A Sparse Group Lasso multi-marker mixed model for Genome-Wide association studies

Yingjie GUO, Chenxi WU*, Ao LI, Xiaoyan LIU, Alon KEINAN, Maozu GUO

April 24, 2017

Abstract

1 Introduction

2 Materials and Methods

2.1 Genotypes and phenotypes

2.2

2.3 Multivariate Linear mixed model

Our approach builds on linear mixed model, explaining the phenotype variability by a sum of individual genetic effects and random confounding effects. Suppose that m individuals of a phenotype are collected, a linear mixed model in association mapping is typically expressed as

$$\boldsymbol{y} = \underbrace{\sum_{j=1}^{n} \beta_j \boldsymbol{x}_j}_{\text{Fixed effect}} + \underbrace{\underbrace{\boldsymbol{Zu}}_{\text{confounding}} + \underbrace{\boldsymbol{\Psi}}_{\text{noise}}}_{\text{Random effect}} \tag{1}$$

Where $\boldsymbol{y} = (y_1, \ldots, y_m)$ is an $m \times 1$ vector of observed phenotypes, and $\boldsymbol{X} = (\boldsymbol{x_1}, \ldots, \boldsymbol{x_q})$ is an $m \times q$ matrix of fixed effects including SNPs and other confounding variables. $\boldsymbol{\beta}$ is a $q \times 1$ vector representing coefficients of the fixed effects. $\boldsymbol{Z}$ is usually a designed matrix which gives group information of the samples. Here we treat each sample belongs to an unique group that $\boldsymbol{Z}$ turns out to be an identity matrix. $\boldsymbol{u}$ is the random effect of the

*Max-Planck Institute of Mathematics, Vivatsgasse 7, 53111 Bonn, Germany, wuchenxi2013@gmail.com

mixed model with $Var(\boldsymbol{u}) = \boldsymbol{\delta_g^2 K}$. To account for confounding by population structure, family structure, and cryptic relatedness in GWAS, $\boldsymbol{K}$ can be reliably estimated from genetic markers, e.g., using the realized relationship matrix(RRM) which captures the overall genetic similarity between all pairs of samples.

The resulting mixed model is typically applied in the context of single candidate SNP, that is, restricting the sum in Equation (1) to a particular SNP while ignoring all others. This independent analysis can be compromised by complex genetic architectures with some genetic factors masking others.

As an alternative, we proposed an efficient approach to carry out joint inference in the model implied by Equation (1). Our approach assesses all the SNPs at the same time while accounting for their interdependencies and without making any assumptions on their ordering. To allow for applications to genome-wide SNP data, we place a XX prior over the fixed effects β_j, assigning zero-effect size to majority of SNPs via bi-level selection method that select non-zero effect SNPs both at the group level and within group level, which as done in the classical sparse group lasso.

We call this approach SGL-LMM(Sparse Group Lasso-Linear Mixed Model), as it combines the advantages of established LMM with sparse group Lasso regression. The resulting model allows for dissecting the explained phenotype variance into individual SNP effects and effects caused by population structure.

2.3.1 Sparse Group Lasso - Linear Mixed Model

Let $\boldsymbol{X} \in \mathbb{R}^{m \times q}$ denotes the matrix of fixed effects for m individuals. This matrix includes the column of 1s corresponding to the offset, the covariances, and all the SNPs to be considered. We model the phenotype for m individuals, $\boldsymbol{y} \in \mathbb{R}^{m \times 1}$ as the sum of fixed effects β_j of SNPs $\boldsymbol{x_j}$ and random effect due to confounding influences $\boldsymbol{u}$ [see Equation (1)]. The fixed effect terms are summed over genome-wide SNPs, where the great majority of SNPs have zero-effect size, that is, $\beta_j = 0$, which is achieved by a XX prior on all weights. The random effect $\boldsymbol{u}$ is not observed directly. Instead, we assume that the distribution of $\boldsymbol{u}$ is Gaussian with covariance $\boldsymbol{K}$, $\boldsymbol{u} \sim \mathcal{N}(\boldsymbol{0}, \sigma_g^2 \boldsymbol{K})$.

