Submitted to the Annals of Applied Statistics

ITERATIVE DISPERSION UPDATE ACHIEVES GLOBAL OPTIMUM FITTING LINEAR MIXED MODEL IN GENETIC ASSOICATIONS

BY YONGTAO GUAN[1,2,a] AND DANIEL LEVY[1,2,b]

[1] *Framingham Heart Study,* , [a] *grant.guan@nih.gov*

[2] *National Heart, Lung, and Blood Institute,* [b] *levyd@nhlbi.nih.gov*

In genetic association studies, the linear mixed model (LMM) has become a standard practice to account for population stratification and relatedness in the samples to reduce false positives. Despite recent progress, fitting LMMs to achieve an exact optimum has no theoretical assurance and remains computationally demanding, particularly in multi-omics studies where tens of thousands of phenotypes are tested for association with millions of genetic markers. The state-of-the-art method used the Newton-Raphson algorithm for optimization. Here we present IDUL that uses iterative dispersion updates to fit LMMs. IDUL requires no evaluation of derivatives of the likelihood function, rather, it fits two weighted least square regressions in each iteration. Applied to a set of protein assays from the Framingham Heart Study, IDUL converged in several iterations and provided consistent estimates between different initial values, which outperformed the Newton-Raphson method. IDUL is highly efficient for additional phenotypes, making it ideal to study genetic determinant of multi-omics phenotypes. Through a theoretical lens, we show that IDUL is a gradient ascent algorithm, and is as efficient as the Newton-Raphson method near the optimum. Most significantly, with a sufficiently large sample, IDUL converges to the global optimum with probability one, achieving the exact optimum. IDUL is freely available at http://www.haplotype.org.

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), and recent works focused on improving efficiency, as exemplified by EMMAX Kang et al. (2010), P3D Zhang et al. (2010), FaST-LMM Lippert et al. (2011), GEMMA Zhou and Stephens (2012), and BOLT-LMM Loh et al. (2015). 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 using approximation in the computation. For example, FaST-LMM, as a side note, proposed to approximate the genetic similarity matrix using a low-rank spectral decomposition, reducing the computational complexity of the mixed model equations. EMMAX and P3D used parameters estimated from the null model for each genetic variant (i.e. single nucleotide polymorphism; SNP) without repeated model fitting under the alternatives. 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. All these approximations produced different ranking of test statistics and hence a potential power loss.

Keywords and phrases: IDUL, gradient ascend, Newton-Raphson, multi-omics.

Another way to improve efficiency is algorithmic innovation. For example, FaST-LMM first rotates the genotypes and phenotypes according to eigenvectors of the genetic relatedness matrix so that the rotated data become uncorrelated. Then it optimizes a single parameter in the variance component using Brent's method. GEMMA also rotates genotypes and phenotypes during optimization, but it does so implicitly. The innovation of GEMMA is its ability to efficiently evaluate the second derivatives, so that it can use the Newton-Raphson method for optimization, which is quadratic and more efficient than Brent's method that is super linear. 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 Iterative Dispersion Update for the Linear mixed model (IDUL) to fit standard LMMs in genetic association studies. The IDUL aims to solve the inconsistency (or starting-point-dependency) observed in the Newton-Raphson algorithm, improves computation efficiency, and makes the LMM 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. Our theoretical analysis revealed that IDUL is a gradient ascent algorithm, and near the optimum it is as efficient as the Newton-Raphson method. Most importantly, we show that the optimum that IDUL attains is the global optimum with probability one as the sample size increases. Therefore, IDUL achieves an exact optimum for LMM in genetic association studies.

2. Results. 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 $\mathbf{D}$. Multiply both sides of (1) by $\mathbf{Q}^t$ to get

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

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. This rotation was done explicitly or implicitly in previous works such as Lippert et al. (2011); Zhou and Stephens (2012) to improve computation efficiency.

Our new iterative algorithm repeats the following two steps until η converges: 1) 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 2) 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 $\mathbf{r}$ (hence the name IDUL). 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 $V(\mathbf{H}^{-1}) = \text{tr}(\mathbf{H}^{-2})/n - \text{tr}(\mathbf{H}^{-1})^2/n^2 > 0$, $v_n = \frac{2\eta}{n^2}\left(\text{tr}(\mathbf{H}^{-1}) - n\right) f'_{ml}(\eta)$, and $f'_{ml}(\eta)$ is the first derivative of the log-likelihood evaluated at η, and can be computed analytically (Methods).

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 when the step size $|\eta^{\dagger} - \eta|$ is not too large, 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, and the second derivative in the neighborhood is negative, the update can overshoot and decrease the likelihood. The overshoot can be resolved by making IDUL idle (Methods).

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 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 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 and its derivatives, there must exist a minimum such that $f'_{ml}(\eta) = 0$ but $f''_{ml}(\eta) > 0$, which produces a contradiction.

When both v_n in Equation (3) and ϵ_n in Equation (4) can be safely ignored (such as when n is large), 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 (even the first derivative can be avoided), 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.

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.

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 $\eta_0 = h_0/(1 - h_0)$.

