A General Framework for Variable Selection in Linear Mixed Models with Applications to Genetic Studies with Structured Populations

Sahir R Bhatnagar[1,2], Karim Oualkacha[3], Yi Yang[4], Marie Forest[2], and Celia MT Greenwood[1,2,5]

[1]Department of Epidemiology, Biostatistics and Occupational Health, McGill University

[2]Lady Davis Institute, Jewish General Hospital, Montréal, QC

[3]Département de Mathématiques, Université de Québec À Montréal

[4]Department of Mathematics and Statistics, McGill University

[5]Departments of Oncology and Human Genetics, McGill University

July 3, 2018

Abstract

Complex traits are thought to be influenced by a combination of environmental factors and rare and common genetic variants. However, detection of such multivariate associations can be compromised by low statistical power and confounding by population structure. Linear mixed effect models (LMM) can account for correlations due to relatedness but are not applicable in high-dimensional (HD) settings where the number

of predictors greatly exceeds the number of samples. False negatives can result from two-stage approaches, where the residuals estimated from a null model adjusted for the subjects' relationship structure are subsequently used as the response in a standard penalized regression model. To overcome these challenges, we develop a general penalized LMM framework that simultaneously selects and estimates variables, accounting for between individual correlations, in one step. Our method can accommodate several sparsity inducing penalties such as the lasso, elastic net and group lasso, and also readily handles prior annotation information in the form of weights. We develop a groupwise-majorization descent algorithm which is highly scalable, computationally efficient and has theoretical guarantees of the convergence. Through simulations, we show that are method has better power over the two-stage approach, particularly for polygenic traits. We apply our method to identify SNPs that predict bone mineral density in the UK Biobank cohort. This approach can also be used to generate genetic risk scores and finding groups of predictors associated with the response, such as variants within a gene or pathway. Our algorithms are available in an R package (https://github.com/sahirbhatnagar/ggmix).

1 Introduction

Genome-wide association studies (GWAS) have become the standard method for analyzing genetic datasets owing to their success in identifying thousands of genetic variants associated with complex diseases (https://www.genome.gov/gwastudies/). Despite these impressive findings, the discovered markers have only been able to explain a small proportion of the phenotypic variance known as the missing heritability problem (1). One plausible explanation is that there are many causal variants that each explain a small amount of variation with small effect sizes (2). Methods such GWAS, which test each variant or single nucleotide polymorphism (SNP) independently, are likely to miss these true associations due to the stringent significance thresholds required to reduce the number of false positives (1). Another

major issue to overcome is that of confounding due to geographic population structure, family and/or cryptic relatedness which can lead to spurious associations (3). For example, there may be subpopulations within a study that differ with respect to their genotype frequencies at a particular locus due to geographical location or their ancestry. This heterogeneity in genotype frequency can cause correlations with other loci and consequently mimic the signal of association even though there is no biological association (4, 5).

To address the first problem, multivariable regression methods have been proposed which simultaneously fit many SNPs in a single model (6, 7). Indeed, the power to detect an association for a given SNP may be increased when other causal SNPs have been accounted for. Conversely, a stronger signal from a causal SNP may weaken false signals when modeled jointly (6).

Confounding by population structure has also received significant attention in the literature (8, 9, 10, 11). There are two main approaches to account for the relatedness between subjects: 1) the principal component (PC) adjustment method and 2) the linear mixed model (LMM). The PC adjustment method includes the top PCs of genome-wide SNP genotypes as additional covariates in the model (12). The LMM uses an estimated covariance matrix from the individuals' genotypes and includes this information in the form of a random effect (3).

While these problems have been addressed in isolation, there has been relatively little progress towards addressing them jointly. Region-based tests of association have been developed where a linear combination of p variants is regressed on the response variable in a mixed model framework (13). In case-control data, a stepwise logistic-regression procedure was used to evaluate the relative importance of variants within a small genetic region (14). These methods however are not applicable in the high-dimensional setting, i.e., when the number of variables p is much larger than the sample size n, as is often the case in genetic studies where millions of variants are measured on thousands of individuals.

In light of this, there has been recent interest in penalized linear mixed models which place a constraint on the magnitude of the effect sizes while controlling for confounding influences such as population structure. For example, the LMM-lasso (15) places a Laplace prior on all main effects while the adaptive mixed lasso (16) uses the L_1 penalty (17) with adaptively chosen weights (18) to allow for differential shrinkage amongst the variables in the model. Another method applied a combination of both the lasso and group lasso penalties in order to select variants within a gene most associated with the response (19). One potential issue with these methods is that they are performed in two steps. First, the variance components are estimated once from a LMM with a single random effect that uses the estimated covariance matrix from the individuals' genotypes to account for the relatedness but assumes no SNP effects. In the second step, these are treated as known quantities by regressing the SNPs on the residuals from the first step, effectively treating the observations as independent. This approach has both computational and practical advantages since existing penalized regression software such as `glmnet` (20) and `gglasso` (21), which assume independent observations, can be applied directly to the residuals. However, recent work has shown that there can be a loss in power if a causal variant is included in the calculation of the covariance matrix as its effect will have been removed in the first step (22). Another issue with the aforementioned methods is that they first require computing the covariance matrix with a computation time of $\mathcal{O}(n^2p)$ followed by a spectral decomposition of this matrix in $\mathcal{O}(n^3)$ time. These methods become prohibitive to use for large cohorts such as the UK Biobank (23) which have collected genetic information on half a million individuals. There is thus a need to develop newer methodologies that reflect the increasing size and genetic heterogeneity of the large cohort studies being assembled today.

