n-c)^2 > 0$, and both mean and variance of ϵ_n decreases linearly $O(\frac{1}{n})$. Thus, $f''_{re}(\eta) < 0$ almost sure, or with probability 1, for a sufficiently large n.

4.7 Computing association p-values

With maximum likelihood estimate and REML estimate of η, we can compute p-values for association. To test the null hypothesis $\beta = 0$, we computed the likelihood ratio test (LRT) p-values using maximum likelihood estimates as suggested by (Yu et al., 2006) and Wald test p-values using REML estimates as suggested by (Kang et al., 2008). Both test statics were described in clean detail in Supplementary of (Zhou and Stephens, 2012).

5 Acknowledgements

This research is supported by the Division of Intramural Research of the National Heart, Lung, and Blood Institute, Bethesda, MD (D.L. Principal Investigator). We thank Nick Martin and the Queensland Institute of Medical Research for making the Australia height data available to us.

6 Author contributions

Y.G. conceived the study, developed the methodology, implemented the computational software and performed experiments, analyzed results, and wrote the manuscript. D.L. supervised the work, provided access to data and computation, edited and approved the final manuscript.

7 Declaration of interests

The authors declare no competing interests.

References

Auton, A., G. R. Abecasis, D. M. Altshuler, R. M. Durbin, D. R. Bentley, A. Chakravarti, A. G. Clark, and et al. (2015). A global reference for human genetic variation. *Nature 526*(7571), 68–74.

Burden, R. L. and J. D. Faires (2010). *Numerical Analysis*. Cengage Learning; 9th edition.

Gilmour, A. R., R. Thompson, and B. R. Cullis (1995). Average information reml: An efficient algorithm for variance parameter estimation in linear mixed models. *Biometrics 51*(4), 1440–1450.

Graham, M. M. and A. J. Storkey (2017). Asymptotically exact inference in differentiable generative models. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Guan, Y. and D. Levy (2024, 02). Estimation of inbreeding and kinship coefficients via latent identity-by-descent states. *Bioinformatics 40*(2), btae082.

Ho, J. E., A. Lyass, P. Courchesne, G. Chen, C. Liu, X. Yin, S.-J. Hwang, J. M. Massaro, M. G. Larson, and D. Levy (2018, July). Protein biomarkers of cardiovascular disease and mortality in the community. *J. Am. Heart Assoc. 7*(14).

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

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

Kang, H. M., N. A. Zaitlen, C. M. Wade, A. Kirby, D. Heckerman, M. J. Daly, and E. Eskin (2008, 03). Efficient Control of Population Structure in Model Organism Association Mapping. *Genetics 178*(3), 1709–1723.

Kannel, W. B., M. Feinleib, P. M. McNamara, R. J. Garrison, and W. P. Castelli (1979, September). An investigation of coronary heart disease in families. the framingham offspring study. *Am. J. Epidemiol. 110*(3), 281–290.

Lippert, C., J. Listgarten, Y. Liu, C. M. Kadie, R. I. Davidson, and D. Heckerman (2011, Oct). Fast linear mixed models for genome-wide association studies. *Nature Methods 8*(10), 833–835.

Loh, P.-R., G. Tucker, B. K. Bulik-Sullivan, B. J. Vilhjálmsson, H. K. Finucane, R. M. Salem, D. I. Chasman, P. M. Ridker, B. M. Neale, B. Berger, N. Patterson, and A. L. Price (2015, Mar). Efficient bayesian mixed-model analysis increases association power in large cohorts. *Nature Genetics 47*(3), 284–290.

Mbatchou, J., L. Barnard, J. Backman, A. Marcketta, J. A. Kosmicki, A. Ziyatdinov, C. Benner, C. O'Dushlaine, M. Barber, B. Boutkov, L. Habegger, M. Ferreira, A. Baras, J. Reid, G. Abecasis, E. Maxwell, and J. Marchini (2021, Jul). Computationally efficient whole-genome regression for quantitative and binary traits. *Nature Genetics 53*(7), 1097–1103.

Svishcheva, G. R., T. I. Axenovich, N. M. Belonogova, C. M. van Duijn, and Y. S. Aulchenko (2012). Rapid variance components–based method for whole-genome association analysis. *Nature Genetics 44*(10), 1166–1170.

Taliun, D., D. N. Harris, M. D. Kessler, J. Carlson, Z. A. Szpiech, R. Torres, and et al. (2021, February). Sequencing of 53,831 diverse genomes from the NHLBI TOPMed program. *Nature 590*(7845), 290–299.

West, B. T., K. B. Welch, and A. T. Galecki (2014). *Linear mixed models: A practical guide using statistical software*. Chapman and Hall/CRC, 2nd edition.

Yang, J., B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders, D. R. Nyholt, P. A. Madden, A. C. Heath, N. G. Martin, G. W. Montgomery, M. E. Goddard, and P. M. Visscher (2010, June). Common SNPs explain a large proportion of the heritability for human height. *Nat Genet 42*(7), 565–569.

Yang, J., S. H. Lee, M. E. Goddard, and P. M. Visscher (2011, 2022/12/25). Gcta: A tool for genome-wide complex trait analysis. *The American Journal of Human Genetics 88*(1), 76–82.

Yu, J., G. Pressoir, W. H. Briggs, I. Vroh Bi, M. Yamasaki, J. F. Doebley, M. D. McMullen, B. S. Gaut, D. M. Nielsen, J. B. Holland, S. Kresovich, and E. S. Buckler (2006, Feb). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics 38*(2), 203–208.

