Fitting linear mixed model in genetic association studies: an efficient and exact method

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

February 22, 2024

Abstract

The linear mixed model (LMM) has become a standard tool 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 computation remained computationally demanding and without theoretical assurance. The computation is particularly challenging for multi-omics studies where tens of thousands of phenotypes are tested for association with millions of genetic markers. We present IDUL that uses iterative dispersion updates to fit LMMs. We demonstrated that IDUL is a gradient ascent algorithm, enjoys the quadratic convergence of the state-of-the-art Newton-Raphson method, but avoids its inconsistency and inefficiency in computation. Moreover, we showed that the likelihood of LMM is asymptotically locally concave at the optimum, which guarantees unimodality of the likelihood function, so that IDUL achieves exact optimum almost sure. Applied to a set of protein assays from the Framingham Heart Study, IDUL demonstrated its superb efficiency and consistency: converged in several iterations and provided consistent estimates between different initial values. IDUL is highly efficient for additional phenotypes, making it ideal to study genetic determinant of multi-omics phenotypes. A software package implementing 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), and 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 multi-omics 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 cheaply, 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 is fitting the LMM once under the null and used it for all SNPs; the other is adopting GRAMMAR-Gamma approach in computing score test statistics; and the third is using hard thresholding to make 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 but maintains exact computation. Exact computation removes the need to consider which approximate computation works best for a given dataset. 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, which converges linearly. The rotation reduced computation complexity in optimization. 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 quadratically. The comparison between FaST-LMM and GEMMA can be found in (Zhou and Stephens, 2012). However, the Newton-Raphson method suffers from inconsistency when the initial values are distant from the optimum, so that GEMMA starts its optimization iterations with Brent's

method and then switches to the Newton-Raphson method.

In this paper, we present IDUL that uses iIterative dispersion update to fit LMMs in genetic association studies, and demonstrate its theoretical and practical advantages. IDUL aims to solve the inconsistency (or starting-point-dependency) observed in the Newton-Raphson algorithm, to improve computation efficiency, and to make the *exact* computation of LMMs applicable to large-scale multi-omics studies. As an iterative method, IDUL only fits two weighted least square regressions in each iteration, so that its per iteration computation is highly efficient compared to the Newton-Raphson method. IDUL is designed to analyzing multi-omics data, and it is highly efficient for additional phenotypes. Our theoretical analysis revealed that IDUL is a gradient ascent algorithm, enjoys the quadratic convergence of the Newton-Raphson method, but avoids its inconsistency and inefficiency in computation. Most importantly, we show that the likelihood function is asymptotically locally concave, which implies unimodality of the likelihood function. Therefore, IDUL achieves the exact optimum almost sure.

2 Results

2.1 The model and the algorithm

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 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

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

where $\mathbf{X_Q} = \mathbf{Q}^t\mathbf{X}$, $\mathbf{y_Q} = \mathbf{Q}^t\mathbf{y}$, and $\mathbf{H} = \eta\mathbf{D} + \mathbf{I}_n$ is a diagonal matrix. IDUL repeats the following two steps until η converges:

S1 For a given η (and hence $\mathbf{H}$), fit (2) using weighted least squares with weight $\mathbf{H}^{-1}$ to obtain residual $\mathbf{r}$ and compute $\mathbf{s}$ such that $\mathbf{s}_i = \mathbf{r}_i^2$; and

S2 Denote $\mathbf{d}$ the vector of the diagonal elements of $\mathbf{D}$, and fit $\mathbf{s} \sim \mathrm{MVN}_n(\mu + \mathbf{d}\gamma, \tau^{-1}\mathbf{H}^2)$ using weighted least square with weight $\mathbf{H}^{-2}$ to obtain estimates $\hat{\gamma}$ and $\hat{\mu}$ and update $\eta = \hat{\gamma}/\hat{\mu}$.

Intuitively, $\hat{\gamma}$ 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).

2.2 Analytic update of the algorithm

IDUL is equivalent to the following analytical update:

$$\eta^\dagger = \eta + \frac{2\eta^2}{n}\frac{1}{V(\mathbf{H}^{-1}) + v_n}\, f'_{ml}(\eta), \tag{3}$$

where $V(\mathbf{H}^{-1}) = \mathrm{tr}(\mathbf{H}^{-2})/n - \mathrm{tr}(\mathbf{H}^{-1})^2/n^2 > 0$, $v_n = \frac{2\eta}{n^2}\left(\mathrm{tr}(\mathbf{H}^{-1}) - n\right) f'_{ml}(\eta)$, and $f'_{ml}(\eta)$ is the first derivative of the log-likelihood evaluated at η, and can be computed analytically (Material and Methods). Note Equation (3) is derived here to study the analytic properties of IDUL algorithm, not for model fitting.

For a sufficiently large n or η in the neighborhood of $f'_{ml}(\eta) = 0$, the term v_n can be safely ignored, so that the increment is positively proportional to $f'_{ml}(\eta)$ such that $\eta^\dagger - \eta = \alpha f'_{ml}(\eta)$ for some $\alpha > 0$. This leads to the following observations: 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, likelihood increases after the update, which can be seen through Taylor expansion of the log-likelihood at η to get $f_{ml}(\eta^\dagger) - f_{ml}(\eta) = \alpha f'_{ml}(\eta)^2 + O(|\eta^\dagger - \eta|^2)$. Of course, when $|\eta^\dagger - \eta|$ is large the update can overshoot to land on the other side of the optimum, causing the algorithm converges like a damping pendulum. The overshoot can be resolved by making IDUL lazy (Material and Methods).