In this paper we develop a general penalized LMM framework that simultaneously selects and estimates variables, accounting for between individual correlations, in one step. Our method can accommodate several sparsity inducing penalties such as the lasso, elastic net and group lasso, and also readily handles prior annotation information in the form of weights. We de-

velop a groupwise-majorization descent algorithm which is highly scalable, computationally efficient and has theoretical guarantees of the convergence.

2 Penalized Mixed Models

SNP genotypes can be coded as dummy variables with homozygotes being assigned a 0.0, heterozygotes being a 0.5, and opposite homozygotes being a 1.0 under an additive model or, for models involving dominance or recessive effects, with heterozygotes being assigned a 0.0 or 1.0, respectively. For the analyses we describe below, we assumed an additive model.

2.1 Model Set-up

Let $i = 1, \ldots, N$ be the grouping index, $j = 1, \ldots, n_i$ the observation index within a group and $N_T = \sum_{i=1}^{N} n_i$ the total number of observations. For each group let $\boldsymbol{y}_i = (y_1, \ldots, y_{n_i})$ be the observed vector of responses, $\mathbf{X}_i$ an $n_i \times (p+1)$ design matrix (with the column of 1s for the intercept), $\boldsymbol{b}_i$ a group-specific random effect vector of length n_i and $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \ldots, \varepsilon_{in_i})$ the individual error terms. Furthermore, denote the stacked vectors $\mathbf{Y} = (\mathbf{y}_i, \ldots, \mathbf{y}_N)^T \in \mathbb{R}^{N_T \times 1}$, $\boldsymbol{b} = (\boldsymbol{b}_i, \ldots, \boldsymbol{b}_N)^T \in \mathbb{R}^{N_T \times 1}$, $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}_i, \ldots, \boldsymbol{\varepsilon}_N)^T \in \mathbb{R}^{N_T \times 1}$, and the stacked matrix $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_N)^T \in \mathbb{R}^{N_T \times (p+1)}$. Furthermore, let $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T \in \mathbb{R}^{(p+1)\times 1}$ a vector of fixed effects regression coefficients corresponding to $\mathbf{X}$. Following (24), we consider the following linear mixed model with a single random effect:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{b} + \boldsymbol{\varepsilon} \tag{1}$$

where the random effect $\boldsymbol{b}$ and the error variance $\boldsymbol{\varepsilon}$ are assigned the distributions

$$\boldsymbol{b} \sim \mathcal{N}(0, \eta\sigma^2\boldsymbol{\Phi}) \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}(0, (1-\eta)\sigma^2\mathbf{I}) \tag{2}$$

Here, $\mathbf{\Phi}_{N_T \times N_T}$ is a known positive semi-definite and symmetric kinship matrix, $\mathbf{I}_{N_T \times N_T}$ is the identity matrix and parameters σ^2 and $\eta \in [0, 1]$ determine how the variance is divided between $\boldsymbol{b}$ and $\boldsymbol{\varepsilon}$. The joint density of $\mathbf{Y}$ is multivariate normal:

$$\mathbf{Y}|(\boldsymbol{\beta}, \eta, \sigma^2) \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \eta\sigma^2\mathbf{\Phi} + (1-\eta)\sigma^2\mathbf{I}) \tag{3}$$

Alternatively we may consider the parameterization in (8):

$$\mathbf{Y}|(\boldsymbol{\beta}, \delta, \sigma_g^2) \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma_g^2(\mathbf{\Phi} + \delta\mathbf{I})) \tag{4}$$

where $\delta = \sigma_e^2/\sigma_g^2$, σ_g^2 is the genetic variance and σ_e^2 is the residual variance. (24) consider the parameterization in (3) since maximization is easier over the compact set $\eta \in [0, 1]$ than over the unbounded interval $\delta \in [0, \infty)$ as is done in (4) by (8).