Zhang, Z., E. Ersoz, C.-Q. Lai, R. J. Todhunter, H. K. Tiwari, M. A. Gore, P. J. Bradbury, J. Yu, D. K. Arnett, J. M. Ordovas, and E. S. Buckler (2010, Apr). Mixed linear model approach adapted for genome-wide association studies. *Nature Genetics 42*(4), 355–360.

Zhou, Q. and Y. Guan (2019). Fast Model-Fitting of Bayesian Variable Selection Regression Using the Iterative Complex Factorization Algorithm. *Bayesian Analysis 14*(2), 573 – 594.

Zhou, X. and M. Stephens (2012, June). Genome-wide efficient mixed-model analysis for association studies. *Nat Genet 44*(7), 821–824.

8 Appendices

8.1 R code for IDUL and IDUL†

```
eG=eigen(G);    #G=ZKZ^t;
D=ifelse(eG$values < 0, 0, eG$values);
xQ=t(eG$vectors) %*% cbind(W, x);  #rotate covariates W and genotype x;
yQ=t(eG$vectors) %*% y;    #rotate phenoytpe y;

idul=function(xQ,yQ,D,eta,epsilon) {
    repeat {
       H = eta * D + 1;
       r2 = lm(yQ~xQ, weights=1/H)$residuals^2;
       tauinv = mean(r2/H);
       fit2 = lm(r2~D, weights=1/H/H);
       param=fit2$coefficients;
       eta1 = max(0,param[2]/tauinv+(1-param[1]/tauinv)*eta);
       print(c(eta,eta1),digits=6);
       if(abs(eta1-eta) < epsilon) {break;}
       eta = eta1;
    };
    return(eta);
}
```

```
idul_plus=function(xQ,yQ,D,eta,epsilon) {
    H = eta * D + 1;
    r2 = lm(yQ~xQ, weights=1/H)$residuals^2;
    tauinv = mean(r2/H);
    like = -sum(log(H)) - length(D) *log(tauinv);
    repeat {
       fit2 = lm(r2~D, weights=1/H/H);
       param=fit2$coefficients;
       eta1 = max(0, param[2]/tauinv+(1-param[1]/tauinv)*eta);
       while(abs(eta1-eta)>epsilon){
           H1= eta1 * D + 1;
           r2 = lm(yQ~xQ, weights=1/H1)$residuals^2;
           tauinv1 = mean(r2/H1);
           like1 = -sum(log(H1)) - length(D)*log(tauinv1);
           if(like1 >= like) {break;}
           eta1 = (eta1+eta)/2;
       }
       print(c(eta,eta1, like, like1));
       if(abs(eta1-eta) < epsilon) {break;}
       eta = eta1; H=H1; like=like1; tauinv=tauinv1;
    };
    return(eta);
}
```

8.2 Proof of Proposition 1

Proof. We first simplify two expressions

$$\begin{aligned} \mathbf{y_Q^t P_x y_Q} &= \mathbf{r}^t\mathbf{H}^{-1}\mathbf{r} \\ \mathbf{y_Q^t P_x D P_x y_Q} &= \frac{1}{\eta}(\mathbf{r}^t\mathbf{H}^{-1}\mathbf{r} - \mathbf{r}^t\mathbf{H}^{-2}\mathbf{r}) \end{aligned} \tag{17}$$

1) By $\mathbf{PP} = \mathbf{P}$ and $\mathbf{r} = \mathbf{P y_Q}$, we have $\mathbf{y_Q^t P_x} \mathbf{y}_Q = \mathbf{y_Q^t H^{-1} P y_Q} = \mathbf{y_Q^t H^{-1} PP y_Q} = \mathbf{y_Q^t P^t H^{-1} P y_Q} = \mathbf{r}^t\mathbf{H}^{-1}\mathbf{r}$. 2) Recall $\mathbf{H} = \eta\mathbf{D} + \mathbf{I}_n$, so $\mathbf{D} = \frac{1}{\eta}(\mathbf{H} - \mathbf{I}_n)$, then by direct computation we have $\mathbf{y_Q^t P_x D P_x y_Q} = \mathbf{y_Q^t P^t H^{-1} D H^{-1} P y_Q} = \mathbf{r}^t\mathbf{H}^{-1}\mathbf{D}\mathbf{H}^{-1}\mathbf{r} = \frac{1}{\eta}\mathbf{r}^t\mathbf{H}^{-1}(\mathbf{H} - \mathbf{I}_n)\mathbf{H}^{-1}\mathbf{r} = \frac{1}{\eta}(\mathbf{r}^t\mathbf{H}^{-1}\mathbf{r} - \mathbf{r}^t\mathbf{H}^{-2}\mathbf{r})$. With these two reduced expressions, the first derivatives can

be transformed as following:

$$\begin{aligned} f'_{ml}(\eta) &= -\frac{1}{2}\mathrm{tr}(\mathbf{H}^{-1}\mathbf{D}) + \frac{n}{2}\frac{\mathbf{y}_\mathbf{Q}^t\mathbf{P}_x\mathbf{D}\mathbf{P}_x\mathbf{y}_\mathbf{Q}}{\mathbf{y}_\mathbf{Q}^t\mathbf{P}_x\mathbf{y}_\mathbf{Q}} \\ &= -\frac{1}{2}\frac{1}{\eta}\mathrm{tr}(\mathbf{H}^{-1}(\mathbf{H}-\mathbf{I}_n)) + \frac{n}{2}\frac{1}{\eta}\frac{\mathbf{r}^t\mathbf{H}^{-1}\mathbf{r} - \mathbf{r}^t\mathbf{H}^{-2}\mathbf{r}}{\mathbf{r}^t\mathbf{H}^{-1}\mathbf{r}} \\ &= \frac{1}{2\eta}(-n + \mathrm{tr}(\mathbf{H}^{-1}) + n - n\frac{\mathbf{r}^t\mathbf{H}^{-2}\mathbf{r}}{\mathbf{r}^t\mathbf{H}^{-1}\mathbf{r}}) \\ &= \frac{1}{2\eta}\left(\mathrm{tr}(\mathbf{H}^{-1}) - n\frac{\mathbf{r}^t\mathbf{H}^{-2}\mathbf{r}}{\mathbf{r}^t\mathbf{H}^{-1}\mathbf{r}}\right), \end{aligned} \tag{18}$$

and

$$\begin{aligned} f'_{re}(\eta) &= -\frac{1}{2}\mathrm{tr}(\mathbf{P}_x\mathbf{D}) + \frac{n-c}{2}\frac{\mathbf{y}_\mathbf{Q}^t\mathbf{P}_x\mathbf{D}\mathbf{P}_x\mathbf{y}_\mathbf{Q}}{\mathbf{y}_\mathbf{Q}^t\mathbf{P}_x\mathbf{y}_\mathbf{Q}} \\ &= -\frac{1}{2}\frac{1}{\eta}\mathrm{tr}(\mathbf{P}^t\mathbf{H}^{-1}(\mathbf{H}-\mathbf{I}_n)) + \frac{n-c}{2}\frac{1}{\eta}\frac{\mathbf{r}^t\mathbf{H}^{-1}\mathbf{r} - \mathbf{r}^t\mathbf{H}^{-2}\mathbf{r}}{\mathbf{r}^t\mathbf{H}^{-1}\mathbf{r}} \\ &= \frac{1}{2\eta}(\mathrm{tr}(\mathbf{P}) + \mathrm{tr}(\mathbf{H}^{-1}\mathbf{P}) + (n-c) - (n-c)\frac{\mathbf{r}^t\mathbf{H}^{-2}\mathbf{r}}{\mathbf{r}^t\mathbf{H}^{-1}\mathbf{r}}) \\ &= \frac{1}{2\eta}\left(\mathrm{tr}(\mathbf{H}^{-1}\mathbf{P}) - (n-c)\frac{\mathbf{r}^t\mathbf{H}^{-2}\mathbf{r}}{\mathbf{r}^t\mathbf{H}^{-1}\mathbf{r}}\right), \end{aligned} \tag{19}$$

where the last equality holds because $\mathrm{tr}(\mathbf{P}) = \mathrm{tr}(\mathbf{I}_n) - \mathrm{tr}(\mathbf{P}_0)$ and $\mathbf{P}_0$ is a projection with rank c thus $\mathrm{tr}(\mathbf{P}_0) = c$.

□

8.3 Proof of Lemma 2

Proof. Step 1 is a weighed linear regression, we can compute $\mathbf{r} = \mathbf{P}\mathbf{y}_\mathbf{Q}$. Step 2 is also a weighted linear regression with two covariates, so that its solution can be directly computed. Let vector $\mathbf{d}$ be the diagonal elements of $\mathbf{D}$ and $\mathbf{1}$ is the vector of 1 and $\mathbf{s}$ is component wise square of $\mathbf{r}$, we have

$$\begin{aligned} (\hat{\gamma}, \hat{\mu})^t &= ((\mathbf{d}, \mathbf{1})^t\mathbf{H}^{-2}(\mathbf{d}, \mathbf{1}))^{-1}(\mathbf{d}, \mathbf{1})^t\mathbf{H}^{-2}\mathbf{s} \\ &= \begin{pmatrix} \mathbf{d}^t\mathbf{H}^{-2}\mathbf{d} & \mathbf{d}^t\mathbf{H}^{-2}\mathbf{1} \\ \mathbf{1}^t\mathbf{H}^{-2}\mathbf{d} & \mathbf{1}^t\mathbf{H}^{-2}\mathbf{1} \end{pmatrix}^{-1} (\mathbf{d}, \mathbf{1})^t\,\mathbf{H}^{-2}\mathbf{s} \\ &= \frac{1}{\Delta}\begin{pmatrix} \mathbf{1}^t\mathbf{H}^{-2}\mathbf{1} & -\mathbf{d}^t\mathbf{H}^{-2}\mathbf{1} \\ -\mathbf{1}^t\mathbf{H}^{-2}\mathbf{d} & \mathbf{d}^t\mathbf{H}^{-2}\mathbf{d} \end{pmatrix} (\mathbf{d}^t\mathbf{H}^{-2}\mathbf{s}, \mathbf{1}^t\mathbf{H}^{-2}\mathbf{s})^t \\ &= \frac{1}{\Delta}\begin{pmatrix} \mathbf{1}^t\mathbf{H}^{-2}\mathbf{1}\cdot\mathbf{d}^t\mathbf{H}^{-2}\mathbf{s} - \mathbf{d}^t\mathbf{H}^{-2}\mathbf{1}\cdot\mathbf{1}^t\mathbf{H}^{-2}\mathbf{s} \\ -\mathbf{1}^t\mathbf{H}^{-2}\mathbf{d}\cdot\mathbf{d}^t\mathbf{H}^{-2}\mathbf{s} + \mathbf{d}^t\mathbf{H}^{-2}\mathbf{d}\cdot\mathbf{1}^t\mathbf{H}^{-2}\mathbf{s} \end{pmatrix}. \end{aligned} \tag{20}$$

