High-dimensional Bayesian Model Selection

Nanwei Wang*
Lunenfeld-Tanenbaum Research Institute
and
Laurent Briollais
Lunenfeld-Tanenbaum Research Institute
Helene Massam
Department of Mathematics and Statistics, York Univeristy

May 28, 2019

Abstract

This is a review of linear mixed models in GWAS analysis. In this review, we will explain some important mathematical concepts regards to linear mixed models. We will mainly focus on the maximum likelihood estimation, restricted maximum likelihood estimation of linear mixed models and the various hypothesis tests in GWAS.

1 Introduction

Given a dataset consist of N samples and p snps, denote y as the continous phenotype variable; X as the $n \times p$ snp matrix and W as other nuisance variables including the intercept. The nuisance variables can be patients' personal information and principle components. To test if one snp X_i is correlated with the phenotype, we can build the following linear mixed model:

$$y = \beta_0 W + \beta_i X_i + z + \epsilon, \tag{1.1}$$

*The authors gratefully acknowledge *please remember to list all relevant funding sources in the unblinded version*

Where $z \sim \mathcal{N}(0, \sigma_g^2 K)$, $K = \frac{1}{p} X X^t$ denotes the genetic relationship matrix, is the genetic variance component and $\epsilon \sim \mathcal{N}(0, \sigma_e^2)$ is the random error. There are two parts of the linear mixed model: $\beta_0 W + \beta_i X_i$ is the fixed effects and $z + \epsilon$ is the random effects. Compared to the simple linear model, the genetic variance component z accounts for sample relatedness and population stratification.

2 Maximum likelihood estimation

In order to simplify the notation, we just write $X\beta$ as the fixed effects. In the linear mixed model, we can get that the phenotype y follows the multivariate normal distribution $\mathcal{N}(X\beta, \Sigma)$, where $\Sigma = \sigma_g^2 K + \sigma_e^2 I_N$. The log-likelihood of the parameters $\theta = (\beta, \sigma_g, \sigma_e)$ is

$$l(\theta|y, X) = -\frac{1}{2}(n \log(2\pi) + \log|\Sigma| + (y - X\beta)^t \Sigma^{-1}(y - X\beta)) \tag{2.1}$$

As the covariance matrix Σ is not diagonal, maximizing the log-likelihood function 2.1 to compute the MLE is very difficult. To simplify the log-likelihood function, we can do a transformation of the original data and get independent samples.

First, get the eigen-decomposition of matrix $K = USU^t$, where S is the diagonal matrix of eigenvalues, the columns of matrix $U$is the corresponding eigenvectors. Second, transform the data:

$$y^* = U^t y; \;\; X^* = U^t X,$$

so we get

$$\begin{aligned} E(y^*) &= X^*\beta; \\ Var(y^*) &= \sigma_g^2 UKU^t + \sigma_e^2 UIU^t = \sigma_g^2 S + \sigma_e^2 I. \end{aligned}$$

Finally, we simplify the log-likelihood fucntion 2.1:

$$l(\theta|y^*, X^*) = -\frac{1}{2}(n \log(2\pi) + \log|\sigma_g^2 S + \sigma_e^2 I| + (y^* - X^*\beta)^t (\sigma_g^2 S + \sigma_e^2 I)^{-1}(y^* - X^*\beta). \tag{2.2}$$

Set $\delta = \frac{\sigma_e^2}{\sigma_g^2}$, the log-likelihood function can be further simplified as follows

$$l(\theta|y^*, X^*) = -\frac{1}{2}(n \log(2\pi) + n \log(\sigma_g^2) + \sum_{i=1}^{n} \log(S_{ii} + \delta) + \frac{1}{\sigma_g^2}(y^* - X^*\beta)^t (S + \delta I)^{-1}(y^* - X^*\beta). \tag{2.3}$$

The score function of parameter β, σ_g^2 in log-likelihood function 2.3 is

$$\begin{aligned} S(\beta) &= \frac{dl}{d\beta} = -\frac{1}{\sigma_g^2}(X^*)^t (S + \delta I)^{-1}(y^* - X^*\beta) \\ S(\sigma_g^2) &= \frac{dl}{d\sigma_g^2} = -\frac{1}{2}(\frac{n}{\sigma_g^2} - \frac{1}{\sigma_g^4}(y^* - X^*\beta)^t (S + \delta I)^{-1}(y^* - X^*\beta)) \end{aligned}$$