2.3 The global optimum

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

$$f''_{ml}(\eta) = \frac{n}{2\eta^2}\left[-V(\mathbf{H}^{-1}) + \epsilon_n\right], \tag{4}$$

where $V(\mathbf{H}^{-1})$ is defined in (3) and both mean and variance of ϵ_n varnish linearly (proportional to $1/n$) as n increases (details in Method). In other words, at local optimum such that $f'_{ml}(\eta) = 0$, $f''_{ml}(\eta) < 0$ almost sure, or with probability 1, for a sufficiently large n. Most significantly, this asymptotically local concaveness implies that with a sufficiently large sample size, the log-likelihood $f_{ml}(\eta)$ attains its unique global maximum at $f'_{ml}(\eta) = 0$. For 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 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 (asymptotically) unimodal, and hence an optimum to which IDUL converges is the global optimum.

2.4 Connection with Newton-Raphson

When both v_n in Equation (3) and ϵ_n in Equation (4) can be safely ignored (such as when n is large), The analytic update of IDUL becomes $\eta^\dagger = \eta - f'_{ml}(\eta)/f''_{ml}(\eta)$, which is the Newton-Raphson update. In other words, IDUL is equivalent to the Newton-Raphson method at the neighborhood of the optimum. But IDUL requires no computation of the second derivative, and outside the neighborhood of the optimum, where the Newton-Raphson method is known to be numerically unstable (Burden and Faires, 2010), IDUL produces consistent results (below).

2.5 Consistency of IDUL

To compare consistency between IDUL and Newton-Raphson, for each phenotype, we fit 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 $\eta_0 = h_0/(1 - h_0)$. The datasets we used for comparison were described in Material and Methods.

It takes IDUL on average 7.3 iterations to converge for each combination of phenotype and initial value, compared to 12.5 of Newton-Raphson. Figure 1 compares the consistency of fitted $\hat{\eta}$ when initial values η_0 were drawn from different V segments. IDUL had perfect

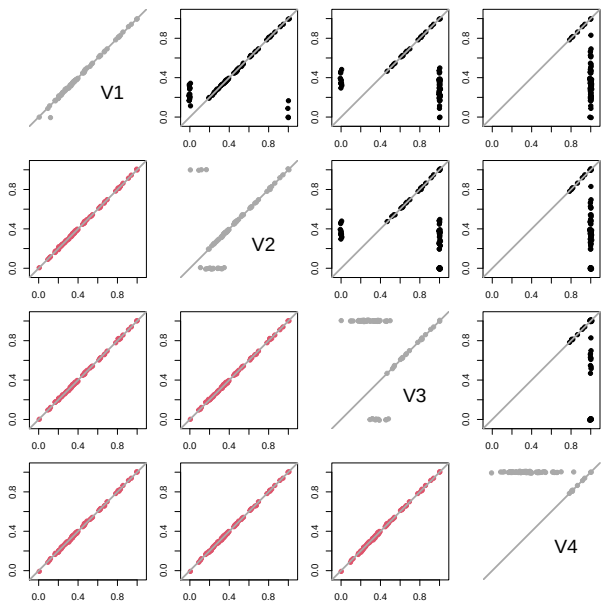

Figure 1: Consistency of IDUL and Newton-Raphson w.r.t different initial values. IDUL and Newton-Raphson were run on the same sets of initial values. For each phenotype, four initial values were generated with seeds randomly selected from four segments. So both IDUL and Newton-Raphson produced 4 columns 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.

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 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 is a trap for 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 S2). To overcome the inconsistency, implementing the Newton-Raphson method to fit LMM requires multiple runs from random initial values and comparing their likelihoods. As a comparison, IDUL only needs to run once for each model fitting, each run takes less iterations, and each iteration takes less computation.

2.6 Efficiency of IDUL

We implemented IDUL into a software package of the same name that fits the LMMs and compute test statistics, and compared with GEMMA, a software package that fits LMMs using the Newton-Raphson method. The multi-omics data we used for comparison was described in Material and Methods. For both likelihood ratio test and Wald test p-values (see Material and Methods), IDUL and GEMMA reached almost perfect agreement (Supplementary Figure S1). This suggested that with primer runs using Brent's method, GEMMA can reliably achieves exact optimum for both ML and REML estimates, albeit with extra computation (Table 1).

Comparing efficiency of algorithms in practice always have caveats, as the details of implementation matter. But even if we ignore the dramatic improvement in efficiency of IDUL over GEMMA on wall time, and focus on the total CPU time spent, IDUL is still markedly more efficient, spending 58% CPU time of that of GEMMA for likelihood ratio test and 51% CPU time for Wald test. Most strikingly, for additional 9 phenotypes, IDUL only spends 11% (1.6 minutes) extra wall time of computing the first phenotype for likelihood ratio test; for additional 78 phenotypes, 43% (6 minutes) extra wall time of computing the first phenotype. Similar high efficiency for additional phenotypes was also observed with Wald test (Table 1). The high efficiency for extra phenotypes makes IDUL ideal to study genetic determinant of multi-omics phenotypes.

2.7 Exact vs approximation

IDUL is an exact method. Zhou and Stephens (2012) classified methods into exact methods and approximate methods and demonstrated that 1) among exact methods available at the time,

		CPU (minutes)	Wall (minute)		
		$p = 1$	$p = 1$	10	79
LRT	IDUL	474.0	14.0	15.6	20.0
	GEMMA	811.4	76.5	-	-
Wald	IDUL	393.7	12.8	16.4	25.3
	GEMMA	779.4	98.9	-	-

Table 1: Comparison between IDUL and GEMMA on one phenotype, and comparisons of IDUL on multiple phenotypes. GEMMA used 112 threads and IDUL used 64. Judged from the ratio between CPU time and Wall time, it is likely GEMMA only used multithreading in part of its computation, while IDUL implemented SNP level parallel computation. Thus we provide IDUL with less threads to strive for a fair comparison. Taking multiple phenotypes for sequential analysis was not implemented in GEMMA, and its wall time for 10 and 79 phenotypes are missing.

