Bolt-lmm

Friday, January 26, 2024 1:16 PM

ReML

reml.pdf (xiuming.info)

Consider a general linear regression model:

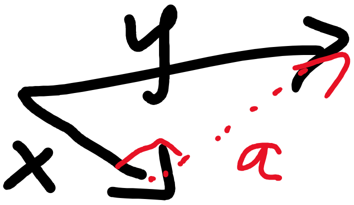
$$y = X\beta + \epsilon$$

Where y is the phenotype, X is the genotype matrix (N inds × M snps) , and ϵ is the error term with the distribution $N(0, H(\theta))$, and $H(\theta)$ is a covariance matrix by parameter θ.

By using Maximum Likelihood, we can get:

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

$$\sigma^2 = \frac{1}{N}(y - X\hat{\beta})^T (y - X\hat{\beta})$$



We set A = $X(X^T X)^{-1} X^T$, then Ay = $X\hat{\beta} = \hat{y}$

If vector a is orthogonal with all columns in X, then $a^T X = 0$, and $a^T y$ is called **error contrast** ω.

$$\omega = A^T y = A^T(X\beta + \epsilon) = A^T\epsilon \sim N(0, A^T H(\theta) A)$$

So we can estimate θ by using Restrict MLE, and we don't need to estimate β, so it becomes **unbiased**.

$$L_\omega(\theta | A^T y)$$

Mahalanobis Distance

If a covariance matrix is:

$$\Sigma = \begin{matrix} cov(X,X) & cov(X,Y) & cov(X,Z) \\ cov(Y,X) & cov(Y,Y) & cov(Y,Z) \\ cov(Z,X) & cov(Z,Y) & cov(Z,Z) \end{matrix}$$

Then **the difference between two variables X, Y** which are in the same distribution and covariance matrix = Σ :

$$D(X, Y) = \sqrt{(X - Y)^T \Sigma^{-1} (X - Y)}$$

Specifically, if Σ is an identical matrix, it becomes the Euclidean Distance.

We minimize the squared Mahalanobis Distance to the residual: Y - Xβ, and we get:

$$\hat{\beta} = (X^T H^{-1} X)^{-1} X^T H^{-1} y$$

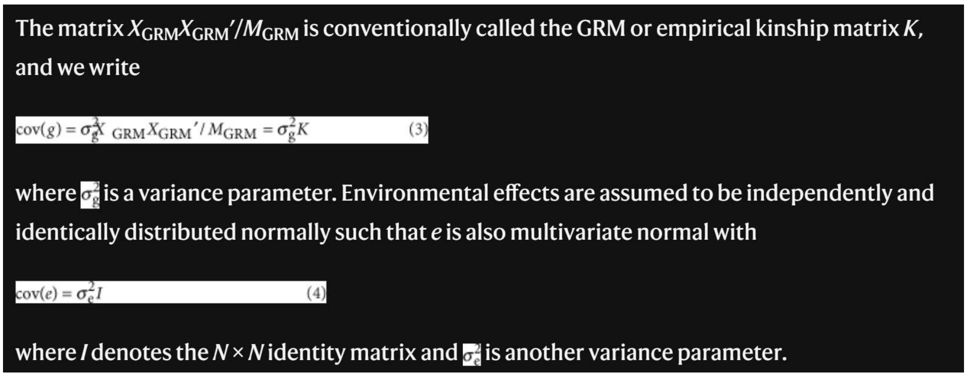
Bolt-lmm

In Bolt-lmm paper, which is:

$$\chi^2_{\text{LMM}} = \frac{(x'_{\text{test}} V^{-1} y)^2}{x'_{\text{test}} V^{-1} x_{\text{test}}} \qquad (5)$$

Where V gotten from X_{LOCO} which are 21 chromosomes, and X_{test} is from the remained chromosome.

$$\text{V} = \text{cov(y)} = \sigma_g^2 K + \sigma_e^2 I$$



The matrix $X_{\text{GRM}}X_{\text{GRM}}'/M_{\text{GRM}}$ is conventionally called the GRM or empirical kinship matrix *K*, and we write

$$\text{cov}(g) = \sigma_g^2 X_{\text{GRM}} X_{\text{GRM}}'/M_{\text{GRM}} = \sigma_g^2 K \qquad (3)$$

where σ_g^2 is a variance parameter. Environmental effects are assumed to be independently and identically distributed normally such that *e* is also multivariate normal with

$$\text{cov}(e) = \sigma_e^2 I \qquad (4)$$

where *I* denotes the *N* × *N* identity matrix and σ_e^2 is another variance parameter.

The representation of Chi-square test is:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

$O_i: Observed\ value$

$E_i: Expected\ value$

We can get P-value, β, SE out of this chisq test.

Bolt-lmm-inf

$$\chi^2_{\text{BOLT-LMM-inf}} = \frac{(x'_{\text{test}} V^{-1}_{\text{LOCO}} y)^2}{c_{\text{inf}}} \qquad (8)$$

The difference between Bolt-lmm-inf and Bolt-lmm is that inf takes fewer SNPs to calculate the denominator.

$$c_{\text{inf}} = \frac{\text{mean}\ (x'_{\text{test}} V^{-1}_{\text{LOCO}} y)^2}{\text{mean}\ \chi^2_{\text{LMM-LOCO}}} \qquad (9)$$

In practice, we take 30 SNPs to estimate the c_{inf} for computational efficiency.

My question:

In the numerator factor, we still need to take LOCO CV for 22 times, and it's still a matrix multiplication calculation, how can it gets speed up?

It's like in time complicity calculation: if in Bolt-lmm we calculate the numerator in O(N^2) time, and we calculate the denominator in O(N^2) time, so the total time is O(N^2). and in Bolt-lmm-inf,

inf

the calculation time of c_{inf} becomes a constant time O(1), the total time is still O(N^2)