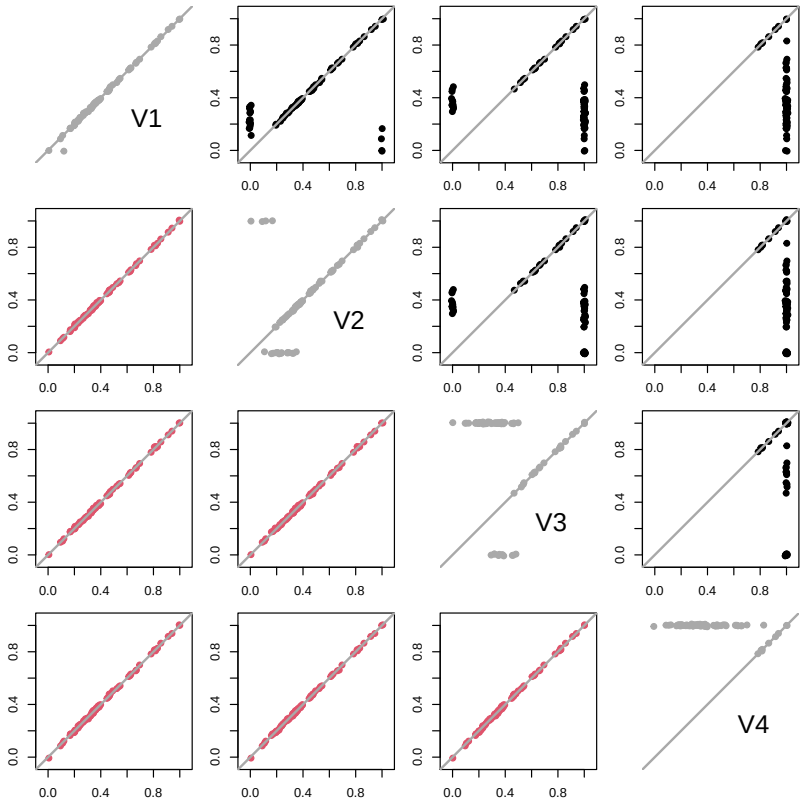

FIG 1. *Consistency of IDUL and Newton-Raphson. IDUL and Newton-Raphson were run on the same sets of initial values. For each phenotype, four initial values were generated with seeds randomly selected from four segments. So both IDUL and Newton-Raphson produced* 4 *columns of estimates each with* 79 *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. 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.*

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 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 Auton et al. (2015) (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.

With maximum likelihood update (3) at hand, we can obtain REML update by substituting $f'_{ml}(\eta)$ with $f'_{re}(\eta)$, and computing $f'_{re}(\eta)$ can be done efficiently (Methods). To test the null

hypothesis $\beta = 0$, we computed the likelihood ratio test (LRT) p-values using ML 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). We implemented both LRT and Wald test in IDUL and compared with GEMMA, a software package for LMM that implements the Newton-Raphson method for optimization.

For both LRT and Wald test p-values, IDUL and GEMMA reached perfect agreement (Supplementary Figure S1). This suggested that with multiple starting points and primer runs using Brent's method, GEMMA can reliably achieves optimum for both ML and REML estimates, albeit at some extra computation costs (Table 1). It is difficult to make a fair comparison on algorithms when the details of implementation contribute significantly to CPU and wall time. But if we ignore the dramatic wall time differences between IDUL and GEMMA, and focus on the total CPU time spent, IDUL spent 58% CPU time as GEMMA for LRT and 51% CPU time for Wald test. Most strikingly, for additional 9 phenotypes, IDUL only spends 11% (1.6 minutes) extra wall time for LRT; for additional 78 phenotypes, 43% (6 minutes) extra wall time. 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.

		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, however, GEMMA only used multithreading in part of its computation. On the other hand, IDUL implemented SNP level parallel computation. Taking multiple phenotypes for sequential analysis was not implemented in GEMMA, and its wall time for 10 *and* 79 *phenotypes are missing.*

3. Discussion. We documented two key contributions in this paper. First, we introduced 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 update near the optimal point, but avoided the inconsistency that Newton-Raphson suffers when the initial point is distant from the optimum. In addition, IDUL is a gradient ascent algorithm as each update it increases the likelihood according to the first directive of the likelihood. 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 more expensive to compute. Second, we showed that the log-likelihood is convex at the optimum with a sufficiently large sample size. (We hypothesize that the same is true for REML likelihood, see Methods). A profound corollary is that the standard linear mixed model has a unique global optimum, which makes the log-likelihood convex. The convexity of the log-likelihood and IDUL being a gradient ascent algorithm jointly underpin its perfect consistency over different initial values. Thus IDUL can reliably achieve an exact optimum.

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 $O(n^3)$, for multiplying Q^t with all SNPs, all phenotypes, and all covariates is $O((m + p + c)n^2)$, and for optimization 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, Newton-Raphson tends to have a larger equivalent t than IDUL because 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.

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.