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 multi-omics dataset in Framingham Heart Study, however, we found the hard threshold approach less satisfactory, presumably due to closer relatedness between samples and larger effect sizes in multi-omics dataset. Figure 2 showed comparison of test statistics of eight protein phenotypes between exact computation and hard thresholding approximation. (These phenotypes are chosen because their large η estimates under the null.) The inconsistency is rather pronounced, suggesting potential difficulty with approximate methods in multi-omics data for closely related samples such as in Framingham Heart Study.

3 Discussion

We documented two key contributions for exact methods in linear mixed model. First, we introduced a novel, consistent, and highly efficient algorithm IDUL to fit the standard linear mixed model for genetic association studies, and demonstrated its superiority over the Newton-Raphson method from both theoretical and practical perspectives. With IDUL, we achieved efficiency of the Newton-Raphson method, but avoided the inconsistency and ineffi-

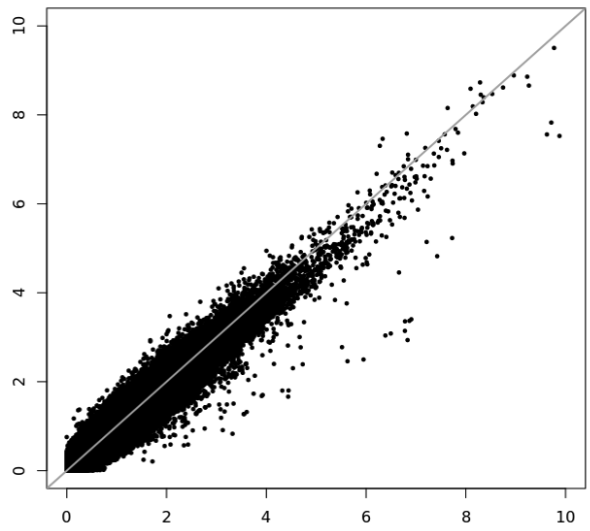

Figure 2: Comparison between exact and approximate ($-\log_{10}$) p-values. The exact p-values are on x-axis, and approximate p-values are on y-axis. Points whose either exact ($-\log_{10}$) p-value or approximate ($-\log_{10}$) p-value larger than 10 were omitted.

ciency Newton-Raphson suffers. Surprisingly, we achieve these benefits with less computation because IDUL only involves fitting weighted linear regression without the need of the second derivative of the log-likelihood, which is expensive to compute.

Second, we showed that the log-likelihood is asymptotically locally concave at the optimum. (We hypothesize that the same is true for REML likelihood, see Material and Methods). A profound corollary is that the standard linear mixed model has a unique global optimum. The local concaveness at the unique optimum and IDUL being a gradient ascent algorithm jointly underpin its perfect consistency over different initial values. Thus, IDUL can reliably achieve the exact optimum. Although much recent effort in linear mixed model focused on developing approximate methods, exact methods are always more satisfactory than approximate methods. As it was noted in (Zhou and Stephens, 2012), one advantage of exact computation is that it obviates the need to consider which approximations work best for your datasets.

IDUL is designed with genetic association testing for multi-omics 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 genome-wide multi-omic association analysis, we have n sample, m genetic markers, p phenotypes, and c covariates, then the total complexity is $O(n^3 + (m+p+c)n^2 + tmpnc^2)$, where the complexity for eigen decomposition for the kinship matrix is $O(n^3)$, the complexity for multiplying Q^t with all SNPs, all phenotypes, and all

covariates is $O((m+p+c)n^2)$, and the complexity for optimization of all genotype-phenotype pairs is $O(tmpnc^2)$, where $O(nc^2)$ is the complexity of solving a single linear regression, and we assume on average it takes t linear regression for IDUL to converge. For a typical multi-omic analysis where $m > p > n$, the dominate term in total complexity is $O(tmpnc^2)$, which is linear in number of samples n, number of genetic markers m, and number of phenotypes p. On a separate but related note, although Newton-Raphson has the same computation complexity, it tends to have a larger equivalent t than IDUL because it essentially 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.

From a different perspective, the upfront costs of eigen decomposition and SNP rotations are entirely free for additional phenotypes. 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.

When testing fixed effects in LMM, MLE is preferred over the REML estimate, because for fixed effects ML estimates are unbiased, but REML estimates are biased. REML estimates, however, is preferred when the interest is the variance component, such as in estimating trait heritability, because it produces unbiased estimate of η at the cost of the biased estimates of the fixed effect. IDUL can be adapted to obtain REML estimate based on its analytical update (Material 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 S3). As a comparison, the Average Information REML implemented in GCTA had difficulty even with 90% down-sampling and produced untenable results.

We demonstrated that IDUL is superior over the Newton-Raphson method in fitting standard LMMs. We would like to point out that IDUL is a specialized algorithm and takes

advantage of a specific dispersion structure that is 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. Finally, 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.

4 Material 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 multi-omics 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 Likelihoods and their 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

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

and 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

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

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

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

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

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

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 Step 1 of 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 $\eta^\dagger = \hat{\gamma}/\hat{\mu}$ in Step 2.

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

$$\eta^\dagger = \eta + \frac{2\eta^2}{n} f'_{ml}(\eta) \frac{1}{V(\mathbf{H}^{-1}) + v_n}, \tag{12}$$

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

Proof of Lemma 2 is deferred to Appendices. With the analytic update (12), it's easy to see that IDUL is a gradient ascend algorithm. 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 $f_{ml}(\eta^\dagger) - f_{ml}(\eta) = \frac{2\eta^2}{n} {f'_{ml}}^2(\eta) \frac{1}{V(\mathbf{H}^{-1}) + v_n} + \frac{1}{2}f''_{ml}(\xi)|\eta^\dagger - \eta|^2$. When n is large v_n can be ignored, and when $|\eta^\dagger - \eta|$ is small or $f''_{ml}(\xi) > 0$, we have $f_{ml}(\eta^\dagger) - f_{ml}(\eta) > 0$.