Assuming Gaussian noise, $\boldsymbol{\psi} \sim \mathcal{N}(\boldsymbol{0}, \sigma_e^2 \boldsymbol{I})$, and marginalizing over the random vector $\boldsymbol{u}$, we can write down the conditional posterior distribution over the weight vector $\boldsymbol{\beta}$:

$$p(\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{K}, \sigma_g^2, \sigma_e^2, \lambda) \propto \Bigg(\underbrace{\mathcal{N}(\boldsymbol{y}| \sum_{j=1}^{q} \beta_j \boldsymbol{x_j}, \, \sigma_g^2 \boldsymbol{K} + \sigma_e^2 \boldsymbol{I})}_{marginallikelihood} \underbrace{\pi(\boldsymbol{\beta})}_{prior} \Bigg) \tag{2}$$

With the prior of the form

$$\pi(\boldsymbol{\beta}) \propto exp(-\lambda \sum_{G \in \mathcal{G}} (\alpha_G \parallel \boldsymbol{\beta}_G \parallel_2 + \parallel \boldsymbol{\beta}_G \parallel_1)) \tag{3}$$

Here, suppose we are given M (possibly overlapping) groups $\mathcal{G} = \{G_1, G_2, \ldots, G_M\}$, so that $G_i \subset \{1, 2, \ldots, p\} \forall i$, of maximum size B. These groups contain sets of "similar" features, such as SNPs within a region of a gene. We assume that all but $k \ll M$ groups are identically zero. Among the active groups, we further assume that at most a fraction $\alpha \in (0, 1)$ of the coefficients per group are non zero. $\alpha_G > 0$ are constants that balance the tradeoff between the group norms and the ℓ_1 norm. λ denotes the sparsity hyperparameter of our prior, σ_g^2 is the magnitude of the random effect component, and σ_e^2 denotes the magnitude of the residual variance.

2.3.2 Parameter inference

Learning the hyperparameters $\theta = \{\sigma_g^2, \sigma_e^2\}$ and the weights $\boldsymbol{\beta}$ jointly is a hard non-convex optimization problems. Here, we propose a combinaton of fitting some of these parameters on the null model with the individual SNP effects excluded and reduction to a sparse overlapping group lasso problem.

We first optimize σ_g^2,σ_e^2 by maximum likelihood under the null model, ignoring the effect of individual SNPs. The analogous procedure is widely used in single-SNP mixed models and has been shown to yield near-identical results to an exact approach.

Evaluation of the log likelihood The log likelihood is parameterized by a weight vector $\boldsymbol{\beta}$ and the variances of the random components, σ_g^2 and σ_e^2.

$$LL(\sigma_g^2, \sigma_e^2, \boldsymbol{\beta}) = log\mathcal{N}(\boldsymbol{y}|\boldsymbol{X}\boldsymbol{\beta},\ \sigma_g^2\boldsymbol{K} + \sigma_e^2\boldsymbol{I}) \tag{4}$$

To speed up the computation needed, we introducing $\delta \equiv \frac{\sigma_e^2}{\sigma_g^2}$, the covariance matrix becomes $\sigma_g^2(\boldsymbol{K} + \delta\boldsymbol{I})$, and the likelihood becomes a function of $\boldsymbol{\beta}$,δ and σ_g^2:

$$LL(\delta, \sigma_g^2, \boldsymbol{\beta}) = log\mathcal{N}(\boldsymbol{y}|\boldsymbol{X}\boldsymbol{\beta}, \sigma_g^2(\boldsymbol{K} + \delta\boldsymbol{I})) \tag{5}$$

Using the formula for the multivariate Normal distribution, we obtain

$$LL(\delta, \sigma_g^2, \boldsymbol{\beta}) = -\frac{1}{2}\Big(mlog(2\pi\sigma_g^2) + log(|\boldsymbol{K} + \delta\boldsymbol{I}|) + \frac{1}{\sigma_g^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{K} + \delta\boldsymbol{I})^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\Big) \tag{6}$$

Since $\boldsymbol{K}$ is symmetric matrix, letting $\boldsymbol{U}\boldsymbol{S}\boldsymbol{U}^T = \boldsymbol{K}$ be the spectral decomposition of $\boldsymbol{K}$, and noting that $\boldsymbol{I} = \boldsymbol{U}\boldsymbol{U}^T$, Equation (6) becomes

$$LL(\delta, \sigma_g^2, \boldsymbol{\beta}) = -\frac{1}{2}\Big(mlog(2\pi\sigma_g^2) + log((|USU^T + \delta\boldsymbol{U}\boldsymbol{U}^T|)) + \frac{1}{\sigma_g^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(USU^T + \delta\boldsymbol{U}\boldsymbol{U}^T)^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\Big)$$

