SPARSE LINEAR MIXED MODEL

Xiao Jiashun
Department of Mathematics
Hong Kong University of Science and Technology
`jxiaoae@connect.ust.hk`

April 25, 2019

ABSTRACT

linear mixed model (LMM) is a powerful tool in genetics application, Including, control the population stratification, individual relatedness and polygenic modeling in GWAS. Here, the polygennic modeling means prediction of phenotype based on genotype and try to capture the "chip heritability" as much as possible. However, the hypothesis of LMM is based on that every genetic variants have effects on the phenotype and the effect size are normally distributed, which is obvious inappropriate in true high dimensional, million of genetics variants, genetic architecture. Motivated by this, we consider a sparse linear mixed model, which assumes that only a relatively small proportion of all genetic variants have effects on the phenotype. To make our model feasible for large-scale data set, we adopt variational approximation to estimate posterior distribution of latent variables and parameters. We will illustrate how to derive the model and test it on both simulated and real GWAS data set. The python source code is available for download at Github (https://github.com/JiaShun-Xiao/SparseLMM)

***K*eywords** Sparse LMM · Variational EM

1 Variational EM: background and notation

Suppose we have dataset $\{X, Z, y\}$ where $X \in \mathbb{R}^{nxp}$ is the design matrix whose columns are normalized with mean 0 and variance 1/p, $y \in \mathbb{R}^n$ is linear combination of the candidate predictors X, covariate matrix $Z \in \mathbb{R}^{nxc}$ and residual noise e

$$Y = Zw + X\beta + e \tag{1}$$

where $\beta \in \mathbb{R}^p$ is the random effect, and $e \sim N(0, \sigma_e^2)$. Let γ_k be the variable indicating whether β_k is zero or not. We assume the following spike-slab prior for β:

$$\begin{cases} \beta_k \sim N(0, \sigma_\beta^2 I_p) & if \gamma_k = 1 \\ \beta_k = 0 & if \gamma_k = 0 \end{cases} \tag{2}$$

where $Pr(\gamma_k = 1) = \pi$ and $Pr(\gamma_k = 0) = 1 - \pi$. There are two components in this model. The first one are the unknown parameters ($\theta = \{\pi, \sigma_\beta^2, \sigma_e^2\}$), the second one are latent variable (β, γ). Here we estimate the posterior of β, γ given θ by minimizing the Kullback-Leibler divergence between an approximating distribution $q(\beta; \gamma)$, and then iteratively estimate $p(\theta|X, y)$ by maximizing the variational lower bound with the variational solution from the first component.

1.1 Posterior of latent variable

The basic idea behind the variational approximation is to find a distribution $q(\beta; \gamma)$ that provide a good approximation to the posterior $p(\beta, \gamma|X, y, \theta)$ by mimimizing the Kullback-Leibler divergence

$$D(q||p) = \int q(\beta, \gamma) log\{\frac{q(\beta; \gamma)}{p(\beta, \gamma)}\} d\beta d\gamma \tag{3}$$

the $q(\beta, \gamma)$ is restricted to be of the form

$$q(\beta, \gamma; \phi) = \prod_{k=1}^{p} q(\beta_k, \gamma_k; \phi_k) \tag{4}$$

where $\phi = (\phi_1, ..., \phi_p)$ are free parameters, each factors have the form

$$q(\beta_k, \gamma_k; \phi_k) = \begin{cases} \alpha_k N(\beta_k | \mu_k, s_k^2) & if \gamma_k = 1 \\ (1 - \alpha_k)\delta_0(\beta_k) & if \gamma_k = 0 \end{cases} \tag{5}$$

where δ_0 is the "spike" at zero, and $\phi_k = (\alpha_k, \mu_k,{}_k^2)$. With probability α_k, the additive effect β_k is normal with mean μ_k and variance ${}_k^2$ (the “slab”), and with probability $1 - \alpha_k$, the variable has no effect on Y.
The coordinate descent updates for this optimization problem can be obtained by taking partial derivatives of the Kullback-Leibler divergence, setting the partial derivatives to zero, and solving for the parameters α_k,μ_k and ${}_k^2$. This yields coordinate updates

$$\begin{aligned} s_k^2 &= \frac{\sigma_e^2}{(X^T X)_{kk} + 1/\sigma_\beta^2} \\ \mu_k &= \frac{s_k^2}{\sigma_e^2}((X^T y)_k - \sum_{j \neq k} (X^T X)_{jk} \alpha_j \mu_j) \\ \frac{\alpha_k}{1 - \alpha_k} &= \frac{\pi}{1 - \pi} \times \frac{s_k}{\sigma_\beta \sigma_e} \times e^{\frac{\mu_k^2}{2 s_k^2}} \end{aligned}$$