When $f''_{ml}(\xi) < 0$ and $|\eta^\dagger - \eta|$ is large, the update can overshoot, landing on the other side of the mode, and converge like a damping pendulum . We can modify Step 2 to obtain an alternative update $\eta^\dagger = \hat{\gamma}\hat{\tau}$, where $\hat{\tau} = \frac{n}{\mathbf{r}^t\mathbf{H}^{-1}\mathbf{r}}$. This leads to an analytical form

$$\eta^\dagger = \eta + \frac{2\eta^2}{n} f'_{ml}(\eta) \frac{1}{V(\mathbf{H}^{-1})} \frac{\mathrm{tr}(\mathbf{H}^{-1})}{n}. \tag{13}$$

Compare updates (13) with (12), the update $\eta_2^\dagger = \hat{\gamma}\hat{\tau}$ takes a smaller step as $\mathrm{tr}(\mathbf{H}^{-1})/n < 1$. These two updates can be combined to get $\eta^\dagger = (\eta_1^\dagger + \eta_2^\dagger)/2$ to prevent overshooting. We call this making IDUL lazy (R code in Appendices).

4.5 From MLE to REML

With maximum likelihood update (12) at hand, we can obtain REML update by substituting $f'_{ml}(\eta)$ with $f'_{re}(\eta)$, to get

$$\eta^\dagger = \eta + \frac{2\eta^2}{n} \frac{1}{V(\mathbf{H}^{-1})} f'_{re}(\eta), \tag{14}$$

where we took the liberty to remove the v_n term. Of course, following the update (13), we can get REML version of it by substituting $f'_{ml}(\eta)$ with $f'_{re}(\eta)$ to make the REML update lazy.

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.

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

$$f''_{ml}(\eta) = \frac{n}{2\eta^2}\left[-V(\mathbf{H}^{-1}) + \epsilon_n\right], \tag{15}$$

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

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

$$f''_{re}(\eta) = \frac{n-c}{2\eta^2}\left[-V(\mathbf{H}^{-1}\mathbf{P}) + \epsilon_n\right], \tag{16}$$

where $V(\mathbf{H}^{-1}\mathbf{P}) = tr(\mathbf{H}^{-1}\mathbf{P}\mathbf{H}^{-1}\mathbf{P})/(n-c) - tr(\mathbf{H}^{-1}\mathbf{P})^2/(n-c)^2 > 0$, 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.

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

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.

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, 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

```
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) {
    h = runif(1); #random init on heritability;
    eta = h/(1-h);
    repeat { print(eta);
       H = eta * D + 1;
       r2 = lm(yQ~xQ, weights=1/H)$residuals^2;
       fit2 = lm(r2~D, weights=1/H/H)
       eta1 = fit2$coefficients[2]/fit2$coefficients[1];
       if(abs(eta1/(1+eta1)-eta/(1+eta)) < 1e-6) {break;}
       eta = eta1;
    };  ##iterative dispersion updates
    return(eta);
}
```

8.2 R code for IDUL lazy

```
idul_lazy=function(xQ,yQ,D) {
    h = runif(1); #random init on heritability;
    eta = h/(1-h);
    repeat { print(eta);
       H = eta * D + 1;
       r2 = lm(yQ~xQ, weights=1/H)$residuals^2;
       fit2 = lm(r2~D, weights=1/H/H)
       ta1 = fit2$coefficients[2]/fit2$coefficients[1];
       ta2 = fit2$coefficients[2]/mean(r2/H);
       eta1 = (ta1 + ta2)/2;
       if(abs(eta1/(1+eta1)-eta/(1+eta)) < 1e-6) {break;}
       eta = eta1;
```

```
    }; ##iterative dispersion updates
    return(eta);
}
```

8.3 Proof of Proposition 1

Proof. We first simplify two expressions

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

1) By $\mathbf{PP} = \mathbf{P}$ and $\mathbf{r} = \mathbf{P y_Q}$, we have $\mathbf{y_Q^t P_x} y_Q = \mathbf{y_Q^t H^{-1} P y_Q} = \mathbf{y_Q^t H^{-1} P P y_Q} = \mathbf{y_Q^t P^t H^{-1} P y_Q} = \mathbf{r^t H^{-1} r}$. 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 $\mathbf{y_Q^t P_x D P_x y_Q} = \mathbf{y_Q^t P^t H^{-1} D H^{-1} P y_Q} = \mathbf{r^t H^{-1} D H^{-1} r} = \frac{1}{\eta}\mathbf{r^t H^{-1}}(\mathbf{H} - \mathbf{I}_n)\mathbf{H^{-1} r} = \frac{1}{\eta}(\mathbf{r^t H^{-1} r} - \mathbf{r^t H^{-2} r})$. With these two reduced expressions, the first derivatives can be transformed as following:

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

and

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

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 $\mathrm{tr}(\mathbf{P}_0) = c$.