Define the complete parameter vector $\mathbf{\Theta} = (\boldsymbol{\beta}, \eta, \sigma^2)$. The negative log-likelihood for (3) is given by

$$-\ell(\mathbf{\Theta}) \propto \frac{N_T}{2}\log(\sigma^2) + \frac{1}{2}\log\left(\det(\mathbf{V})\right) + \frac{1}{2\sigma^2}\left(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\right)^T \mathbf{V}^{-1}\left(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\right) \tag{5}$$

where $\mathbf{V} = \eta\mathbf{\Phi} + (1-\eta)\mathbf{I}$ and $\det(\mathbf{V})$ is the determinant of $\mathbf{V}$. Let $\mathbf{\Phi} = \mathbf{U}\mathbf{D}\mathbf{U}^T$ be the eigen (spectral) decomposition of the kinship matrix $\mathbf{\Phi}$, where $\mathbf{U}_{N_T \times N_T}$ is an orthonormal matrix of eigenvectors (i.e. $\mathbf{U}\mathbf{U}^T = \mathbf{I}$) and $\mathbf{D}_{N_T \times N_T}$ is a diagonal matrix of eigenvalues Λ_i. $\mathbf{V}$ can

then be further simplified (24)

$$\begin{aligned}\mathbf{V} &= \eta\mathbf{\Phi} + (1-\eta)\mathbf{I} \\ &= \eta\mathbf{U}\mathbf{D}\mathbf{U}^T + (1-\eta)\mathbf{U}\mathbf{I}\mathbf{U}^T \\ &= \mathbf{U}\eta\mathbf{D}\mathbf{U}^T + \mathbf{U}(1-\eta)\mathbf{I}\mathbf{U}^T \\ &= \mathbf{U}\left(\eta\mathbf{D} + (1-\eta)\mathbf{I}\right)\mathbf{U}^T \\ &= \mathbf{U}\widetilde{\mathbf{D}}\mathbf{U}^T \end{aligned} \tag{6}$$

where

$$\widetilde{\mathbf{D}} = \eta\mathbf{D} + (1-\eta)\mathbf{I} \tag{7}$$

$$= \eta \begin{bmatrix} \Lambda_1 & & & \\ & \Lambda_2 & & \\ & & \ddots & \\ & & & \Lambda_{N_T} \end{bmatrix} + (1-\eta) \begin{bmatrix} 1 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 1+\eta(\Lambda_1-1) & & & \\ & 1+\eta(\Lambda_2-1) & & \\ & & \ddots & \\ & & & 1+\eta(\Lambda_{N_T}-1) \end{bmatrix}$$

$$= \text{diag}\left\{1+\eta(\Lambda_1-1), 1+\eta(\Lambda_2-1), \ldots, 1+\eta(\Lambda_{N_T}-1)\right\} \tag{8}$$

Since (7) is a diagonal matrix, its inverse is also a diagonal matrix:

$$\widetilde{\mathbf{D}}^{-1} = \text{diag}\left\{\frac{1}{1+\eta(\Lambda_1-1)}, \frac{1}{1+\eta(\Lambda_2-1)}, \ldots, \frac{1}{1+\eta(\Lambda_{N_T}-1)}\right\} \tag{9}$$

From (6) and (8), $\log(\det(\mathbf{V}))$ simplifies to

$$\begin{aligned}\log(\det(\mathbf{V})) &= \log\left(\det(\mathbf{U})\det\left(\widetilde{\mathbf{D}}\right)\det(\mathbf{U}^T)\right)\\ &= \log\left\{\prod_{i=1}^{N_T}(1+\eta(\Lambda_i-1))\right\}\\ &= \sum_{i=1}^{N_T}\log(1+\eta(\Lambda_i-1))\end{aligned} \tag{10}$$

since $\det(\mathbf{U}) = 1$. It also follows from (6) that

$$\begin{aligned}\mathbf{V}^{-1} &= \left(\mathbf{U}\widetilde{\mathbf{D}}\mathbf{U}^T\right)^{-1}\\ &= \left(\mathbf{U}^T\right)^{-1}\left(\widetilde{\mathbf{D}}\right)^{-1}\mathbf{U}^{-1}\\ &= \mathbf{U}\widetilde{\mathbf{D}}^{-1}\mathbf{U}^T\end{aligned} \tag{11}$$

since for an orthonormal matrix $\mathbf{U}^{-1} = \mathbf{U}^T$. Substituting (9), (10) and (11) into (5) the negative log-likelihood becomes

$$-\ell(\boldsymbol{\Theta}) \propto \frac{N_T}{2}\log(\sigma^2) + \frac{1}{2}\sum_{i=1}^{N_T}\log(1+\eta(\Lambda_i-1)) + \frac{1}{2\sigma^2}\left(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}\right)^T\mathbf{U}\widetilde{\mathbf{D}}^{-1}\mathbf{U}^T\left(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}\right) \tag{12}$$

$$\begin{aligned}&= \frac{N_T}{2}\log(\sigma^2) + \frac{1}{2}\sum_{i=1}^{N_T}\log(1+\eta(\Lambda_i-1)) + \frac{1}{2\sigma^2}\left(\mathbf{U}^T\mathbf{Y}-\mathbf{U}^T\mathbf{X}\boldsymbol{\beta}\right)^T\widetilde{\mathbf{D}}^{-1}\left(\mathbf{U}^T\mathbf{Y}-\mathbf{U}^T\mathbf{X}\boldsymbol{\beta}\right)\\ &= \frac{N_T}{2}\log(\sigma^2) + \frac{1}{2}\sum_{i=1}^{N_T}\log(1+\eta(\Lambda_i-1)) + \frac{1}{2\sigma^2}\left(\widetilde{\mathbf{Y}}-\widetilde{\mathbf{X}}\boldsymbol{\beta}\right)^T\widetilde{\mathbf{D}}^{-1}\left(\widetilde{\mathbf{Y}}-\widetilde{\mathbf{X}}\boldsymbol{\beta}\right)\\ &= \frac{N_T}{2}\log(\sigma^2) + \frac{1}{2}\sum_{i=1}^{N_T}\log(1+\eta(\Lambda_i-1)) + \frac{1}{2\sigma^2}\sum_{i=1}^{N_T}\frac{\left(\widetilde{Y}_i-\sum_{j=0}^{p}\widetilde{X}_{ij+1}\beta_j\right)^2}{1+\eta(\Lambda_i-1)}\end{aligned} \tag{13}$$