Next, we factor out $\boldsymbol{U}$ and $\boldsymbol{U}^T$ from the covariance of the Normal, so that it becomes the diagonal matrix $(\boldsymbol{S} + \delta I)$, obtaining

$$-\frac{1}{2}\Big(mlog(2\pi\sigma_g^2)+log(|\boldsymbol{U}(\boldsymbol{S}+\delta\boldsymbol{I})\boldsymbol{U}^T|)+\frac{1}{\sigma_g^2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{U}(\boldsymbol{S}+\delta\boldsymbol{I})\boldsymbol{U}^T|)^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})\Big) \quad (7)$$

The determinant of $|\boldsymbol{U}(\boldsymbol{S}+\delta\boldsymbol{I})\boldsymbol{U}^T|$ can be written as $|(\boldsymbol{S}+\delta\boldsymbol{I})|$ using the properties that $|\boldsymbol{AB}=|\boldsymbol{A}||\boldsymbol{B}||$, and that $|\boldsymbol{U}|=|\boldsymbol{U}^T|=1$. The inverse of this part can be rewritten as $\boldsymbol{U}(\boldsymbol{S}+\delta\boldsymbol{I})^{-1}\boldsymbol{U}^T$ using the properties that $(\boldsymbol{AB}^{-1})=\boldsymbol{B}^{-1}\boldsymbol{A}^{-1}$, that $\boldsymbol{U}^{-1}=\boldsymbol{U}^T$, and that $\boldsymbol{U}^{-T}=\boldsymbol{U}$. Thus, after additionally moving out $\boldsymbol{U}$ from the covariance term so that it now acts as a rotation matrix on the inputs $\boldsymbol{X}$ and targets $\boldsymbol{y}$, we obtain

$$-\frac{1}{2}\Big(mlog(2\pi\sigma_g^2)+log(|\boldsymbol{U}||(\boldsymbol{S}+\delta\boldsymbol{I})||\boldsymbol{U}^T|)+\frac{1}{\sigma_g^2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^T\boldsymbol{U}(\boldsymbol{S}+\delta\boldsymbol{I})^{-1}\boldsymbol{U}^T(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})\Big)$$

$$=-\frac{1}{2}\Big(mlog(2\pi\sigma_g^2)+log(|(\boldsymbol{S}+\delta\boldsymbol{I})|)+\frac{1}{\sigma_g^2}\big((\boldsymbol{U}^T\boldsymbol{y})-(\boldsymbol{U}^T\boldsymbol{X})\boldsymbol{\beta}\big)^T(\boldsymbol{S}+\delta\boldsymbol{I})^{-1}\big((\boldsymbol{U}^T\boldsymbol{y})-(\boldsymbol{U}^T\boldsymbol{X})\boldsymbol{\beta}\big)\Big) \quad (8)$$

Finding the maximum likelihood genetic variance with null model We start by eliminating $\boldsymbol{\beta}$ from Equation (8), and set derivative with respect to σ_g^2 to zero. As the covariance matrix of the Normal distribution is now a diagonal matrix $(\boldsymbol{S}+\delta\boldsymbol{I})$, we could obtain

$$0=-\frac{1}{2}\left(\frac{n}{\hat{\sigma}_g^2}-\frac{1}{\hat{\sigma}_g^4}\sum_{i=1}^{m}\frac{([\boldsymbol{U}^T\boldsymbol{y}]_i)^2}{[\boldsymbol{S}]_{ii}+\delta}\right)$$

Multiplllying both sides by $2\hat{\sigma}_g^4$ and solving for $\hat{\sigma}_g^2$, we get

$$\hat{\sigma}_g^2=\frac{1}{n}\sum_{i=1}^{m}\frac{([\boldsymbol{U}^T\boldsymbol{y}]_i)^2}{[\boldsymbol{S}]_{ii}+\delta}$$

This equation can be evaluated in $\mathcal{O}(m)$.