□

8.4 Proof of Lemma 1

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

$$\begin{aligned}
(\hat{\gamma}, \hat{\mu})^t &= ((\mathbf{d}, \mathbf{1})^t \mathbf{H}^{-2} (\mathbf{d}, \mathbf{1}))^{-1} (\mathbf{d}, \mathbf{1})^t \mathbf{H}^{-2} \mathbf{s} \\
&= \begin{matrix} \mathbf{d}^t\mathbf{H}^{-2}\mathbf{d} & \mathbf{d}^t\mathbf{H}^{-2}\mathbf{1} \\ \mathbf{1}^t\mathbf{H}^{-2}\mathbf{d} & \mathbf{1}^t\mathbf{H}^{-2}\mathbf{1} \end{matrix}^{-1} (\mathbf{d}, \mathbf{1})^t \mathbf{H}^{-2} \mathbf{s} \\
&= \frac{1}{\Delta} \begin{matrix} \mathbf{1}^t\mathbf{H}^{-2}\mathbf{1} & -\mathbf{d}^t\mathbf{H}^{-2}\mathbf{1} \\ -\mathbf{1}^t\mathbf{H}^{-2}\mathbf{d} & \mathbf{d}^t\mathbf{H}^{-2}\mathbf{d} \end{matrix} (\mathbf{d}^t\mathbf{H}^{-2}\mathbf{s}, \mathbf{1}^t\mathbf{H}^{-2}\mathbf{s})^t \\
&= \frac{1}{\Delta} \begin{matrix} \mathbf{1}^t\mathbf{H}^{-2}\mathbf{1} \cdot \mathbf{d}^t\mathbf{H}^{-2}\mathbf{s} - \mathbf{d}^t\mathbf{H}^{-2}\mathbf{1} \cdot \mathbf{1}^t\mathbf{H}^{-2}\mathbf{s} \\ -\mathbf{1}^t\mathbf{H}^{-2}\mathbf{d} \cdot \mathbf{d}^t\mathbf{H}^{-2}\mathbf{s} + \mathbf{d}^t\mathbf{H}^{-2}\mathbf{d} \cdot \mathbf{1}^t\mathbf{H}^{-2}\mathbf{s} \end{matrix} .
\end{aligned} \tag{18}$$

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

$$\begin{aligned}
\hat{\mu} =& \frac{1}{\Delta} \left(-\mathbf{1}^t\mathbf{D}\mathbf{H}^{-2}\mathbf{1} \cdot \mathbf{r}^t\mathbf{D}\mathbf{H}^{-2}\mathbf{r} + \mathbf{1}^t\mathbf{D}^2\mathbf{H}^{-2}\mathbf{1} \cdot \mathbf{r}^t\mathbf{H}^{-2}\mathbf{r} \right) \\
=& \frac{1}{\Delta}\frac{1}{\eta^2} \left(-\mathbf{1}^t\mathbf{H}^{-2}(\mathbf{H} - \mathbf{I})\mathbf{1} \cdot \mathbf{r}^t(\mathbf{H} - \mathbf{I})\mathbf{H}^{-2}\mathbf{r} + \mathbf{1}^t(\mathbf{H} - \mathbf{I})^2\mathbf{H}^{-2}\mathbf{1} \cdot \mathbf{r}^t\mathbf{H}^{-2}\mathbf{r} \right) \\
=& \frac{1}{\Delta}\frac{1}{\eta^2} \left(-\mathrm{tr}(\mathbf{H}^{-1} - \mathbf{H}^{-2}) \cdot \mathbf{r}^t(\mathbf{H}^{-1} - \mathbf{H}^{-2})\mathbf{r} + \mathrm{tr}(\mathbf{I}_n - 2\mathbf{H}^{-1} + \mathbf{H}^{-2}) \cdot \mathbf{r}^t\mathbf{H}^{-2}\mathbf{r} \right) \\
=& \frac{1}{\Delta}\frac{1}{\eta^2} \left(-\mathrm{tr}(\mathbf{H}^{-1})\mathbf{r}^t\mathbf{H}^{-1}\mathbf{r} + n\mathbf{r}^t\mathbf{H}^{-2}\mathbf{r} + \mathrm{tr}(\mathbf{H}^{-2})\mathbf{r}^t\mathbf{H}^{-1}\mathbf{r} - \mathrm{tr}(\mathbf{H}^{-1})\mathbf{r}^t\mathbf{H}^{-2}\mathbf{r} \right) \\
=& \frac{1}{\Delta}\frac{1}{\eta^2}\frac{\mathbf{r}^t\mathbf{H}^{-1}\mathbf{r}}{n} \left[n\ \mathrm{tr}(\mathbf{H}^{-2}) - \mathrm{tr}(\mathbf{H}^{-1})\ n\frac{\mathbf{r}^t\mathbf{H}^{-2}\mathbf{r}}{\mathbf{r}^t\mathbf{H}^{-1}\mathbf{r}} - n \left(\mathrm{tr}(\mathbf{H}^{-1}) - n\frac{\mathbf{r}^t\mathbf{H}^{-2}\mathbf{r}}{\mathbf{r}^t\mathbf{H}^{-1}\mathbf{r}} \right) \right],
\end{aligned} \tag{19}$$

where

$$\begin{aligned}\Delta =&\mathbf{d}^t\mathbf{H}^{-2}\mathbf{d}\cdot\mathbf{1}^t\mathbf{H}^{-2}\mathbf{1}-\mathbf{d}^t\mathbf{H}^{-2}\mathbf{1}\cdot\mathbf{1}^t\mathbf{H}^{-2}\mathbf{d}\\ =&\mathrm{tr}(\mathbf{D}^2\mathbf{H}^{-2})\cdot\mathrm{tr}(\mathbf{H}^{-2})-\mathrm{tr}(\mathbf{D}\mathbf{H}^{-2})^2\\ =&\frac{1}{\eta^2}\left(\mathrm{tr}((\mathbf{H}-\mathbf{I})^2\mathbf{H}^{-2})\mathrm{tr}(\mathbf{H}^{-2})-\mathrm{tr}((\mathbf{H}-\mathbf{I})\mathbf{H}^{-2})^2\right)\\ =&\frac{1}{\eta^2}\left(n\ \mathrm{tr}(\mathbf{H}^{-2})-\mathrm{tr}(\mathbf{H}^{-1})^2\right).\end{aligned} \tag{20}$$