When testing fixed effects in LMM, MLE is preferred over the REML estimate, because ML estimates of fixed effects 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. 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 the height data from Yang et al. (2010), we performed down-sampling at $90\%, 70\%$, and 50% of total 3925 samples 100 times each, fit the LMM using IDUL for different estimates of kinship matrices without any issue (Supplementary Figure S3). As a comparison, 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, it is prudent to examine phenotypes and perform quantile normalization when they show severe departure from normality. IDUL is freely available at http://www.haplotype.org.

4. Methods. 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

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

(8)

$$f''_{ml}(\eta) = \frac{1}{2}\text{tr}(\mathbf{H}^{-1}\mathbf{D}\mathbf{H}^{-1}\mathbf{D}) - \frac{n}{2}\frac{(2\mathbf{y_Q^t}\mathbf{P}_x\mathbf{D}\mathbf{P}_x\mathbf{D}\mathbf{P}_x\mathbf{y_Q})(\mathbf{y}_Q^t\mathbf{P}_x\mathbf{y_Q}) - \left(\mathbf{y_Q^t}\mathbf{P}_x\mathbf{D}\mathbf{P}_x\mathbf{y_Q}\right)^2}{(\mathbf{y_Q^t}\mathbf{P}_x\mathbf{y_Q})^2}.$$

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

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

(10)

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

First we simplify the directive of likelihood functions using residuals from Step 1 of IDUL.

PROPOSITION 4.1. *Let* $\mathbf{r} = \mathbf{P}\mathbf{y_Q}$, *the following hold:*

$$f'_{ml}(\eta) = \frac{1}{2\eta}\left(tr(\mathbf{H}^{-1}) - n\frac{\mathbf{r}^t\mathbf{H}^{-2}\mathbf{r}}{\mathbf{r}^t\mathbf{H}^{-1}\mathbf{r}}\right), \tag{11}$$

$$f'_{re}(\eta) = \frac{1}{2\eta}\left(tr(\mathbf{H}^{-1}\mathbf{P}) - (n-c)\frac{\mathbf{r}^t\mathbf{H}^{-2}\mathbf{r}}{\mathbf{r}^t\mathbf{H}^{-1}\mathbf{r}}\right). \tag{12}$$

Because $\mathbf{H}$ is diagonal, evaluation of f'_{ml} has a complexity of $O(n)$. To evaluate f'_{re} note $\text{tr}(\mathbf{H}^{-1}\mathbf{P}) = \text{tr}(\mathbf{H}) - \text{tr}(\mathbf{H}\mathbf{P}_0)$, while $\text{tr}(\mathbf{H}\mathbf{P}_0) = \text{tr}(\mathbf{H}^{-1}\mathbf{X}(\mathbf{X}^t\mathbf{H}^{-1}\mathbf{X})^{-1}\mathbf{X}^t\mathbf{H}^{-1}) = \text{tr}(\frac{\mathbf{X}^t\mathbf{H}^{-2}\mathbf{X}}{\mathbf{X}^t\mathbf{H}^{-2}\mathbf{X}})$. 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 $O(n + nc + c^3)$.

Then we derive the analytic form of the IDUL by computing $\eta^\dagger = \hat{\gamma}/\hat{\mu}$ in Step 2.

LEMMA 4.2. *The update of IDUL for maximum likelihood estimate is*

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

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.

The Taylor expansion of f_{ml} at η is $f_{ml}(\eta+x) = f_{ml}(\eta) + f'_{ml}(\eta)x + \frac{1}{2}f''_{ml}(\xi)x^2$ where $\xi \in (\eta - x, \eta + x)$. 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 $f_{ml}(\eta^\dagger) - f_{ml}(\eta) > 0$. But when $f''_{ml}(\xi) < 0$ and $|\eta^\dagger - \eta|$ is large, then we may overshoot, landing on the other side of the mode. 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

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

Compare updates 14 with 13, the update $\eta_2^\dagger = \hat{\gamma}\hat{\tau}$ takes a smaller step as $\mathrm{tr}(\mathbf{H}^{-1})/n < 1$, but it is more stable compared to $\eta_1^\dagger = \hat{\gamma}/\hat{\mu}$ because the latter can overshoot causing oscillating behavior (such as when phenotype is strongly correlated with an leading eigenvector), and numerically $\hat{\mu}$ can be negative. These two updates can be combined to to get $\eta^\dagger = (\eta_1^\dagger + \eta_2^\dagger)/2$. We call this making IDUL idle.

With maximum likelihood update (13) at hand, we can obtain REML update by substituting $f'_{ml}(\eta)$ with $f'_{re}(\eta)$, to get

$$\eta^\dagger = \eta + \frac{2\eta^2}{n}\frac{1}{V(\mathbf{H}^{-1})} f'_{re}(\eta), \tag{15}$$

where we took the liberty to remove the v_n term. Of course, following the udpate (14), we can get REML version of it by substituting $f'_{ml}(\eta)$ with $f'_{re}(\eta)$ to make the REML update idle.

Finally, we quantify the second derivative of the log-likelihood function.