Since $\mathbf{H}$ and $\mathbf{D}$ are diagonal, we have

$$\begin{aligned}
\hat{\gamma} &= \frac{1}{\Delta}\left(\mathbf{1}^t\mathbf{H}^{-2}\mathbf{1}\cdot\mathbf{d}^t\mathbf{H}^{-2}\mathbf{s} - \mathbf{d}^t\mathbf{H}^{-2}\mathbf{1}\cdot\mathbf{1}^t\mathbf{H}^{-2}\mathbf{s}\right)\\
&= \frac{1}{\Delta}\left(\mathrm{tr}(\mathbf{H}^{-2})\cdot\mathbf{r}^t\mathbf{D}\mathbf{H}^{-2}\mathbf{r} - \mathrm{tr}(\mathbf{D}\mathbf{H}^{-2})\mathbf{r}^t\mathbf{H}^{-2}\mathbf{r}\right)\\
&= \frac{1}{\Delta}\frac{1}{\eta}\left(\mathrm{tr}(\mathbf{H}^{-2})\cdot\mathbf{r}^t(\mathbf{H}-\mathbf{I}_n)\mathbf{H}^{-2}\mathbf{r} - \mathrm{tr}((\mathbf{H}-\mathbf{I}_n)\mathbf{H}^{-2})\mathbf{r}^t\mathbf{H}^{-2}\mathbf{r}\right)\\
&= \frac{1}{\Delta}\frac{1}{\eta}\left(\mathrm{tr}(\mathbf{H}^{-2})\cdot\mathbf{r}^t\mathbf{H}^{-1}\mathbf{r} - \mathrm{tr}(\mathbf{H}^{-2})\cdot\mathbf{r}^t\mathbf{H}^{-2}\mathbf{r} - \mathrm{tr}(\mathbf{H}^{-1})\mathbf{r}^t\mathbf{H}^{-2}\mathbf{r} + \mathrm{tr}(\mathbf{H}^{-2})\mathbf{r}^t\mathbf{H}^{-2}\mathbf{r}\right)\\
&= \frac{1}{\Delta}\frac{1}{\eta}\frac{\mathbf{r}^t\mathbf{H}^{-1}\mathbf{r}}{n}\left(n\,\mathrm{tr}(\mathbf{H}^{-2}) - \mathrm{tr}(\mathbf{H}^{-1})\,n\,\frac{\mathbf{r}^t\mathbf{H}^{-2}\mathbf{r}}{\mathbf{r}^t\mathbf{H}^{-1}\mathbf{r}}\right),
\end{aligned}$$

and

$$\begin{aligned}
\hat{\mu} =& \frac{1}{\Delta}\left(-\mathbf{1}^t\mathbf{D}\mathbf{H}^{-2}\mathbf{1}\cdot\mathbf{r}^t\mathbf{D}\mathbf{H}^{-2}\mathbf{r} + \mathbf{1}^t\mathbf{D}^2\mathbf{H}^{-2}\mathbf{1}\cdot\mathbf{r}^t\mathbf{H}^{-2}\mathbf{r}\right)\\
=& \frac{1}{\Delta}\frac{1}{\eta^2}\left(-\mathbf{1}^t\mathbf{H}^{-2}(\mathbf{H}-\mathbf{I})\mathbf{1}\cdot\mathbf{r}^t(\mathbf{H}-\mathbf{I})\mathbf{H}^{-2}\mathbf{r} + \mathbf{1}^t(\mathbf{H}-\mathbf{I})^2\mathbf{H}^{-2}\mathbf{1}\cdot\mathbf{r}^t\mathbf{H}^{-2}\mathbf{r}\right)\\
=& \frac{1}{\Delta}\frac{1}{\eta^2}\left(-\mathrm{tr}(\mathbf{H}^{-1}-\mathbf{H}^{-2})\cdot\mathbf{r}^t(\mathbf{H}^{-1}-\mathbf{H}^{-2})\mathbf{r} + \mathrm{tr}(\mathbf{I}_n - 2\mathbf{H}^{-1} + \mathbf{H}^{-2})\cdot\mathbf{r}^t\mathbf{H}^{-2}\mathbf{r}\right)\\
=& \frac{1}{\Delta}\frac{1}{\eta^2}\left(-\mathrm{tr}(\mathbf{H}^{-1})\mathbf{r}^t\mathbf{H}^{-1}\mathbf{r} + n\mathbf{r}^t\mathbf{H}^{-2}\mathbf{r} + \mathrm{tr}(\mathbf{H}^{-2})\mathbf{r}^t\mathbf{H}^{-1}\mathbf{r} - \mathrm{tr}(\mathbf{H}^{-1})\mathbf{r}^t\mathbf{H}^{-2}\mathbf{r}\right)\\
=& \frac{1}{\Delta}\frac{1}{\eta^2}\frac{\mathbf{r}^t\mathbf{H}^{-1}\mathbf{r}}{n}\left[n\,\mathrm{tr}(\mathbf{H}^{-2}) - \mathrm{tr}(\mathbf{H}^{-1})\,n\frac{\mathbf{r}^t\mathbf{H}^{-2}\mathbf{r}}{\mathbf{r}^t\mathbf{H}^{-1}\mathbf{r}} - n\left(\mathrm{tr}(\mathbf{H}^{-1}) - n\frac{\mathbf{r}^t\mathbf{H}^{-2}\mathbf{r}}{\mathbf{r}^t\mathbf{H}^{-1}\mathbf{r}}\right)\right],
\end{aligned} \tag{21}$$