where $\widetilde{\mathbf{Y}} = \mathbf{U}^T\mathbf{Y}$, $\widetilde{\mathbf{X}} = \mathbf{U}^T\mathbf{X}$, $\widetilde{Y}_i$ denotes the i^{th} element of $\widetilde{\mathbf{Y}}$, $\widetilde{X}_{ij}$ is the i,j^{th} entry of $\widetilde{\mathbf{X}}$ and $\mathbf{1}$ is a column vector of N_T ones.

2.2 Penalized Maximum Likelihood Estimator

We define the $p+3$ length vector of parameters $\boldsymbol{\Theta} \coloneqq (\Theta_0, \Theta_1, \ldots, \Theta_{p+1}, \Theta_{p+2}, \Theta_{p+3}) = (\boldsymbol{\beta}, \eta, \sigma^2)$ where $\boldsymbol{\beta} \in \mathbb{R}^{p+1}, \eta \in [0,1], \sigma^2 > 0$. In what follows, $p+2$ and $p+3$ are the indices in $\boldsymbol{\Theta}$ for η and σ^2, respectively. Define the objective function:

$$Q_\lambda(\boldsymbol{\Theta}) = f(\boldsymbol{\Theta}) + \lambda \sum_{j \neq 0} v_j P_j(\beta_j) \tag{14}$$

where $f(\boldsymbol{\Theta}) \coloneqq -\ell(\boldsymbol{\Theta})$ is defined in (13), $P_j(\cdot)$ is a penalty term on the fixed regression coefficients $\beta_1, \ldots, \beta_{p+1}$ (we do not penalize the intercept), controlled by the nonnegative regularization parameter λ, and v_j is the penalty factor for jth covariate. These penalty factors serve as a way of allowing parameters to be penalized differently. Note that we do not penalize η or σ^2. The penalty term is a necessary constraint because in our applications, the sample size is much smaller than the number of predictors. An estimate of the regression parameters $\widehat{\boldsymbol{\Theta}}_\lambda$ is obtained by

$$\widehat{\boldsymbol{\Theta}}_\lambda = \arg\min_{\boldsymbol{\Theta}} Q_\lambda(\boldsymbol{\Theta}) \tag{15}$$

3 Computational Algorithm version 1

To solve for (15) we use a block relaxation technique (25) given by Algorithm 1

Algorithm 1: Block Relaxation Algorithm

Set the iteration counter $k \leftarrow 0$, initial values for the parameter vector $\boldsymbol{\Theta}^{(0)}$ and convergence threshold ϵ;

for $\lambda \in \{\lambda_{max}, \ldots, \lambda_{min}\}$ **do**

repeat

$$\boldsymbol{\beta}^{(k+1)} \leftarrow \arg\min_{\boldsymbol{\beta}} Q_\lambda \left(\boldsymbol{\beta}, \eta^{(k)}, \sigma^{2\ (k)}\right)$$

$$\eta^{(k+1)} \leftarrow \arg\min_{\eta} Q_\lambda \left(\boldsymbol{\beta}^{(k+1)}, \eta, \sigma^{2\ (k)}\right)$$

$$\sigma^{2\ (k+1)} \leftarrow \arg\min_{\sigma^2} Q_\lambda \left(\boldsymbol{\beta}^{(k+1)}, \eta^{(k+1)}, \sigma^2\right)$$

$k \leftarrow k + 1$

until *convergence criterion is satisfied:* $\left\|\boldsymbol{\Theta}^{(k+1)} - \boldsymbol{\Theta}^{(k)}\right\|_2 < \epsilon$;

end

Below we discuss the specifics of Algorithm 1

3.1 Updates for the β parameter

Recall that the part of the objective function that depends on $\boldsymbol{\beta}$ has the form

$$Q_\lambda(\boldsymbol{\Theta}) = \frac{1}{2}\sum_{i=1}^{N_T} w_i \left(\widetilde{Y}_i - \sum_{j=0}^{p} \widetilde{X}_{ij+1}\beta_j\right)^2 + \lambda \sum_{j=1}^{p} v_j|\beta_j| \tag{16}$$

where

$$w_i \coloneqq \frac{1}{\sigma^2\left(1+\eta(\Lambda_i - 1)\right)} \tag{17}$$

However `glmnet` solves the following problem:

$$\boldsymbol{\beta}^{(k+1)} \leftarrow \arg\min_{\boldsymbol{\beta}} \frac{1}{2\sum_{i=1}^{N_T}\widetilde{w}_i^{(k)}} \sum_{i=1}^{N_T} \widetilde{w}_i^{(k)} \left(\widetilde{Y}_i - \sum_{j=0}^{p} \widetilde{X}_{ij+1}\beta_j\right)^2 + \lambda \sum_{j=1}^{p} v_j|\beta_j| \tag{18}$$

where

$$\widetilde{w}_i^{(k)} = N_T \cdot \frac{w_i^{(k)}}{\sum_{i=1}^{N_T} w_i^{(k)}} \tag{19}$$