THEOREM 4.3. *Let η be an optimum of log-likelihood function such that $f'_{ml}(\eta) = 0$, then*

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

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 for Proposition 4.1, Lemma 4.2, and Theorem 4.3 can be found in Supplementary. 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 4.3.1. *Let η be an optimum of log-likelihood function such that $f'_{re}(\eta) = 0$, then*

$$f''_{re}(\eta) = \frac{n-c}{2\eta^2}\left[-V(\mathbf{H}^{-1}\mathbf{P}) + \epsilon_n\right], \tag{17}$$

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.

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.

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).

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.

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.

5. Supplementary Materials.

5.1. *Figures.*

5.1.1. *IDUL and GEMMA.* From 5757 FHS samples we randomly selected $500,000$ common SNPs, and produced a genotype input in bimbam mean genotype format that can be used by both IDUL and GEMMA. The missing genotypes were replaced by the mean dosage of that SNP. Both IDUL and GEMMA used the kinship matrix that was previously estimated. The two phenotypes are quantile normalized HDL level and quantile normalized lipoprotein expression.

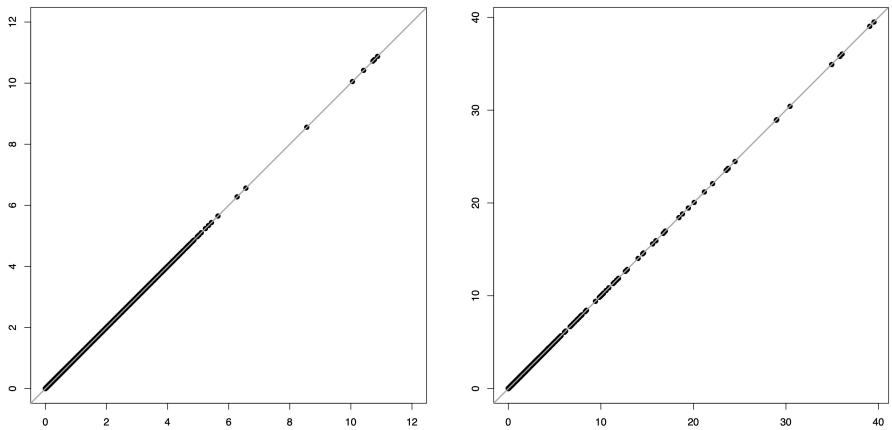

FIG S1. *Comparison of* $-\log_{10} p$ *between IDUL (x-axis) and GEMMA (y-axis). The left panel plot Wald test p-values for a quantile normalized HDL level, and the right panel is for LRT p-values for a quantile normalized lipoprotein expression. IDUL and GEMMA achieved identical results.*

5.1.2. *1000 Genomes dataset.* The 1000 Genomes project contains 2504 samples from diverse populations. While Sabre dataset in the main text is an example dataset that dominated by relatedness, the 1000 Genomes samples is an example dataset that is dominated by population stratification. We simulated 100 phenotypes with nominal heritability of 0.80. For each phenotype, we fit LMM using IDUL and Newton-Raphson with the same sets of initial values (see main text).

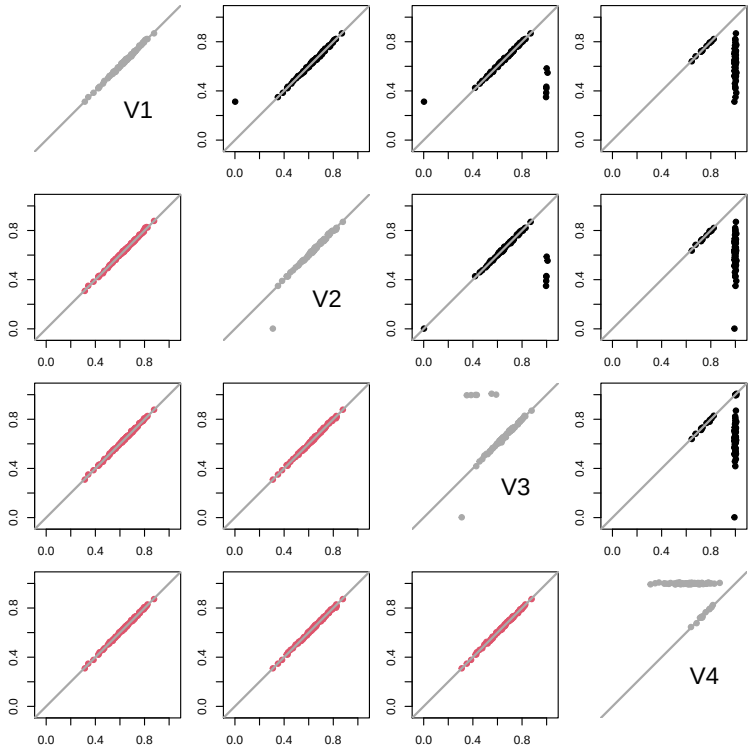

FIG S2. *Consistency of IDUL and Newton-Raphson. IDUL and Newton-Raphson were run on the same sets of initial values. For each phenotype, four initial values were generated with seeds randomly selected from four segments. So both IDUL and Newton-Raphson produced* 4 *columns of estimates each with* 100 *rows. The pairwise plots of the four columns of IDUL estimates were shown in the lower-triangle and colored in red. Those of Newton-Raphson estimates were shown in the upper-triangle and in black. 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.*

5.1.3. *IDUL is reliable for small samples.* We performed down-sampling studies of a dataset containing 3925 samples. The proportions for down-sampling are $90\%, 70\%,$ and 50%. For each proportion, we repeated the experiments 100 times, where in each experiments samples were drawn randomly without replacement. Heritabilities were estimated based on two different kinship estimates for each experiment.