Set them to zero, we can get

$$\begin{aligned} \beta &= ((X^*)^t(S+\delta I)^{-1}X^*)^{-1}(X^*)^t(S+\delta I)^{-1}y^* \\ \sigma_g^2 &= \frac{1}{n}(y^*-X^*\beta)^t(S+\delta I)^{-1}(y^*-X^*\beta)) \end{aligned} \tag{2.4}$$

Plug the equations 2.4 back the to log-likelihood 2.3, we can write the log-likelihood function as a one-dimensional function of parameter δ. It will be easy to get the MLE of δ first, then the MLE of β and δ_g.

3 Restricted maximum likelihood estimation

As the MLE of the variance parameter σ_g^2 is biased, the restricted maximum likelihood estimation is used more often in the literature. We will talk more about RMLE in this review.

First,we can study a simple case: linear mixed model without fixed effect part:

$$y \sim \mathcal{N}(0,\Sigma), \quad \text{where } \Sigma = \sigma_g^2(K+\delta I).$$

The log-likelihood function is

$$l(\theta|y,X) = -\frac{1}{2}(N\log(2\pi) + \log|\Sigma| + y^t\Sigma^{-1}y)$$

Plug the $K = USU^t$ into the the log-likelihood function ,or equvaliently to say, perform the transformation of data: $y^* = U^ty$; $X^* = U^tX$, we can simplify the log-likelihood function:

$$l(\theta|y^*,X^*) = -\frac{1}{2}(N\log(2\pi) + n\log(\sigma_g^2) + \log|S+\delta I| + \frac{1}{\sigma_g^2}(y^*)^t(S+\delta I)^{-1}y^*)$$

The score function is

$$\begin{aligned} S(\sigma_g^2) &= \frac{n}{\sigma_g^2} - \frac{1}{(\sigma_g^2)^2}\sum_{i=1}^n \frac{(y_i^*)^2}{\lambda_i+\delta} \\ S(\delta) &= \sum_{i=1}^n \frac{1}{\lambda_i+\delta} - \frac{1}{\sigma_g^2}\sum_{i=1}^n \frac{(y_i^*)^2}{(\lambda_i+\delta)^2}, \end{aligned}$$

and the fisher information matrix is

$$F = \begin{bmatrix} \frac{dS(\sigma_g^2)}{d\sigma_g^2} & \frac{dS(\sigma_g^2)}{d\delta} \\ \frac{dS(\delta)}{d\sigma_g^2} & \frac{dS(\delta)}{d\delta} \end{bmatrix} = \begin{bmatrix} -\frac{n}{(\sigma_g^2)^2} + \frac{2}{(\sigma_g^2)^3}\sum_{i=1}^n \frac{(y_i^*)^2}{\lambda_i+\delta} & \frac{1}{(\sigma_g^2)^2}\sum_{i=1}^n \frac{(y_i^*)^2}{\lambda_i+\delta} \\ \frac{1}{(\sigma_g^2)^2}\sum_{i=1}^n \frac{(y_i^*)^2}{\lambda_i+\delta} & -\sum_{i=1}^n \frac{1}{(\lambda_i+\delta)^2} + \frac{2}{\sigma_g^2}\sum_{i=1}^n \frac{(y_i^*)^2}{(\lambda_i+\delta)^3} \end{bmatrix}.$$

We can easily extend the above computation to linear mixed model with fixed effects. We need to find a projection matrix A to transform the data such that $AX = 0$. The residual matrix $R = I - X(X^tX)^{-1}X^t$. We give the following lemma about the residual matrix.

Lemma 3.1.

we show mathematically how one can get the restricted log-likelihood function from the regular log-likelihood function. Denote $P = \Sigma^{-1}(I - X(X^t\Sigma^{-1}X)^{-1}X^t\Sigma^{-1})$ The restricted log-likelihood is

$$l_r(\theta|X, y) = -\tfrac{1}{2}\big((N-k)\log(2\pi) - \log|X^tX| + \log|\Sigma| + \log|X^t\Sigma^{-1}X| + y^TPy\big) \quad (3.1)$$