Note that $\sum_i \widetilde{w}_i^{(k)} = N_T$. We can simplify (18) to be:

$$\boldsymbol{\beta}^{(k+1)} \leftarrow \arg\min_{\boldsymbol{\beta}} \frac{1}{2N_T} \sum_{i=1}^{N_T} N_T \cdot \frac{w_i^{(k)}}{\sum_{i=1}^{N_T} w_i^{(k)}} \left(\widetilde{Y}_i - \sum_{j=0}^{p} \widetilde{X}_{ij+1}\beta_j\right)^2 + \lambda \sum_{j=1}^{p} v_j|\beta_j|$$

$$\boldsymbol{\beta}^{(k+1)} \leftarrow \arg\min_{\boldsymbol{\beta}} \frac{1}{2\sum_{i=1}^{N_T} w_i^{(k)}} \sum_{i=1}^{N_T} w_i^{(k)} \left(\widetilde{Y}_i - \sum_{j=0}^{p} \widetilde{X}_{ij+1}\beta_j\right)^2 + \lambda \sum_{j=1}^{p} v_j|\beta_j| \tag{20}$$

In order to make (16) to be in the form of (20), we must scale the lambda accordingly:

$$\boldsymbol{\beta}^{(k+1)} \leftarrow \arg\min_{\boldsymbol{\beta}} \frac{1}{2} \sum_{i=1}^{N_T} w_i^{(k)} \left(\widetilde{Y}_i - \sum_{j=0}^{p} \widetilde{X}_{ij+1}\beta_j\right)^2 + \frac{\lambda}{\sum_{i=1}^{N_T} w_i^{(k)}} \sum_{j=1}^{p} v_j|\beta_j| \tag{21}$$

Conditional on $\eta^{(k)}$ and $\sigma^{2\,(k)}$, it can be shown that the solution for $\boldsymbol{\beta}$ is a weighted lasso problem with observation weights given by (17).

The full derivation is given in Section 7.1. Therefore, $\boldsymbol{\beta}^{(k+1)}$ can be efficiently solved using the `glmnet` algorithm (20). Note that the rescaling of the weights to sum to N_T is what is being done in `glmnet`.

3.2 Updates for the η paramter

$$\eta^{(k+1)} \leftarrow \arg\min_{\eta} \frac{1}{2} \sum_{i=1}^{N_T} \log(1+\eta(\Lambda_i - 1)) + \frac{1}{2\sigma^{2\,(k)}} \sum_{i=1}^{N_T} \frac{\left(\widetilde{Y}_i - \sum_{j=0}^{p} \widetilde{X}_{ij+1} \beta_j^{(k+1)}\right)^2}{1+\eta(\Lambda_i - 1)} \tag{22}$$

Given $\boldsymbol{\beta}^{(k+1)}$ and $\sigma^{2\,(k)}$, solving for $\eta^{(k+1)}$ becomes a univariate optimization problem. We use a bound constrained optimization algorithm (26) implemented in the `optim` function in `R` and set the lower and upper bounds to be 0 and 1, respectively.

3.3 Updates for the σ^2 parameter

$$\sigma^{2\,(k+1)} \leftarrow \arg\min_{\sigma^2} \frac{N_T}{2} \log(\sigma^2) + \frac{1}{2\sigma^2} \sum_{i=1}^{N_T} \frac{\left(\widetilde{Y}_i - \sum_{j=0}^{p} \widetilde{X}_{ij+1} \beta_j\right)^2}{1+\eta(\Lambda_i - 1)} \tag{23}$$

Conditional on $\boldsymbol{\beta}^{(k+1)}$ and $\eta^{(k+1)}$, there exists an analytic solution for $\sigma^{2\,(k+1)}$:

$$\begin{aligned} \frac{\partial}{\partial \sigma^2} Q_\lambda(\boldsymbol{\Theta}) &= \frac{N_T}{2\sigma^2} - \frac{1}{2\sigma^4} \sum_{i=1}^{N_T} \frac{\left(\widetilde{Y}_i - \sum_{j=0}^{p} \widetilde{X}_{ij+1} \beta_j^{(k+1)}\right)^2}{1+\eta^{(k+1)}(\Lambda_i - 1)} = 0 \\ \sigma^{2\,(k+1)} &= \frac{1}{N_T} \sum_{i=1}^{N_T} \frac{\left(\widetilde{Y}_i - \sum_{j=0}^{p} \widetilde{X}_{ij+1} \beta_j^{(k+1)}\right)^2}{1+\eta^{(k+1)}(\Lambda_i - 1)} \end{aligned} \tag{24}$$