Putting together

$$\begin{aligned}\eta^\dagger=\hat{\gamma}/\hat{\mu}=\eta&\frac{n\ \mathrm{tr}(\mathbf{H}^{-2})-\mathrm{tr}(\mathbf{H}^{-1})\ n\ \frac{\mathbf{r}^t\mathbf{H}^{-2}\mathbf{r}}{\mathbf{r}^t\mathbf{H}^{-1}\mathbf{r}}}{n\ \mathrm{tr}(\mathbf{H}^{-2})-\mathrm{tr}(\mathbf{H}^{-1})\ n\ \frac{\mathbf{r}^t\mathbf{H}^{-2}\mathbf{r}}{\mathbf{r}^t\mathbf{H}^{-1}\mathbf{r}}-n\left(\mathrm{tr}(\mathbf{H}^{-1})-n\frac{\mathbf{r}^t\mathbf{H}^{-2}\mathbf{r}}{\mathbf{r}^t\mathbf{H}^{-1}\mathbf{r}}\right)}\\ =\eta&\left[1+\frac{n\left(\mathrm{tr}(\mathbf{H}^{-1})-n\ \frac{\mathbf{r}^t\mathbf{H}^{-2}\mathbf{r}}{\mathbf{r}^t\mathbf{H}^{-1}\mathbf{r}}\right)}{n\ \mathrm{tr}(\mathbf{H}^{-2})-\mathrm{tr}(\mathbf{H}^{-1})\ n\ \frac{\mathbf{r}^t\mathbf{H}^{-2}\mathbf{r}}{\mathbf{r}^t\mathbf{H}^{-1}\mathbf{r}}-n\left(\mathrm{tr}(\mathbf{H}^{-1})-n\frac{\mathbf{r}^t\mathbf{H}^{-2}\mathbf{r}}{\mathbf{r}^t\mathbf{H}^{-1}\mathbf{r}}\right)}\right]\\ =\eta&+\frac{2\eta^2}{n}f'_{ml}(\eta)\frac{1}{\frac{1}{n}\mathrm{tr}(\mathbf{H}^{-2})-\frac{1}{n^2}\mathrm{tr}(\mathbf{H}^{-1})\ n\ \frac{\mathbf{r}^t\mathbf{H}^{-2}\mathbf{r}}{\mathbf{r}^t\mathbf{H}^{-1}\mathbf{r}}-\frac{1}{n}\left(\mathrm{tr}(\mathbf{H}^{-1})-n\frac{\mathbf{r}^t\mathbf{H}^{-2}\mathbf{r}}{\mathbf{r}^t\mathbf{H}^{-1}\mathbf{r}}\right)}\\ =\eta&+\frac{2\eta^2}{n}f'_{ml}(\eta)\frac{1}{V(\mathbf{H}^{-1})+\frac{1}{n^2}\left(\mathrm{tr}(\mathbf{H}^{-1})-n\right)\left(\mathrm{tr}(\mathbf{H}^{-1})-n\frac{\mathbf{r}^t\mathbf{H}^{-2}\mathbf{r}}{\mathbf{r}^t\mathbf{H}^{-1}\mathbf{r}}\right)}\\ =\eta&+\frac{2\eta^2}{n}f'_{ml}(\eta)\frac{1}{V(\mathbf{H}^{-1})+\frac{2\eta}{n^2}\left(\mathrm{tr}(\mathbf{H}^{-1})-n\right)f'_{ml}(\eta)}\\ =\eta&+\frac{2\eta^2}{n}f'_{ml}(\eta)\frac{1}{V(\mathbf{H}^{-1})+v_n},\end{aligned} \tag{21}$$

where $v_n=\frac{2\eta}{n^2}\left(\mathrm{tr}(\mathbf{H}^{-1})-n\right)f'_{ml}(\eta)$, which can be safely ignored when n is large and/or $f'_{ml}(\eta)$ is small. Alternatively, noting $\hat{\tau}=\frac{n}{\mathbf{r}^t\mathbf{H}^{-1}\mathbf{r}}$, and we get

$$\begin{aligned}\eta^\dagger=\hat{\gamma}\hat{\tau}=\eta&\frac{n\ \mathrm{tr}(\mathbf{H}^{-2})-\mathrm{tr}(\mathbf{H}^{-1})\ n\ \frac{\mathbf{r}^t\mathbf{H}^{-2}\mathbf{r}}{\mathbf{r}^t\mathbf{H}^{-1}\mathbf{r}}}{n\ \mathrm{tr}(\mathbf{H}^{-2})-\mathrm{tr}(\mathbf{H}^{-1})^2}\\ =\eta&\left[1+\frac{\mathrm{tr}(\mathbf{H}^{-1})^2-\mathrm{tr}(\mathbf{H}^{-1})\ n\ \frac{\mathbf{r}^t\mathbf{H}^{-2}\mathbf{r}}{\mathbf{r}^t\mathbf{H}^{-1}\mathbf{r}}}{n\ \mathrm{tr}(\mathbf{H}^{-2})-\mathrm{tr}(\mathbf{H}^{-1})^2}\right]\\ =\eta&+\frac{2\eta^2}{n}f'_{ml}(\eta)\frac{1}{V(\mathbf{H}^{-1})}\cdot\frac{\mathrm{tr}(\mathbf{H}^{-1})}{n}.\end{aligned} \tag{22}$$