where

$$\begin{aligned}
\Delta =& \mathbf{d}^t\mathbf{H}^{-2}\mathbf{d}\cdot\mathbf{1}^t\mathbf{H}^{-2}\mathbf{1} - \mathbf{d}^t\mathbf{H}^{-2}\mathbf{1}\cdot\mathbf{1}^t\mathbf{H}^{-2}\mathbf{d}\\
=& \mathrm{tr}(\mathbf{D}^2\mathbf{H}^{-2})\cdot\mathrm{tr}(\mathbf{H}^{-2}) - \mathrm{tr}(\mathbf{D}\mathbf{H}^{-2})^2\\
=& \frac{1}{\eta^2}\left(\mathrm{tr}((\mathbf{H}-\mathbf{I})^2\mathbf{H}^{-2})\mathrm{tr}(\mathbf{H}^{-2}) - \mathrm{tr}((\mathbf{H}-\mathbf{I})\mathbf{H}^{-2})^2\right)\\
=& \frac{1}{\eta^2}\left(n\,\mathrm{tr}(\mathbf{H}^{-2}) - \mathrm{tr}(\mathbf{H}^{-1})^2\right).
\end{aligned} \tag{22}$$

Note $\hat{t} = \frac{\mathbf{r}^t\mathbf{H}^{-1}\mathbf{r}}{n}$ and $2\eta f'_{ml}(\eta) = \mathrm{tr}(\mathbf{H}^{-1}) - n\frac{\mathbf{r}^t\mathbf{H}^{-2}\mathbf{r}}{\mathbf{r}^t\mathbf{H}^{-1}\mathbf{r}}$, we have

$$\hat{\gamma}/\hat{t} = \frac{1}{\Delta\eta}\left(n^2V + \mathrm{tr}(\mathbf{H}^{-1})2\eta f'_{ml}(\eta)\right), \tag{23}$$

and

$$\hat{\mu}/\hat{t} = \frac{1}{\Delta\eta^2}\left(n^2V + [\mathrm{tr}(\mathbf{H}^{-1}) - n]2\eta f'_{ml}(\eta)\right). \tag{24}$$

Putting together and plug in $\Delta = \frac{n^2 V}{\eta^2}$ to get

$$\eta^\dagger = \hat{\gamma}/\hat{t} + (1 - \hat{\mu}/\hat{t}\,)\eta = \eta + \frac{2\eta^2}{nV} f'_{ml}(\eta) \tag{25}$$

Alternatively, note

$$\begin{aligned}
\eta^\dagger = \hat{\gamma}/\hat{t} &= \eta \frac{n\,\mathrm{tr}(\mathbf{H}^{-2}) - \mathrm{tr}(\mathbf{H}^{-1})\; n\; \frac{\mathbf{r}^t \mathbf{H}^{-2}\mathbf{r}}{\mathbf{r}^t \mathbf{H}^{-1}\mathbf{r}}}{n\,\mathrm{tr}(\mathbf{H}^{-2}) - \mathrm{tr}(\mathbf{H}^{-1})^2} \\
&= \eta \left[1 + \frac{\mathrm{tr}(\mathbf{H}^{-1})^2 - \mathrm{tr}(\mathbf{H}^{-1})\; n\; \frac{\mathbf{r}^t \mathbf{H}^{-2}\mathbf{r}}{\mathbf{r}^t \mathbf{H}^{-1}\mathbf{r}}}{n\,\mathrm{tr}(\mathbf{H}^{-2}) - \mathrm{tr}(\mathbf{H}^{-1})^2} \right] \\
&= \eta + \frac{2\eta^2}{n} f'_{ml}(\eta) \frac{1}{V(\mathbf{H}^{-1})} \cdot \frac{\mathrm{tr}(\mathbf{H}^{-1})}{n} \\
&= \eta + \frac{2\eta^2}{nV} f'_{ml}(\eta) \cdot \frac{\mathrm{tr}(\mathbf{H}^{-1})}{n}
\end{aligned} \tag{26}$$

has a fractional step size with fraction being $\frac{\mathrm{tr}(\mathbf{H}^{-1})}{n}$, so that we can combine updates (26) and (25) to get a lazy update

$$\begin{aligned}
\eta^\dagger = \frac{1}{2}\left(\hat{\gamma}/\hat{t} + (1 - \hat{\mu}/\hat{t}\,)\eta\right) + \frac{1}{2}\hat{\gamma}/\hat{t} &= \hat{\gamma}/\hat{t} + \frac{1}{2}(1 - \hat{\mu}/\hat{t}\,)\eta \\
&= \eta + \frac{2\eta^2}{nV} f'_{ml}(\eta) \frac{1 + \mathrm{tr}(\mathbf{H}^{-1})/n}{2}.
\end{aligned} \tag{27}$$

Finally, note that $\mathbf{H} = \eta \mathbf{D} + \mathbf{I}_n$ and since $\eta > 0$ and $\mathbf{D}_j > 0$ so $\mathbf{H}_j > 1$. Denote h_j the j-th diagonal element of $\mathbf{H}^{-1}$, we have

$$\begin{aligned}
n^2 V &= n\,\mathrm{tr}(\mathbf{H}^{-2}) - \mathrm{tr}(\mathbf{H}^{-1})\mathrm{tr}(\mathbf{H}^{-1}) \\
&= \sum_{ij} h_i^2 - \sum h_i h_j \\
&= \frac{1}{2}\left(\sum_{ij} (h_i^2 + h_j^2) - \sum_{i,j} 2 h_i h_j \right) \\
&= \sum_{i,j} (h_i - h_j)^2 > 0.
\end{aligned} \tag{28}$$