3.4 Regularization path

Recall that our objective function has the form

$$Q_\lambda(\mathbf{\Theta}) = \frac{N_T}{2}\log(\sigma^2) + \frac{1}{2}\sum_{i=1}^{N_T}\log(1+\eta(\Lambda_i-1)) + \frac{1}{2}\sum_{i=1}^{N_T} w_i\left(\widetilde{Y}_i - \sum_{j=0}^{p}\widetilde{X}_{ij+1}\beta_j\right)^2 + \lambda\sum_{j=1}^{p} v_j|\beta_j| \tag{25}$$

The Karush-Kuhn-Tucker (KKT) optimality conditions for (25) are given by:

$$\begin{aligned}
\frac{\partial}{\partial\beta_1,\ldots,\beta_p}Q_\lambda(\mathbf{\Theta}) &= \mathbf{0}_p \\
\frac{\partial}{\partial\beta_0}Q_\lambda(\mathbf{\Theta}) &= 0 \\
\frac{\partial}{\partial\eta}Q_\lambda(\mathbf{\Theta}) &= 0 \\
\frac{\partial}{\partial\sigma^2}Q_\lambda(\mathbf{\Theta}) &= 0
\end{aligned} \tag{26}$$

The equations in (26) are equivalent to

$$\begin{aligned}
\sum_{i=1}^{N_T} w_i\widetilde{X}_{i1}\left(\widetilde{Y}_i - \sum_{j=0}^{p}\widetilde{X}_{ij+1}\beta_j\right) &= 0 \\
\frac{1}{v_j}\sum_{i=1}^{N_T} w_i\widetilde{X}_{ij}\left(\widetilde{Y}_i - \sum_{j=0}^{p}\widetilde{X}_{ij+1}\beta_j\right) &= \lambda\gamma_j, \\
\gamma_j \in \begin{cases} \text{sign}(\hat{\beta}_j) & \text{if} \quad \hat{\beta}_j \neq 0 \\ [-1,1] & \text{if} \quad \hat{\beta}_j = 0 \end{cases}, &\qquad \text{for } j = 1,\ldots,p \\
\frac{1}{2}\sum_{i=1}^{N_T}\frac{\Lambda_i - 1}{1+\eta(\Lambda_i-1)}\left(1 - \frac{\left(\widetilde{Y}_i - \sum_{j=0}^{p}\widetilde{X}_{ij+1}\beta_j\right)^2}{\sigma^2(1+\eta(\Lambda_i-1))}\right) &= 0 \\
\sigma^2 - \frac{1}{N_T}\sum_{i=1}^{N_T}\frac{\left(\widetilde{Y}_i - \sum_{j=0}^{p}\widetilde{X}_{ij+1}\beta_j\right)^2}{1+\eta(\Lambda_i-1)} &= 0
\end{aligned} \tag{27}$$

where w_i is given by (17), $\widetilde{\mathbf{X}}_{-1}^T$ is $\widetilde{\mathbf{X}}^T$ with the first column removed, $\widetilde{\mathbf{X}}_1^T$ is the first column

of $\widetilde{\mathbf{X}}^T$, and $\boldsymbol{\gamma} \in \mathbb{R}^p$ is the subgradient function of the ℓ_1 norm evaluated at $(\hat{\beta}_1, \ldots, \hat{\beta}_p)$. Therefore $\widehat{\boldsymbol{\Theta}}$ is a solution in (15) if and only if $\widehat{\boldsymbol{\Theta}}$ satisfies (27) for some γ.

we find the solutinon for the other paramters such that the KKT conditions are verified. page 17 of sswith learning

Therefore we can determine a decreasing sequence of tuning parameters by starting at a maximal value for $\lambda = \lambda_{max}$ for which $\hat{\beta}_j = 0$ for $j = 1, \ldots, p$. In this case, the KKT conditions in (27) are equivalent to

$$\begin{aligned} \frac{1}{v_j}\sum_{i=1}^{N_T}\left|w_i\widetilde{X}_{ij}\left(\widetilde{Y}_i - \widetilde{X}_{i1}\beta_0\right)\right| \leq \lambda, \quad \forall j = 1, \ldots, p \\ \beta_0 = \frac{\sum_{i=1}^{N_T} w_i\widetilde{X}_{i1}\widetilde{Y}_i}{\sum_{i=1}^{N_T} w_i\widetilde{X}_{i1}^2} \\ \frac{1}{2}\sum_{i=1}^{N_T}\frac{\Lambda_i - 1}{1+\eta(\Lambda_i - 1)}\left(1 - \frac{\left(\widetilde{Y}_i - \widetilde{X}_{i1}\beta_0\right)^2}{\sigma^2(1+\eta(\Lambda_i-1))}\right) = 0 \\ \sigma^2 = \frac{1}{N_T}\sum_{i=1}^{N_T}\frac{\left(\widetilde{Y}_i - \widetilde{X}_{i1}\beta_0\right)^2}{1+\eta(\Lambda_i-1)} \end{aligned} \tag{28}$$