Finally, note that $\mathbf{H}=\eta\mathbf{D}+\mathbf{I}_n$ and since $\eta>0$ and $Diag(\mathbf{D})>0$ so $Diag(\mathbf{H})>0$. Denote

h_j the j-th diagonal element of $\mathbf{H}^{-1}$, we have

$$\begin{aligned} n^2 V(\mathbf{H}^{-1}) &= n\,\mathrm{tr}(\mathbf{H}^{-2}) - \mathrm{tr}(\mathbf{H}^{-1})\mathrm{tr}(\mathbf{H}^{-1}) \\ &= \sum_{ij} h_i^2 - \sum h_i h_j \\ &= \frac{1}{2}\left(\sum_{ij}(h_i^2 + h_j^2) - \sum_{i,j} 2h_i h_j\right) \\ &= \sum_{i,j}(h_i - h_j)^2 > 0. \end{aligned} \tag{23}$$

□

8.5 Proof of Theorem 1

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

$$\begin{aligned} \mathbf{y}_\mathbf{Q}^t \mathbf{P}_\mathbf{x} \mathbf{D} \mathbf{P}_\mathbf{x} \mathbf{D} \mathbf{P}_\mathbf{x} \mathbf{y}_\mathbf{Q} &= \mathbf{y}_\mathbf{Q}^t \mathbf{H}^{-1} \mathbf{P} \mathbf{D} \mathbf{P}_\mathbf{x} \mathbf{D} \mathbf{H}^{-1} \mathbf{P} \mathbf{y}_\mathbf{Q} \\ &= \mathbf{r}^t \mathbf{H}^{-1} \mathbf{D} \mathbf{P}_\mathbf{x} \mathbf{D} \mathbf{H}^{-1} \mathbf{r} \\ &= \frac{1}{\eta^2} \mathbf{r}^t \mathbf{H}^{-1} (\mathbf{H} - I_n) \mathbf{P}_x (\mathbf{H} - \mathbf{I}_n) \mathbf{H}^{-1} \mathbf{r} \\ &= \frac{1}{\eta^2}\left(\mathbf{r}^t \mathbf{P}_x \mathbf{r} - \mathbf{r}^t \mathbf{P}_x \mathbf{H}^{-1} \mathbf{r} - \mathbf{r}^t \mathbf{H}^{-1} \mathbf{P}_\mathbf{x} \mathbf{r} + \mathbf{r}^t \mathbf{H}^{-1} \mathbf{P}_\mathbf{x} \mathbf{H}^{-1} \mathbf{r}\right) \\ &= \frac{1}{\eta^2}\left(\mathbf{r}^t \mathbf{H}^{-1} \mathbf{r} - 2\mathbf{r}^t \mathbf{H}^{-2} \mathbf{r} + \mathbf{r}^t \mathbf{H}^{-1} \mathbf{P}_\mathbf{x} \mathbf{H}^{-1} r\right), \end{aligned} \tag{24}$$

where in the last equality, we used $\mathbf{r}^t \mathbf{P}_x \mathbf{H}^{-1} \mathbf{r} = \mathbf{r}^t \mathbf{P}^t \mathbf{H}^{-1} \mathbf{H}^{-1} \mathbf{r} = \mathbf{r}^t \mathbf{H}^{-2} \mathbf{r}$. Combing this and 17 to transform the second derivate to get

$$
\begin{aligned}
f''_{ml}(\eta) &= \frac{1}{2}\mathrm{tr}(\mathbf{H}^{-1}\mathbf{D}\mathbf{H}^{-1}\mathbf{D}) - \frac{n}{1}\frac{\mathbf{y}_{\mathbf{Q}}^t\mathbf{P}_x\mathbf{D}\mathbf{P}_x\mathbf{D}\mathbf{P}_{\mathbf{x}}\mathbf{y}_{\mathbf{Q}}}{\mathbf{y}_{\mathbf{Q}}^t\mathbf{P}_{\mathbf{x}}\mathbf{y}_{\mathbf{Q}}} + \frac{n}{2}\frac{(\mathbf{y}_{\mathbf{Q}}^t\mathbf{P}_{\mathbf{x}}\mathbf{D}\mathbf{P}_{\mathbf{x}}\mathbf{y}_{\mathbf{Q}})^2}{(\mathbf{y}_{\mathbf{Q}}^t\mathbf{P}_{\mathbf{x}}\mathbf{y}_{\mathbf{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\mathbf{H}^{-1}\mathbf{r} - 2\mathbf{r}^t\mathbf{H}^{-2}\mathbf{r} + \mathbf{r}^t\mathbf{H}^{-1}\mathbf{P}_{\mathbf{x}}\mathbf{H}^{-1}\mathbf{r}}{\mathbf{r}^t\mathbf{H}^{-1}\mathbf{r}} + \frac{n}{2\eta^2}\left(\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}}\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\mathbf{H}^{-2}\mathbf{r}}{\mathbf{r}^t\mathbf{H}^{-1}\mathbf{r}} - 2\frac{\mathbf{r}^t\mathbf{H}^{-1}\mathbf{P}_{\mathbf{x}}\mathbf{H}^{-1}\mathbf{r}}{\mathbf{r}^t\mathbf{H}^{-1}\mathbf{r}} + \left(\frac{\mathbf{r}^t\mathbf{H}^{-2}\mathbf{r}}{\mathbf{r}^t\mathbf{H}^{-1}\mathbf{r}}\right)^2\right) \\
&= \frac{1}{2\eta^2}\left(\mathrm{tr}(\mathbf{H}^{-2}) - 2n\frac{\mathbf{r}^t\mathbf{H}^{-3}\mathbf{r}}{\mathbf{r}^t\mathbf{H}^{-1}\mathbf{r}} + n\left(\frac{\mathbf{r}^t\mathbf{H}^{-2}\mathbf{r}}{\mathbf{r}^t\mathbf{H}^{-1}\mathbf{r}}\right)^2 + 2n\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}}\right) \\
&= \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\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}}\right) + 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}}\right].
\end{aligned}
\tag{25}
$$