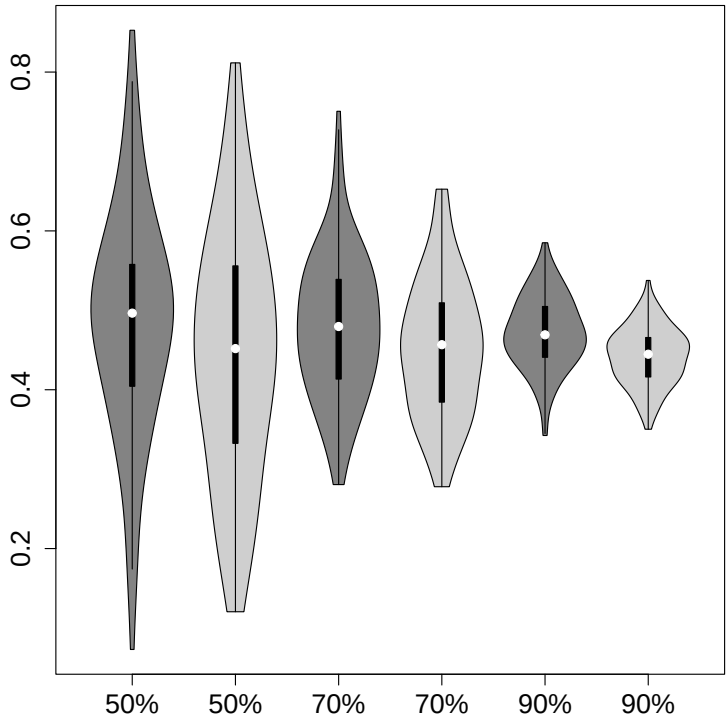

FIG S3. *Violin plots of heritability estimates using IDUL. This is to show feasibility of IDUL on small samples in heritability estimates. Darkgray is for one kinship estimate, and light gray is for another. The downsampling proportions are marked below violin plots. The total number of samples is* 3925.

5.2. *Proofs.*

5.2.1. *Proof of Proposition* 4.1.

PROOF. We first simplify two expressions

$$\begin{aligned} \mathbf{y_Q^t P_x y_Q} &= \mathbf{r^t H^{-1} r} \\ \mathbf{y_Q^t P_x D P_x y_Q} &= \frac{1}{\eta}(\mathbf{r^t H^{-1} r} - \mathbf{r^t H^{-2} r}) \end{aligned} \tag{18}$$

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:

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

and

$$\begin{aligned} f'_{re}(\eta) &= -\frac{1}{2}\text{tr}(\mathbf{P}_x \mathbf{D}) + \frac{n-c}{2}\frac{\mathbf{y_Q^t P}_x \mathbf{D P}_x \mathbf{y_Q}}{\mathbf{y_Q^t P}_x \mathbf{y_Q}} \\ &= -\frac{1}{2}\frac{1}{\eta}\text{tr}(\mathbf{P^t H^{-1}}(\mathbf{H} - \mathbf{I}_n)) + \frac{n-c}{2}\frac{1}{\eta}\frac{\mathbf{r^t H^{-1} r} - \mathbf{r^t H^{-2} r}}{r^t \mathbf{H}^{-1} r} \\ &= \frac{1}{2\eta}(\text{tr}(\mathbf{P}) + \text{tr}(\mathbf{H^{-1} P}) + (n-c) - (n-c)\frac{\mathbf{r^t H^{-2} r}}{\mathbf{r^t H^{-1} r}}) \\ &= \frac{1}{2\eta}\left(\text{tr}(\mathbf{H^{-1} P}) - (n-c)\frac{\mathbf{r^t H^{-2} r}}{\mathbf{r^t H^{-1} r}}\right), \end{aligned} \tag{20}$$

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

□

5.2.2. *Proof of Lemma* 4.2.