We can solve the KKT system of equations in (28) (with a numerical solution for η) in order to have an explicit form of the stationary point $\widehat{\boldsymbol{\Theta}}_0 = \left\{\hat{\beta}_0, \mathbf{0}_p, \hat{\eta}, \hat{\sigma}^2\right\}$. Once we have $\widehat{\boldsymbol{\Theta}}_0$, we can solve for the smallest value of λ such that the entire vector $(\hat{\beta}_1, \ldots, \hat{\beta}_p)$ is 0:

$$\lambda_{max} = \max_j\left\{\left|\frac{1}{v_j}\sum_{i=1}^{N_T}\hat{w}_i\widetilde{X}_{ij}\left(\widetilde{Y}_i - \widetilde{X}_{i1}\hat{\beta}_0\right)\right|\right\}, \quad j = 1, \ldots, p \tag{29}$$

Following (20), we choose $\tau\lambda_{max}$ to be the smallest value of tuning parameters λ_{min}, and construct a sequence of K values decreasing from λ_{max} to λ_{min} on the log scale. The defaults are set to $K = 100$, $\tau = 0.01$ if $n < p$ and $\tau = 0.001$ if $n \geq p$.

3.5 Warm Starts

The way in which we have derived the sequence of tuning parameters using the KKT conditions, allows us to implement warm starts. That is, the solution $\widehat{\boldsymbol{\Theta}}$ for λ_k is used as the initial value $\boldsymbol{\Theta}^{(0)}$ for λ_{k+1}.

3.6 Prediction of the random effects

We use an empirical Bayes approach (e.g. (27)) to predict the random effects $\boldsymbol{b}$. Let the maximum a posteriori (MAP) estimate be defined as

$$\widehat{\boldsymbol{b}} = \arg\max_{\boldsymbol{b}} f(\boldsymbol{b}|\mathbf{Y}, \boldsymbol{\beta}, \eta, \sigma^2) \tag{30}$$

where, by using Bayes rule, $f(\boldsymbol{b}|\mathbf{Y}, \boldsymbol{\beta}, \eta, \sigma^2)$ can be expressed as

$$\begin{aligned} f(\boldsymbol{b}|\mathbf{Y}, \boldsymbol{\beta}, \eta, \sigma^2) &= \frac{f(\mathbf{Y}|\boldsymbol{b}, \boldsymbol{\beta}, \eta, \sigma^2)\pi(\boldsymbol{b}|\eta, \sigma^2)}{f(\mathbf{Y}|\boldsymbol{\beta}, \eta, \sigma^2)} \\ &\propto f(\mathbf{Y}|\boldsymbol{b}, \boldsymbol{\beta}, \eta, \sigma^2)\pi(\boldsymbol{b}|\eta, \sigma^2) \\ &\propto \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \boldsymbol{b})^T\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \boldsymbol{b}) - \frac{1}{2\eta\sigma^2}\boldsymbol{b}^T\boldsymbol{\Phi}^{-1}\boldsymbol{b}\right\} \\ &= \exp\left\{-\frac{1}{2\sigma^2}\left[(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \boldsymbol{b})^T\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \boldsymbol{b}) + \frac{1}{\eta}\boldsymbol{b}^T\boldsymbol{\Phi}^{-1}\boldsymbol{b}\right]\right\} \end{aligned} \tag{31}$$

Solving for (30) is equivalent to minimizing the exponent in (31):

$$\widehat{\boldsymbol{b}} = \arg\min_{\boldsymbol{b}} \left\{(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \boldsymbol{b})^T\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \boldsymbol{b}) + \frac{1}{\eta}\boldsymbol{b}^T\boldsymbol{\Phi}^{-1}\boldsymbol{b}\right\} \tag{32}$$

Taking the derivative of (32) with respect to $\boldsymbol{b}$ and setting it to 0 we get:

$$\begin{aligned}
0 &= -2\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \boldsymbol{b}) + \frac{2}{\eta}\boldsymbol{\Phi}^{-1}\boldsymbol{b} \\
&= -\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + \left(\mathbf{V}^{-1} + \frac{1}{\eta}\boldsymbol{\Phi}^{-1}\right)\boldsymbol{b} \\
\widehat{\boldsymbol{b}} &= \left(\mathbf{V}^{-1} + \frac{1}{\widehat{\eta}}\boldsymbol{\Phi}^{-1}\right)^{-1}\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) \\
&= \left(\mathbf{U}\widetilde{\mathbf{D}}^{-1}\mathbf{U}^T + \frac{1}{\widehat{\eta}}\mathbf{U}\mathbf{D}^{-1}\mathbf{U}^T\right)^{-1}\mathbf{U}\widetilde{\mathbf{D}}^{-1}\mathbf{U}^T(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) \\
&= \left(\mathbf{U}\left[\widetilde{\mathbf{D}}^{-1} + \frac{1}{\widehat{\eta}}\mathbf{D}^{-1}\right]\mathbf{U}^T\right)^{-1}\mathbf{U}\widetilde{\mathbf{D}}^{-1}(\widetilde{\mathbf{Y}} - \widetilde{\mathbf{X}}\widehat{\boldsymbol{\beta}}) \\
&= \mathbf{U}\left[\widetilde{\mathbf{D}}^{-1} + \frac{1}{\widehat{\eta}}\mathbf{D}^{-1}\right]^{-1}\mathbf{U}^T\mathbf{U}\widetilde{\mathbf{D}}^{-1}(\widetilde{\mathbf{Y}} - \widetilde{\mathbf{X}}\widehat{\boldsymbol{\beta}})
\end{aligned}$$