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 23 the first term can be transformed to

$$
\begin{aligned}
\frac{1}{n}\mathrm{tr}(\mathbf{H}^{-2}) - \frac{1}{n^2}\mathrm{tr}(\mathbf{H}^{-1})^2 &= \frac{1}{2n^2}\sum_{i,j}(h_i - h_j)^2 \\
&= \frac{1}{2n^2}\sum_{ij}(h_i - \theta + \theta - h_j)^2 \\
&= \frac{1}{2n^2}\sum_{ij}(h_i - \theta)^2 + \frac{1}{2n^2}\sum_{ij}(\theta - h_j)^2 + \frac{1}{n^2}\sum_{ij}(h_i - \theta)(h_j - \theta) \\
&= \frac{1}{n}\sum_{i}(h_i - \theta)^2,
\end{aligned}
\tag{26}
$$

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

$$
\begin{aligned}
2\left(\frac{\mathbf{r}^t\mathbf{H}^{-2}\mathbf{r}}{\mathbf{r}^t\mathbf{H}^{-1}\mathbf{r}}\right)^2 - 2\frac{\mathbf{r}^t\mathbf{H}^{-3}\mathbf{r}}{\mathbf{r}^t\mathbf{H}^{-1}\mathbf{r}} =& 2\left(\frac{\mathbf{r}^t\mathbf{H}^{-2}\mathbf{r}\cdot\mathbf{r}^t\mathbf{H}^{-2}\mathbf{r} - \mathbf{r}^t\mathbf{H}^{-3}\mathbf{r}\cdot\mathbf{r}^t\mathbf{H}^{-1}\mathbf{r}}{\mathbf{r}^t\mathbf{H}^{-1}\mathbf{r}\cdot\mathbf{r}^t\mathbf{H}^{-1}\mathbf{r}}\right) \\
&= 2\frac{\sum_{i,j}\left(h_i^2h_j^2r_i^2r_j^2 - h_i^3h_jr_i^2r_j^2\right)}{\sum_{i,j}h_ih_jr_i^2r_j^2} \\
&= 2\frac{\sum_{i,j}\left(h_ih_jr_i^2r_j^2(h_ih_j - h_i^2)\right)}{\sum_{i,j}h_ih_jr_i^2r_j^2} \\
&= \frac{\sum_{i,j}\left(h_ih_jr_i^2r_j^2(2h_ih_j - h_i^2 - h_j^2)\right)}{\sum_{i,j}h_ih_jr_i^2r_j^2} \\
&= -\frac{\sum_{i,j}(h_i-h_j)^2s_i^2s_j^2}{\sum_{i,j}s_i^2s_j^2} \\
&= -\frac{\sum_{i,j}(h_i-\theta+\theta-h_j)^2s_i^2s_j^2}{\sum_{i,j}s_i^2s_j^2} \\
&= -\frac{\sum_{i,j}\left[(h_i-\theta)^2+(\theta-h_j)^2+2(h_i-\theta)(\theta-h_j)\right]s_i^2s_j^2}{\sum_{i,j}s_i^2s_j^2} \\
&= -2\frac{\sum_i(h_i-\theta)^2s_i^2}{\sum_i s_i^2} + 2\frac{\sum_i(h_i-\theta)s_i^2}{\sum_i s_i^2}\frac{\sum_j(h_j-\theta)s_j^2}{\sum_j s_j^2},
\end{aligned}
\tag{27}
$$

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 $E[|Z_i-(h_i-\theta)/n|^3] = |h_i-\theta|^3E(|B_i-\frac{1}{n}|^3) = |h_i-\theta|^3O(\frac{1}{n^3})$, 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

$$
\begin{aligned}
2\frac{\mathbf{r}^t\mathbf{H}^{-1}\mathbf{H}^{-1}\mathbf{P}_0\mathbf{H}^{-1}\mathbf{r}}{\mathbf{r}^t\mathbf{H}^{-1}\mathbf{r}} &= 2n\frac{\mathbf{r}^t(\mathbf{H}^{-1}-\theta\mathbf{I}_n)\mathbf{H}^{-1}\mathbf{P}_0(\mathbf{H}^{-1}-\theta\mathbf{I}_n)r}{\mathbf{r}^t\mathbf{H}^{-1}\mathbf{r}} \\
&= 2\frac{\mathbf{r}^t(\mathbf{H}^{-1}-\theta\mathbf{I}_n)\mathbf{H}^{-\frac{1}{2}}U\Lambda U^t\mathbf{H}^{-\frac{1}{2}}(\mathbf{H}^{-1}-\theta\mathbf{I}_n)\mathbf{r}}{\mathbf{r}^t\mathbf{H}^{-1}\mathbf{r}} \\
&= 2\frac{\sum_{j=1}^{c}(\mathbf{r}^t(\mathbf{H}^{-1}-\theta\mathbf{I}_n)\mathbf{H}^{-1/2}u_{\cdot j})^2}{\mathbf{r}^t\mathbf{H}^{-1}\mathbf{r}} \\
&= 2\frac{\sum_{j=1}^{c}\left(\sum_i r_i h_i^{1/2}(h_i-\theta)u_{ij}\right)^2}{\sum_i r_i^2 h_i} \\
&= 2\frac{\sum_{j=1}^{c} w_j^2}{\sum_i r_i^2 h_i} \\
&< 2\frac{\chi_c^2}{\chi_n^2}
\end{aligned}
\tag{28}
$$

where 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, let $F = \frac{\chi_c^2}{\chi_n^2}$, then nF follows F-distribution with d.f. 1 and n, whose mean and variance is $O(1)$, and thus $F = O(\frac{1}{n})$ goes to 0 as n increases. Putting together,

$$
f''_{ml}(\eta) = \frac{n}{2\eta^2}\left[-V(\mathbf{H}^{-1}) + \epsilon_n\right], \tag{29}
$$

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