MATH6450E Project 2: Heritability Estimation based on Linear Mixed Model

FAN Min *
Department of Mathematics
Hong Kong University of Science and Technology
mfanab@connect.ust.hk

Abstract

One application of linear mixed model is to estimate heritability of phenotype based on genotype. In linear mixed model, covariates effects such as age and sex are represented as fix effects part, while heritability is modeled in random effect part. The heritability is estimated by the variance of random effects part over total variance. In this project, we did simulation on GWAS.RData dataset to estimate the heritability of 4 phenotypes.

1 Introduction

GWAS refers to Genome-wide association study. GWAS.RData collects $n = 5123$ individuals genotypes and phenotypes. $\mathbf{G} = [g_{im}] \in \mathbb{R}^{n\times p}$ is the genotype matrix where $p = 319147$ and each column corresponds to a genetic marker. $\mathbf{y}$ is a $n \times 4$ phenotype matrix, where each column indicates one phenotype. Because each time we only care about one phenotype, without loss of generality, the column we are considering is denoted as $\mathbf{y}$.

Linear mixed model is to estimate the heritability of each of the four phenotypes as follows

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{W}\mathbf{u} + \mathbf{e}. \tag{1}$$

$\mathbf{X}\beta$ is about fixed effects. $\mathbb{X} \in \mathbb{R}^{n\times(10+1)}$ includes the principal components scores corresponding to the first ten leading principal components and one column of ones, instead of covariates matrix in general setting. β is the vector of coefficients corresponding to fixed effects.

$\mathbf{W}\mathbf{u}$ is about random effects. $\mathbf{u} \sim \mathcal{N}(0, \sigma_u^2\mathbf{I})$ is the vector of random effect size. $\mathbf{W}$ is the standardized genotype matrix with zero mean and unit variance, that is,

$$w_{im} = \frac{g_{im} - 2p_m}{\sqrt{2p_m(1-p_m)p}}, \tag{2}$$

where p_m is the frequency of the reference allele.

$\mathbf{e} \sim \mathcal{N}(0, \sigma_e^2\mathbf{I})$.

The parameters $\theta = \left\{\beta, \sigma_u^2, \sigma_e^2\right\}$ can be estimated by maximum likelihood estimation. The heritability is calculated by $\hat{h}^2 = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2+\hat{\sigma}_e^2}$.

*https://github.com/ProteusFAN/MATH6450E/tree/master/Heritability

2 Method

We can easily have

$$P(\mathbf{y}|\beta, \sigma_u^2, \sigma_e^2) = \mathcal{N}(\mathbf{X}\beta, \sigma_u^2\mathbf{W}\mathbf{W}^{\mathrm{T}} + \sigma_e^2\mathbf{I}). \tag{3}$$

And the log-likelihood function is

$$\ell(\beta, \sigma_u^2, \sigma_e^2) = -\frac{1}{2}(n\log(2\pi\sigma_u^2) + \log(|\mathbf{K} + \delta\mathbf{I}|) + \frac{1}{\sigma_u^2}(\mathbf{y} - \mathbf{X}\beta)^{\mathrm{T}}(\mathbf{K} + \delta\mathbf{I})^{-1}(\mathbf{y} - \mathbf{X}\beta)), \tag{4}$$

where $\mathbf{K} = \mathbf{W}\mathbf{W}^{\mathrm{T}}$ and $\delta = \frac{\sigma_e^2}{\sigma_u^2}$.

By spectral decomposition, $\mathbf{K} = \mathbf{U}\mathbf{S}\mathbf{U}^{\mathrm{T}}$ and $\mathbf{I} = \mathbf{U}\mathbf{U}^{\mathrm{T}}$. Equation 4 becomes

$$\ell(\beta, \sigma_u^2, \delta) = -\frac{1}{2}\left(n\log(2\pi\sigma_u^2) + \log(|\mathbf{S} + \delta\mathbf{I}|) + \frac{1}{\sigma_u^2}(\mathbf{U}^{\mathrm{T}}\mathbf{y} - \mathbf{U}^{\mathrm{T}}\mathbf{X}\beta)^{\mathrm{T}}(\mathbf{S} + \delta\mathbf{I})^{-1}(\mathbf{U}^{\mathrm{T}}\mathbf{y} - \mathbf{U}^{\mathrm{T}}\mathbf{X}\beta)^{\mathrm{T}}\right) \tag{5}$$

$$= -\frac{1}{2}\left(n\log(2\pi\sigma_u^2) + \sum_{i=1}^{n}\log([\mathbf{S}]_{ii} + \delta) + \frac{1}{\sigma_u^2}\sum_{i=1}^{n}\frac{([\mathbf{U}^{\mathrm{T}}\mathbf{y}]_i - [\mathbf{U}^{\mathrm{T}}\mathbf{X}]_{i:}\beta)^2}{[\mathbf{S}]_{ii} + \delta}\right). \tag{6}$$

Taking the derivative of equation 5 w.r.t. β and setting it to zero, we can have

$$\hat{\beta} = [(\mathbf{U}^{\mathrm{T}}\mathbf{X})^{\mathrm{T}}(\mathbf{S} + \delta\mathbf{I})^{-1}(\mathbf{U}^{\mathrm{T}}\mathbf{X})]^{-1}(\mathbf{U}^{\mathrm{T}}\mathbf{X})^{\mathrm{T}}(\mathbf{S} + \delta\mathbf{I})^{-1}(\mathbf{U}^{\mathrm{T}}\mathbf{y}) \tag{7}$$

$$= \left[\sum_{i=1}^{n}\frac{1}{[\mathbf{S}]_{ii} + \delta}[\mathbf{U}^{\mathrm{T}}\mathbf{X}]_{i:}^{\mathrm{T}}[\mathbf{U}^{\mathrm{T}}\mathbf{X}]_{i:}\right]^{-1}\left[\sum_{i=1}^{n}\frac{1}{[\mathbf{S}]_{ii} + \delta}[\mathbf{U}^{\mathrm{T}}\mathbf{X}]_{i:}^{\mathrm{T}}[\mathbf{U}^{\mathrm{T}}\mathbf{y}]_i\right] \tag{8}$$

Substituting $\hat{\beta}$ in equation 5 and taking derivative w.r.t. σ_u^2, we can have

$$\hat{\sigma_u^2} = \frac{1}{n}\sum_{i=1}^{n}\frac{([\mathbf{U}^{\mathrm{T}}\mathbf{y}]_i - [\mathbf{U}^{\mathrm{T}}\mathbf{X}]_{i:}\hat{\beta})^2}{[\mathbf{S}]_{ii} + \delta} \tag{9}$$

Plugging $\hat{\sigma}_u^2$ and $\hat{\beta}$ into equation 5, we have

$$\ell(\delta) = -\frac{1}{2}\left(n\log(2\pi) + \sum_{i=1}^{n}\log([\mathbf{S}]_{ii} + \delta) + n + n\log\frac{1}{n}\sum_{i=1}^{n}\frac{([\mathbf{U}^{\mathrm{T}}\mathbf{y}]_i - [\mathbf{U}^{\mathrm{T}}\mathbf{X}]_{i:}\hat{\beta})^2}{[\mathbf{S}]_{ii} + \delta}\right), \tag{10}$$

which is a function only related to δ. Brent's method can be implemented to find the maximum of equation 10.

3 Experiment and Result

We first select 100 samples with 1000 genetic markers and implement MLE algorithms. But the result is unsatisfactory. δ tends to be $+\infty$, which means $\frac{\sigma_e^2}{\sigma_u^2}$ goes to $+\infty$ and heritability is 0.