On the other hand,

$$\begin{aligned}
V &= \frac{1}{n}\,\mathrm{tr}(\mathbf{H}^{-2}) - \frac{1}{n^2}\mathrm{tr}(\mathbf{H}^{-1})\mathrm{tr}(\mathbf{H}^{-1}) \\
&< \frac{1}{n}\,\mathrm{tr}(\mathbf{H}^{-2}) \\
&< 1.
\end{aligned} \tag{29}$$

□

8.4 Proof of Theorem 3

Proof. We first simplify a long term $\mathbf{y_Q^t P_x D P_x D P_x y_Q}$. Using $\mathbf{D} = \frac{1}{\eta}(\mathbf{H} - \mathbf{I}_n)$ twice, we get:

$$\begin{aligned}
\mathbf{y_Q^t P_x D P_x D P_x y_Q} &= \mathbf{y_Q^t H^{-1} P D P_x D H^{-1} P y_Q} \\
&= \mathbf{r^t H^{-1} D P_x D H^{-1} r} \\
&= \frac{1}{\eta^2}\mathbf{r^t H^{-1}}(\mathbf{H} - I_n)\mathbf{P}_x(\mathbf{H} - \mathbf{I}_n)\mathbf{H^{-1} r} \\
&= \frac{1}{\eta^2}\left(\mathbf{r^t P}_x\mathbf{r} - \mathbf{r^t P}_x\mathbf{H^{-1} r} - \mathbf{r^t H^{-1} P_x r} + \mathbf{r^t H^{-1} P_x H^{-1} r}\right) \\
&= \frac{1}{\eta^2}\left(\mathbf{r^t H^{-1} r} - 2\mathbf{r^t H^{-2} r} + \mathbf{r^t H^{-1} P_x H^{-1}} r\right),
\end{aligned} \tag{30}$$

where in the last equality, we used $\mathbf{r^t P}_x\mathbf{H^{-1} r} = \mathbf{r^t P^t H^{-1} H^{-1} r} = \mathbf{r^t H^{-2} r}$. Combing this and 17 and the assumption that $f'_{ml}(\eta) = 0$ to get

$$\begin{aligned}
f''_{ml}(\eta) &= \frac{1}{2}\mathrm{tr}(\mathbf{H^{-1} D H^{-1} D}) - \frac{n}{1}\frac{\mathbf{y_Q^t P}_x\mathbf{D P}_x\mathbf{D P_x y_Q}}{\mathbf{y_Q^t P_x y_Q}} + \frac{n}{2}\frac{(\mathbf{y_Q^t P_x D P_x y_Q})^2}{(\mathbf{y_Q^t P_x y_Q})^2} \\
&= \frac{1}{2\eta^2}\mathrm{tr}(\mathbf{H}^{-2}(\mathbf{H} - \mathbf{I}_n)^2) - \frac{n}{\eta^2}\frac{\mathbf{r^t H^{-1} r} - 2\mathbf{r^t H^{-2} r} + \mathbf{r^t H^{-1} P_x H^{-1} r}}{\mathbf{r^t H^{-1} r}} + \frac{n}{2\eta^2}\left(\frac{\mathbf{r^t H^{-1} r} - \mathbf{r^t H^{-2} r}}{\mathbf{r^t H^{-1} r}}\right)^2 \\
&= \frac{1}{2\eta^2}\mathrm{tr}(\mathbf{I}_n - 2\mathbf{H^{-1}} + \mathbf{H^{-2}}) + \frac{n}{2\eta^2}\left(-1 + 2\frac{\mathbf{r^t H^{-2} r}}{\mathbf{r^t H^{-1} r}} - 2\frac{\mathbf{r^t H^{-1} P_x H^{-1} r}}{\mathbf{r^t H^{-1} r}} + \left(\frac{\mathbf{r^t H^{-2} r}}{\mathbf{r^t H^{-1} r}}\right)^2\right) \\
&= \frac{1}{2\eta^2}\left(\mathrm{tr}(\mathbf{H^{-2}}) - 2n\frac{\mathbf{r^t H^{-3} r}}{\mathbf{r^t H^{-1} r}} + n\left(\frac{\mathbf{r^t H^{-2} r}}{\mathbf{r^t H^{-1} r}}\right)^2 + 2n\frac{\mathbf{r^t H^{-1} H^{-1} P}_0\mathbf{H^{-1} r}}{\mathbf{r^t H^{-1} r}}\right) - \frac{1}{\eta}f'_{ml}(\eta) \\
&= \frac{n}{2\eta^2}\left[\left(\frac{1}{n}\mathrm{tr}(\mathbf{H^{-2}}) - \frac{1}{n^2}\mathrm{tr}(\mathbf{H^{-1}})^2\right) + \left(2\left(\frac{\mathbf{r^t H^{-2} r}}{\mathbf{r^t H^{-1} r}}\right)^2 - 2\frac{\mathbf{r^t H^{-3} r}}{\mathbf{r^t H^{-1} r}}\right) + 2\frac{\mathbf{r^t H^{-1} H^{-1} P}_0\mathbf{H^{-1} r}}{\mathbf{r^t H^{-1} r}}\right].
\end{aligned} \tag{31}$$