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

$$
\begin{aligned}
(\hat{\gamma}, \hat{\mu})^t &= ((\mathbf{d}, \mathbf{1})^t \mathbf{H}^{-2} (\mathbf{d}, \mathbf{1}))^{-1} (\mathbf{d}, \mathbf{1})^t \mathbf{H}^{-2} \mathbf{s} \\
&= \begin{pmatrix} \mathbf{d}^t \mathbf{H}^{-2} \mathbf{d} & \mathbf{d}^t \mathbf{H}^{-2} \mathbf{1} \\ \mathbf{1}^t \mathbf{H}^{-2} \mathbf{d} & \mathbf{1}^t \mathbf{H}^{-2} \mathbf{1} \end{pmatrix}^{-1} (\mathbf{d}, \mathbf{1})^t \mathbf{H}^{-2} \mathbf{s} \\
&= \frac{1}{\Delta} \begin{pmatrix} \mathbf{1}^t \mathbf{H}^{-2} \mathbf{1} & -\mathbf{d}^t \mathbf{H}^{-2} \mathbf{1} \\ -\mathbf{1}^t \mathbf{H}^{-2} \mathbf{d} & \mathbf{d}^t \mathbf{H}^{-2} \mathbf{d} \end{pmatrix} (\mathbf{d}^t \mathbf{H}^{-2} \mathbf{s}, \mathbf{1}^t \mathbf{H}^{-2} \mathbf{s})^t \\
&= \frac{1}{\Delta} \begin{pmatrix} \mathbf{1}^t \mathbf{H}^{-2} \mathbf{1} \cdot \mathbf{d}^t \mathbf{H}^{-2} \mathbf{s} - \mathbf{d}^t \mathbf{H}^{-2} \mathbf{1} \cdot \mathbf{1}^t \mathbf{H}^{-2} \mathbf{s} \\ -\mathbf{1}^t \mathbf{H}^{-2} \mathbf{d} \cdot \mathbf{d}^t \mathbf{H}^{-2} \mathbf{s} + \mathbf{d}^t \mathbf{H}^{-2} \mathbf{d} \cdot \mathbf{1}^t \mathbf{H}^{-2} \mathbf{s} \end{pmatrix}.
\end{aligned} \tag{21}
$$

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

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

and

$$
\begin{aligned}
\hat{\mu} =& \frac{1}{\Delta} \left(-\mathbf{1}^t \mathbf{D} \mathbf{H}^{-2} \mathbf{1} \cdot \mathbf{r}^t \mathbf{D} \mathbf{H}^{-2} \mathbf{r} + \mathbf{1}^t \mathbf{D}^2 \mathbf{H}^{-2} \mathbf{1} \cdot \mathbf{r}^t \mathbf{H}^{-2} \mathbf{r} \right) \\
=& \frac{1}{\Delta} \frac{1}{\eta^2} \left(-\mathbf{1}^t \mathbf{H}^{-2} (\mathbf{H} - \mathbf{I}) \mathbf{1} \cdot \mathbf{r}^t (\mathbf{H} - \mathbf{I}) \mathbf{H}^{-2} \mathbf{r} + \mathbf{1}^t (\mathbf{H} - \mathbf{I})^2 \mathbf{H}^{-2} \mathbf{1} \cdot \mathbf{r}^t \mathbf{H}^{-2} \mathbf{r} \right) \\
=& \frac{1}{\Delta} \frac{1}{\eta^2} \left(-\mathrm{tr}(\mathbf{H}^{-1} - \mathbf{H}^{-2}) \cdot \mathbf{r}^t (\mathbf{H}^{-1} - \mathbf{H}^{-2}) \mathbf{r} + \mathrm{tr}(\mathbf{I}_n - 2\mathbf{H}^{-1} + \mathbf{H}^{-2}) \cdot \mathbf{r}^t \mathbf{H}^{-2} \mathbf{r} \right) \\
=& \frac{1}{\Delta} \frac{1}{\eta^2} \left(-\mathrm{tr}(\mathbf{H}^{-1}) \mathbf{r}^t \mathbf{H}^{-1} \mathbf{r} + n \mathbf{r}^t \mathbf{H}^{-2} \mathbf{r} + \mathrm{tr}(\mathbf{H}^{-2}) \mathbf{r}^t \mathbf{H}^{-1} \mathbf{r} - \mathrm{tr}(\mathbf{H}^{-1}) \mathbf{r}^t \mathbf{H}^{-2} \mathbf{r} \right) \\
=& \frac{1}{\Delta} \frac{1}{\eta^2} \frac{\mathbf{r}^t \mathbf{H}^{-1} \mathbf{r}}{n} \left[n \, \mathrm{tr}(\mathbf{H}^{-2}) - \mathrm{tr}(\mathbf{H}^{-1}) \, n \frac{\mathbf{r}^t \mathbf{H}^{-2} \mathbf{r}}{\mathbf{r}^t \mathbf{H}^{-1} \mathbf{r}} - n \left(\mathrm{tr}(\mathbf{H}^{-1}) - n \frac{\mathbf{r}^t \mathbf{H}^{-2} \mathbf{r}}{\mathbf{r}^t \mathbf{H}^{-1} \mathbf{r}} \right) \right],
\end{aligned} \tag{23}
$$