1.2 Posterior of hyperparameters

In M step, by plug the posterior expectation of latent variable from variational infernce step $\{\beta, \gamma\}$ into the lower bound of Kullback-Leibler divergence, we could derive the M-step updates for σ_β and σ_e in the standard way by solving for roots σ_β and σ_e of the gradient, yielding

$$\begin{aligned} \sigma_\beta^2 &= \frac{\sum_{k=1}^{p} \alpha_k (s_k^2 + \mu_k^2)}{\sigma_e^2 \sum_{k=1}^{p} \alpha_k} \\ \sigma_e^2 &= \frac{\|y - Xr\|_2^2 + \sum_{k=1}^{p} (X^T X)_{kk} Var(\beta_k) + \sum_{k=1}^{p} \alpha_k (s_k^2 + \mu_k^2)/\sigma_\beta^2}{n + \sum_{k=1}^{p} \alpha_k} \end{aligned}$$

where $Var(\beta_k) = \alpha_k (s_k^2 + \mu_k^2) - (\alpha_k \mu_k)^2$

2 Results

2.1 Simulations on Quantitative Phenotype Prediction

In this section, we compare the performance of sparse LMM with lasso in a small simulated dataset with 20,000 candidate predictors (SNP), 3 covariate with big effects on phenotype and 1000 individuals. In addition, we generated the simulated data in 9 different scenario. The causal SNP numbers are categorized into three group: 1,10,100, and the "chip heritability" are categorized into three group: 0.2,0.5,0.8. As we can see from Fig1, SpaseLMM always have high prediction R2 score than lasso in five-fold cross- validation, except when the causal SNP number is 100 and chip heritability is larger than 0.5.

2.2 Predicting Quantitative Phenotype in real data

The real data we performed five-fold cross-validation is provided by Prof. Yang Can from Hong Kong University of Science and Technology. There are 4 different quantitative phenotype, and each of them are linear response of 319,147 SNP and 5123 individual. In addition, ten leading principal component and a column of ones were included as covariate. SparseLMM always have better performance than Lasso, especially in phenotype2 and phenotype3 (Fig.2). However, for phenotype4, both methods show no prediction power indicate that the phenotype may have a small chip heritability.

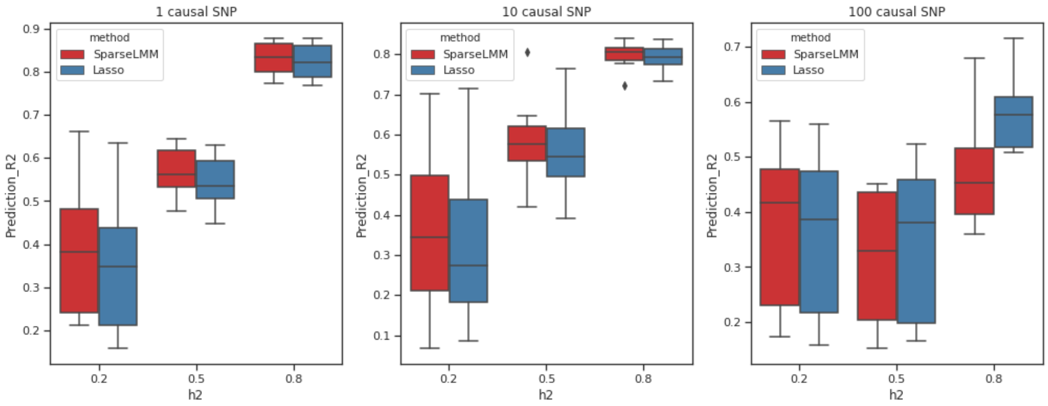

Figure 1: Comparison the accuracy between Sparse LMM and Lasso under different simulation scenarios

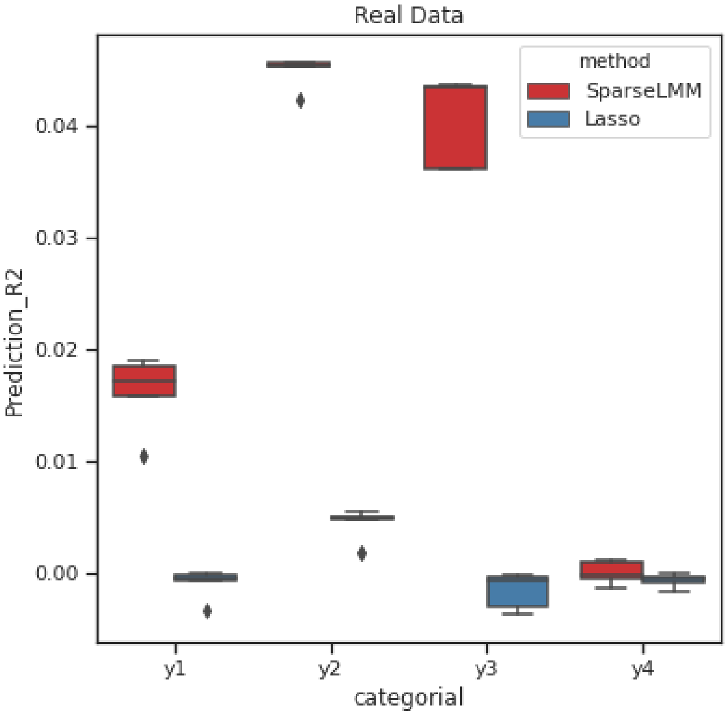

Figure 2: Comparison the accuracy between Sparse LMM and Lasso in real data

References

[1] Carbonetto, P., Zhou, X., Stephens, M. varbvs: Fast Variable Selection for Large-scale Regression. In *arXiv preprint arXiv:1709.06597*, 2017.

[2] Carbonetto, P., Stephens, M. Scalable variational inference for Bayesian variable selection in regression, and its accuracy in genetic association studies. In *Bayesian analysis*, 7(1) 73-108. 2012.