We examine three terms in the square bracket in turn. Denote h_j the j-th diagonal element of $\mathbf{H}^{-1}$, with reference to Equation 28 the first term can be transformed to

$$\begin{aligned}\frac{1}{n}\mathrm{tr}(\mathbf{H}^{-2}) - \frac{1}{n^2}\mathrm{tr}(\mathbf{H}^{-1})^2 &= \frac{1}{2n^2}\sum_{i,j}(h_i - h_j)^2 \\ &= \frac{1}{2n^2}\sum_{ij}(h_i - \theta + \theta - h_j)^2 \\ &= \frac{1}{2n^2}\sum_{ij}(h_i-\theta)^2 + \frac{1}{2n^2}\sum_{ij}(\theta - h_j)^2 + \frac{1}{n^2}\sum_{ij}(h_i-\theta)(h_j-\theta) \\ &= \frac{1}{n}\sum_{i}(h_i-\theta)^2,\end{aligned} \tag{32}$$

where $\theta = \frac{1}{n}\sum_i h_i$ so that $\frac{1}{n^2}\sum_{ij}(h_i-\theta)(h_j-\theta) = 0$. This is the average of the squared errors and we denoted it by $V(\mathbf{H}^{-1})$.

The second term can be transformed to

$$\begin{aligned}2\left(\frac{\mathbf{r}^t\mathbf{H}^{-2}\mathbf{r}}{\mathbf{r}^t\mathbf{H}^{-1}\mathbf{r}}\right)^2 - 2\frac{\mathbf{r}^t\mathbf{H}^{-3}\mathbf{r}}{\mathbf{r}^t\mathbf{H}^{-1}\mathbf{r}} &= 2\left(\frac{\mathbf{r}^t\mathbf{H}^{-2}\mathbf{r}\cdot\mathbf{r}^t\mathbf{H}^{-2}\mathbf{r} - \mathbf{r}^t\mathbf{H}^{-3}\mathbf{r}\cdot\mathbf{r}^t\mathbf{H}^{-1}\mathbf{r}}{\mathbf{r}^t\mathbf{H}^{-1}\mathbf{r}\cdot\mathbf{r}^t\mathbf{H}^{-1}\mathbf{r}}\right) \\ &= 2\frac{\sum_{i,j}\left(h_i^2h_j^2r_i^2r_j^2 - h_i^3h_jr_i^2r_j^2\right)}{\sum_{i,j}h_ih_jr_i^2r_j^2} \\ &= 2\frac{\sum_{i,j}\left(h_ih_jr_i^2r_j^2(h_ih_j - h_i^2)\right)}{\sum_{i,j}h_ih_jr_i^2r_j^2} \\ &= \frac{\sum_{i,j}\left(h_ih_jr_i^2r_j^2(2h_ih_j - h_i^2 - h_j^2)\right)}{\sum_{i,j}h_ih_jr_i^2r_j^2} \\ &= -\frac{\sum_{i,j}(h_i-h_j)^2s_i^2s_j^2}{\sum_{i,j}s_i^2s_j^2} \\ &= -\frac{\sum_{i,j}(h_i-\theta+\theta-h_j)^2s_i^2s_j^2}{\sum_{i,j}s_i^2s_j^2} \\ &= -\frac{\sum_{i,j}\left[(h_i-\theta)^2 + (\theta-h_j)^2 + 2(h_i-\theta)(\theta-h_j)\right]s_i^2s_j^2}{\sum_{i,j}s_i^2s_j^2} \\ &= -2\frac{\sum_i(h_i-\theta)^2s_i^2}{\sum_i s_i^2} + 2\frac{\sum_i(h_i-\theta)s_i^2}{\sum_i s_i^2}\frac{\sum_j(h_j-\theta)s_j^2}{\sum_j s_j^2},\end{aligned} \tag{33}$$

where we denote $s_i^2 = r_i^2h_i$. To this end, the term $-2\frac{\sum_i(h_i-\theta)^2s_i^2}{\sum_i s_i^2}$ can be seen as a stochastic average of squared errors, note $s_i^2 \sim \chi_1^2$, and $s_i^2/\sum_i s_i^2 \sim Beta(\frac{1}{2}, \frac{n-1}{2})$ with mean $\frac{1}{n}$ and variance $\frac{2(n-1)}{n^2(n+2)}$. Therefore, the mean of the term is $-2V(\mathbf{H}^{-1})$, and the variance of the term

is $< \frac{2}{n^2}\sum_i (h_i-\theta)^4 < \frac{2}{n^2}\sum_i (h_i-\theta)^2 = \frac{2}{n}V(\mathbf{H}^{-1})$, which goes to 0 as n increases.