where

$$
\begin{aligned}
\Delta =& \mathbf{d}^t \mathbf{H}^{-2} \mathbf{d} \cdot \mathbf{1}^t \mathbf{H}^{-2} \mathbf{1} - \mathbf{d}^t \mathbf{H}^{-2} \mathbf{1} \cdot \mathbf{1}^t \mathbf{H}^{-2} \mathbf{d} \\
=& \mathrm{tr}(\mathbf{D}^2 \mathbf{H}^{-2}) \cdot \mathrm{tr}(\mathbf{H}^{-2}) - \mathrm{tr}(\mathbf{D} \mathbf{H}^{-2})^2 \\
=& \frac{1}{\eta^2} \left(\mathrm{tr}((\mathbf{H} - \mathbf{I})^2 \mathbf{H}^{-2}) \mathrm{tr}(\mathbf{H}^{-2}) - \mathrm{tr}((\mathbf{H} - \mathbf{I}) \mathbf{H}^{-2})^2 \right) \\
=& \frac{1}{\eta^2} \left(n \, \mathrm{tr}(\mathbf{H}^{-2}) - \mathrm{tr}(\mathbf{H}^{-1})^2 \right).
\end{aligned} \tag{24}
$$

Putting together

(25)
$$
\begin{aligned}
\eta^{\dagger} = \hat{\gamma}/\hat{\mu} &= \eta \frac{n\,\mathrm{tr}(\mathbf{H}^{-2}) - \mathrm{tr}(\mathbf{H}^{-1})\, n\, \frac{\mathbf{r}^t\mathbf{H}^{-2}\mathbf{r}}{\mathbf{r}^t\mathbf{H}^{-1}\mathbf{r}}}{n\,\mathrm{tr}(\mathbf{H}^{-2}) - \mathrm{tr}(\mathbf{H}^{-1})\, n\, \frac{\mathbf{r}^t\mathbf{H}^{-2}\mathbf{r}}{\mathbf{r}^t\mathbf{H}^{-1}\mathbf{r}} - n\left(\mathrm{tr}(\mathbf{H}^{-1}) - n\frac{\mathbf{r}^t\mathbf{H}^{-2}\mathbf{r}}{\mathbf{r}^t\mathbf{H}^{-1}\mathbf{r}}\right)} \\
&= \eta \left[1 + \frac{n\left(\mathrm{tr}(\mathbf{H}^{-1}) - n\, \frac{\mathbf{r}^t\mathbf{H}^{-2}\mathbf{r}}{\mathbf{r}^t\mathbf{H}^{-1}\mathbf{r}}\right)}{n\,\mathrm{tr}(\mathbf{H}^{-2}) - \mathrm{tr}(\mathbf{H}^{-1})\, n\, \frac{\mathbf{r}^t\mathbf{H}^{-2}\mathbf{r}}{\mathbf{r}^t\mathbf{H}^{-1}\mathbf{r}} - n\left(\mathrm{tr}(\mathbf{H}^{-1}) - n\frac{\mathbf{r}^t\mathbf{H}^{-2}\mathbf{r}}{\mathbf{r}^t\mathbf{H}^{-1}\mathbf{r}}\right)}\right] \\
&= \eta + \frac{2\eta^2}{n} f'_{ml}(\eta) \frac{1}{\frac{1}{n}\mathrm{tr}(\mathbf{H}^{-2}) - \frac{1}{n^2}\mathrm{tr}(\mathbf{H}^{-1})\, n\, \frac{\mathbf{r}^t\mathbf{H}^{-2}\mathbf{r}}{\mathbf{r}^t\mathbf{H}^{-1}\mathbf{r}} - \frac{1}{n}\left(\mathrm{tr}(\mathbf{H}^{-1}) - n\frac{\mathbf{r}^t\mathbf{H}^{-2}\mathbf{r}}{\mathbf{r}^t\mathbf{H}^{-1}\mathbf{r}}\right)} \\
&= \eta + \frac{2\eta^2}{n} f'_{ml}(\eta) \frac{1}{V(\mathbf{H}^{-1}) + \frac{1}{n^2}\left(\mathrm{tr}(\mathbf{H}^{-1}) - n\right)\left(\mathrm{tr}(\mathbf{H}^{-1}) - n\frac{\mathbf{r}^t\mathbf{H}^{-2}\mathbf{r}}{\mathbf{r}^t\mathbf{H}^{-1}\mathbf{r}}\right)} \\
&= \eta + \frac{2\eta^2}{n} f'_{ml}(\eta) \frac{1}{V(\mathbf{H}^{-1}) + \frac{2\eta}{n^2}\left(\mathrm{tr}(\mathbf{H}^{-1}) - n\right) f'_{ml}(\eta)} \\
&= \eta + \frac{2\eta^2}{n} f'_{ml}(\eta) \frac{1}{V(\mathbf{H}^{-1}) + v_n}
\end{aligned}
$$

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

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

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

(27)
$$
\begin{aligned}
n^2 V(\mathbf{H}^{-1}) &= n\,\mathrm{tr}(\mathbf{H}^{-2}) - \mathrm{tr}(\mathbf{H}^{-1})\mathrm{tr}(\mathbf{H}^{-1}) \\
&= \sum_{ij} h_i^2 - \sum h_i h_j \\
&= \frac{1}{2}\left(\sum_{ij}(h_i^2 + h_j^2) - \sum_{i,j} 2h_i h_j\right) \\
&= \sum_{i,j}(h_i - h_j)^2 > 0
\end{aligned}
$$

□

5.2.3. *Proof of Theorem* 4.3.