Optimization of δ Plugging in $\hat{\sigma}_g^2$ into Equation(8) with null model, the log likelihood becomes a function only of δ, $LL(\delta,\hat{\sigma}_g^2(\delta))=LL(\delta)$:

$$LL(\delta)=-\frac{1}{2}\left(nlog(2\pi)+\sum_{i=1}^{m}log([\boldsymbol{S}]_{ii}+\delta)+n+nlog\frac{1}{n}\left(\sum_{i=1}^{m}\frac{([\boldsymbol{U}^T\boldsymbol{y}]_i)^2}{[\boldsymbol{S}]_{ii}+\delta}\right)\right)$$

We optimize this function of δ using a one-dimensional numerical optimizer to fine the maximum likelihood value of δ, from which the maximum likelihood values of all the parameters can be directly computed.

Reduction to sparse overlapping group lasso problem Having fixed δ, we use the spectral decomposition of $\boldsymbol{K}$ again to rotate our data such that the covariance matrix becomes isotropic

$$p(\boldsymbol{\beta}|\tilde{\boldsymbol{y}}, \tilde{\boldsymbol{X}}, \sigma_g^2) \propto \mathcal{N}(\tilde{\boldsymbol{y}}|\sum_{j=1}^{q} \beta_j \tilde{\boldsymbol{x_j}}, \sigma_g^2 \boldsymbol{I}) \pi(\boldsymbol{\beta}) \tag{9}$$

Here, $\tilde{\boldsymbol{X}}$ denotes the rotated and resealed genotypes, and $\tilde{\boldsymbol{y}}$ denotes the respective phenotypes

$$\tilde{\boldsymbol{X}} = (\boldsymbol{S} + \delta \boldsymbol{I})^{-\frac{1}{2}} \boldsymbol{U}^T \boldsymbol{X}$$
$$\tilde{\boldsymbol{y}} = (\boldsymbol{S} + \delta \boldsymbol{I})^{-\frac{1}{2}} \boldsymbol{U}^T \boldsymbol{y}$$

Sparse group lasso problem Using the transformation above, we can obtain the Maximum-A-Posteriori parameter estimation of the most probable weights in Equation (9)

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{MAP} = \underset{\boldsymbol{\beta}}{\text{argmax}}\ p(\boldsymbol{\beta}|\tilde{\boldsymbol{y}}, \tilde{\boldsymbol{X}}, \sigma_g^2) &= \underset{\boldsymbol{\beta}}{\text{argmax}} p(\tilde{\boldsymbol{y}}|\boldsymbol{\beta}, \tilde{\boldsymbol{X}}, \sigma_g^2) \pi(\boldsymbol{\beta}) \\ &= \underset{\boldsymbol{\beta}}{\text{argmax}}\ log\ p(\tilde{\boldsymbol{y}}|\boldsymbol{\beta}, \tilde{\boldsymbol{X}}, \sigma_g^2) \pi(\boldsymbol{\beta}) \\ &= \underset{\boldsymbol{\beta}}{\text{argmax}}\ \{log\ \pi(\boldsymbol{\beta}) + \ log \prod_{j=1}^{m} p(\tilde{\boldsymbol{y_j}}|\boldsymbol{\beta}, \tilde{\boldsymbol{x}}^j, \sigma_g^2)\} \end{aligned} \tag{10}$$

The log likelihood is as follow:

$$\begin{aligned} log \prod_{j=1}^{m} p(\tilde{\boldsymbol{y_j}}|\boldsymbol{\beta}, \tilde{\boldsymbol{x}}^j, \sigma_g^2)\} &= \sum_{j=1}^{m} log\ p(\tilde{\boldsymbol{y_j}}|\boldsymbol{\beta}, \tilde{\boldsymbol{x}}^j, \sigma_g^2) \\ &= \sum_{j=1}^{m} log \left(\frac{1}{(2\pi)^{1/2}|\Sigma|^{1/2}} exp\Big(-\frac{1}{2} (\tilde{\boldsymbol{y_j}} - \boldsymbol{\beta}^T \tilde{\boldsymbol{x}}^{\boldsymbol{j}})^T \Sigma^{-1} (\tilde{\boldsymbol{y_j}} - \boldsymbol{\beta}^T \tilde{\boldsymbol{x}}^{\boldsymbol{j}}) \Big) \right) \end{aligned}$$

Where $|\Sigma|^{1/2} = \sigma_g^m$, and $\Sigma^{-1} = \frac{1}{\sigma_g^2} \boldsymbol{I}$

$$log \prod_{j=1}^{m} p(\tilde{\boldsymbol{y_j}}|\boldsymbol{\beta}, \tilde{\boldsymbol{x}}^j, \sigma_g^2)\} =$$