Identifying the second term as a square of sum of random variables, and we apply central limit theorem to show it's a square of a normal random variable whose mean and variance both vanish as n increases. Let $Z_i = \frac{(h_i-\theta)s_i^2}{\sum_i s_i^2} = (h_i-\theta)B_i$, where $B_i \sim Beta(\frac{1}{2}, \frac{n-1}{2})$, $Var(Z_i) \approx (h_i-\theta)^2\frac{2}{n^2}$. Denote $Z = \sum_i Z_i$ and $Var(Z) = \sum_i Var(Z_i) = \frac{2}{n}V(\mathbf{H}^{-1})$. We are to apply Lyapunov central limit theorem, so let us check that Lyapunov's condition holds: $E(Z_i) = (h_i-\theta)/n$ and $E[|Z_i - (h_i-\theta)/n|^3] = |h_i-\theta|^3 E(|B_i - \frac{1}{n}|^3) = |h_i-\theta|^3 O(\frac{1}{n^3})$, so that $\sum_i E[|Z_i-(h_i-\theta)/n|^3] = O(\frac{1}{n^2})$, and $\frac{1}{Var(Z)}\sum_i E[|Z_i-(h_i-\theta)/n|^3] = O(\frac{1}{n})$, which satisfy Lyapunov's condition. Then by Lyapunov central limit theorem $\frac{1}{Var(Z)}\sum_i (Z_i - E(Z_i)) \to N(0,1)$, and equivalently $Z \sim N(0, \frac{2}{n}V(\mathbf{H}^{-1}))$. Thus, the second term has mean $E(Z^2) = \frac{2}{n}V(\mathbf{H}^{-1}) = O(\frac{1}{n})$, and variance (computed via scaled χ_1^2) is $Var(Z^2) = 2(\frac{2}{n}V(\mathbf{H}^{-1}))^2 = O(\frac{1}{n^2})$, which go to 0 as n increases.

The third term can be transformed to

$$
\begin{aligned}
2\frac{\mathbf{r}^t\mathbf{H}^{-1}\mathbf{H}^{-1}\mathbf{P}_0\mathbf{H}^{-1}\mathbf{r}}{\mathbf{r}^t\mathbf{H}^{-1}\mathbf{r}} &= 2n\frac{\mathbf{r}^t(\mathbf{H}^{-1}-\theta\mathbf{I}_n)\mathbf{H}^{-1}\mathbf{P}_0(\mathbf{H}^{-1}-\theta\mathbf{I}_n)r}{\mathbf{r}^t\mathbf{H}^{-1}\mathbf{r}} \\
&= 2\frac{\mathbf{r}^t(\mathbf{H}^{-1}-\theta\mathbf{I}_n)\mathbf{H}^{-\frac{1}{2}}U\Lambda U^t\mathbf{H}^{-\frac{1}{2}}(\mathbf{H}^{-1}-\theta\mathbf{I}_n)\mathbf{r}}{\mathbf{r}^t\mathbf{H}^{-1}\mathbf{r}} \\
&= 2\frac{\sum_{j=1}^c(\mathbf{r}^t(\mathbf{H}^{-1}-\theta\mathbf{I}_n)\mathbf{H}^{-1/2}u_{\cdot j})^2}{\mathbf{r}^t\mathbf{H}^{-1}\mathbf{r}} \\
&= 2\frac{\sum_{j=1}^c\left(\sum_i r_i h_i^{1/2}(h_i-\theta)u_{ij}\right)^2}{\sum_i r_i^2 h_i} \\
&= 2\frac{\sum_{j=1}^c w_j^2}{\sum_i r_i^2 h_i} \\
&< 2F
\end{aligned}
\tag{34}
$$

where $F \sim \frac{\chi_c^2}{\chi_n^2}$. The first equality holds because $\mathbf{r} = (\mathbf{I}_n - \mathbf{P}_0)\mathbf{y}_\mathbf{Q}$ and $\mathbf{r}^t\mathbf{H}^{-1}\mathbf{P}_0\mathbf{v} = 0$ and $\mathbf{v}\mathbf{H}^{-1}\mathbf{P}_0\mathbf{r}^t = 0$ for any $\mathbf{v}$ (or in other words, we add terms equal to 0); the second equality holds because $\mathbf{P}_2 = \mathbf{H}^{-1/2}\mathbf{X}_\mathbf{Q}(\mathbf{X}_\mathbf{Q}^t\mathbf{H}^{-1}\mathbf{X}_\mathbf{Q})^{-1}\mathbf{X}_\mathbf{Q}^t\mathbf{H}^{-1/2}$ is also a projection, and $\mathbf{P}_2$ is symmetric and has the same rank and trace as $\mathbf{P}_0$, therefore $\mathbf{P}_2 = U\Lambda U^t$ where U is orthonormal; the third equality holds because Λ has c eigenvalues 1 and $n-c$ eigenvalues 0; the fourth equality holds by definition; in the fifth equality, we define $w_j = \sum_i (r_i h_i^{1/2}(h_i-\theta)u_{ij})$, and because $r_i h_i^{1/2}$ is standard normal, w_j is a weighted sum of normal random variables, and itself a normal with mean 0 and variance $v_j = \sum_i (h_i-\theta)^2 u_{ij}^2 < \max_i(h_i-\theta)^2 \sum_i u_{ij}^2 = \max_i(h_i-\theta)^2 < 1$, which gives the last inequality. (Note that $u_{\cdot j}$ is an orthonormal basis, and $\sum_i u_{ij}^2 = 1$.) Finally, $F \sim \frac{\chi_c^2}{\chi_n^2}$, and $\frac{n}{c}F$ follows F-distribution with d.f. c and n ($c << n$), whose mean and variance is $O(1)$, and thus $F = O(\frac{1}{n})$ goes to 0 as n increases.

Putting together,

$$f''_{ml}(\eta) = \frac{n}{2\eta^2}\left[-V(\mathbf{H}^{-1}) + \epsilon_n\right], \tag{35}$$

with both mean and variance of ϵ_n decreases linearly $O(\frac{1}{n})$. Thus, $f''_{ml}(\eta) < 0$ asymptotically almost sure, or with probability 1, at where $f'_{ml}(\eta) = 0$. □