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:

$$
\begin{aligned}
\mathbf{y_Q^t P_x D P_x D P_x y_Q} &= \mathbf{y_Q^t H^{-1} P D P_x D H^{-1} P y_Q} \\
&= \mathbf{r}^t \mathbf{H}^{-1} \mathbf{D} \mathbf{P_x} \mathbf{D} \mathbf{H}^{-1} \mathbf{r} \\
&= \frac{1}{\eta^2} \mathbf{r}^t \mathbf{H}^{-1} (\mathbf{H} - I_n) \mathbf{P}_x (\mathbf{H} - \mathbf{I}_n) \mathbf{H}^{-1} \mathbf{r} \\
&= \frac{1}{\eta^2} \left(\mathbf{r}^t \mathbf{P}_x \mathbf{r} - \mathbf{r}^t \mathbf{P}_x \mathbf{H}^{-1} \mathbf{r} - \mathbf{r}^t \mathbf{H}^{-1} \mathbf{P_x} \mathbf{r} + \mathbf{r}^t \mathbf{H}^{-1} \mathbf{P_x} \mathbf{H}^{-1} \mathbf{r} \right) \\
&= \frac{1}{\eta^2} \left(\mathbf{r}^t \mathbf{H}^{-1} \mathbf{r} - 2 \mathbf{r}^t \mathbf{H}^{-2} \mathbf{r} + \mathbf{r}^t \mathbf{H}^{-1} \mathbf{P_x} \mathbf{H}^{-1} r \right),
\end{aligned} \tag{28}
$$

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 (18) 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_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 \mathbf{H}^{-1} \mathbf{r} - 2 \mathbf{r}^t \mathbf{H}^{-2} \mathbf{r} + \mathbf{r}^t \mathbf{H}^{-1} \mathbf{P_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_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{29}
$$

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 (27) the first term can be transformed to

$$
\begin{aligned}
\frac{1}{n} \mathrm{tr}(\mathbf{H}^{-2}) - \frac{1}{n^2} \mathrm{tr}(\mathbf{H}^{-1})^2 &= \frac{1}{2n^2} \sum_{i,j} (h_i - h_j)^2 \\
&= \frac{1}{2n^2} \sum_{ij} (h_i - \theta + \theta - h_j)^2 \\
&= \frac{1}{2n^2} \sum_{ij} (h_i - \theta)^2 + \frac{1}{2n^2} \sum_{ij} (\theta - h_j)^2 + \frac{1}{n^2} \sum_{ij} (h_i - \theta)(h_j - \theta) \\
&= \frac{1}{n} \sum_i (h_i - \theta)^2,
\end{aligned} \tag{30}
$$

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

(31)

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

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

$$
\begin{aligned}
2\frac{\mathbf{r}^t\mathbf{H}^{-1}\mathbf{H}^{-1}\mathbf{P}_0\mathbf{H}^{-1}\mathbf{r}}{\mathbf{r}^t\mathbf{H}^{-1}\mathbf{r}} &= 2n\frac{\mathbf{r}^t(\mathbf{H}^{-1}-\theta\mathbf{I}_n)\mathbf{H}^{-1}\mathbf{P}_0(\mathbf{H}^{-1}-\theta\mathbf{I}_n)r}{\mathbf{r}^t\mathbf{H}^{-1}\mathbf{r}} \\
&= 2\frac{\mathbf{r}^t(\mathbf{H}^{-1}-\theta\mathbf{I}_n)\mathbf{H}^{-\frac{1}{2}}U\Lambda U^t\mathbf{H}^{-\frac{1}{2}}(\mathbf{H}^{-1}-\theta\mathbf{I}_n)\mathbf{r}}{\mathbf{r}^t\mathbf{H}^{-1}\mathbf{r}} \\
&= 2\frac{\sum_{j=1}^{c}(\mathbf{r}^t(\mathbf{H}^{-1}-\theta\mathbf{I}_n)\mathbf{H}^{-1/2}u_{\cdot j})^2}{\mathbf{r}^t\mathbf{H}^{-1}\mathbf{r}} \\
&= 2\frac{\sum_{j=1}^{c}\left(\sum_i r_i h_i^{1/2}(h_i-\theta)u_{ij}\right)^2}{\sum_i r_i^2 h_i} \\
&= 2\frac{\sum_{j=1}^{c} w_j^2}{\sum_i r_i^2 h_i} \\
&< 2\frac{\chi_c^2}{\chi_n^2}
\end{aligned}
\tag{32}
$$

where the first equality holds because $\mathbf{r} = (\mathbf{I}_n - \mathbf{P}_0)\mathbf{y}_\mathbf{Q}$ andd $\mathbf{r}^t\mathbf{H}^{-1}\mathbf{P}_0\mathbf{v} = 0$ 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,

$$
f''_{ml}(\eta) = \frac{n}{2\eta^2}\left[-V(\mathbf{H}^{-1}) + \epsilon_n\right], \tag{33}
$$

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 $f'_{ml}(\eta) = 0$. □