where $\mathbf{V}^{-1}$ is given by (11), and $(\widehat{\boldsymbol{\beta}}, \widehat{\eta})$ are the estimates obtained from Algorithm 1.

3.7 Choice of the tuning parameter

We use the BIC:

$$BIC_\lambda = -2\ell(\widehat{\boldsymbol{\beta}}, \widehat{\sigma}^2, \widehat{\eta}) + c \cdot \widehat{df}_\lambda \tag{33}$$

where $\widehat{df}_\lambda$ is the number of non-zero elements in $\widehat{\boldsymbol{\beta}}_\lambda$ (28) plus two (representing the variance parameters η and σ^2). Several authors have used this criterion for variable selection in mixed models with $c = \log N_T$ (29, 30) and $c = \log N$ (31) (where N is the number of groups). Other authors have proposed $c = \log(\log(N_T)) * \log(N_T)$ (32).

4 Low rank similarity matrix

Let $\mathbf{K} \in \mathbb{R}^{N_T \times k}$ be the matrix containing the k SNPs used to compute the factored kinship matrix $\boldsymbol{\Phi}$ given by

$$\boldsymbol{\Phi} = \mathbf{K}\mathbf{K}^T \tag{34}$$

Furthermore, let $\mathbf{K} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^T$ be the singular value decomposition (SVD) of $\mathbf{K}$. Plugging this into (34) we get

$$\begin{aligned}\mathbf{\Phi} &= \left(\mathbf{U}\mathbf{\Lambda}\mathbf{V}^T\right)\left(\mathbf{U}\mathbf{\Lambda}\mathbf{V}^T\right)^T \\ &= \mathbf{U}\mathbf{\Lambda}\mathbf{V}^T\mathbf{V}\mathbf{\Lambda}\mathbf{U}^T \\ &= \mathbf{U}\mathbf{\Lambda}\mathbf{\Lambda}\mathbf{U}^T \\ &= \mathbf{U}\mathbf{\Sigma}\mathbf{U}^T, \end{aligned} \tag{35}$$

Therefore, the eigenvectors of $\mathbf{\Phi}$ are equal to the singular vectors of $\mathbf{K}$ (denoted by $\mathbf{U}$), and the eigenvalues of $\mathbf{\Phi}$ (denoted by the diagonal matrix $\mathbf{\Sigma}$) are equal to the square of the singular values of $\mathbf{K}$ (33). This allows us to bypass the explicit computation of the kinship matrix by directly applying SVD on the SNP matrix $\mathbf{W}$. (8) noted that the computational time for fitting the LMM can be reduced if the matrix $\mathbf{K}$ is not full rank, i.e., when $k < N_T$. This is due to the fact that the matrix $\mathbf{D}_{N_T \times N_T}$ contains k non-zero eigenvalues followed by $N_T - k$ zeros on the diagonal. Let $\mathbf{U} \equiv [\mathbf{U}_1\ \mathbf{U}_2]$, where $\mathbf{U}_1 \in \mathbb{R}^{N_T \times k}$ and $\mathbf{U}_2 \in \mathbb{R}^{N_T \times (n-k)}$ are the matrices of singular vectors corresponding to the k non-zero and $N_T - k$ zero eigenvalues, respectively. Then (35) can be written as

$$\mathbf{\Phi} = \mathbf{U}_1\mathbf{\Sigma}\mathbf{U}_1^T \tag{36}$$

We now try to simplify the log-likelihood (12). Since there are $N_T - k$ zero eigenvalues, the second term in (12) reduces to

$$\frac{1}{2}\left(\sum_{i=1}^{k}\log(1+\eta(\Sigma_i - 1)) + (N_T - k)\log(1-\eta)\right) \tag{37}$$

where $\Sigma_i = \Lambda_i^2$, and Λ_i is the i^{th} singular value of $\mathbf{W}$. Let $a \equiv (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$. The third term

in (12) can be written as

$$\begin{aligned}\frac{1}{2\sigma^2}a^T\left[\eta\mathbf{\Phi}+(1-\eta)\mathbf{I}_n\right]^{-1}a &= \frac{1}{2\sigma^2}a^T\left[\eta\mathbf{U}_1\mathbf{\Sigma}_1\mathbf{U}_1^T+(1-\eta)\mathbf{I}_n\right]^{-1}a \\ &= \frac{1}{2\sigma^2}a^T\left[\mathbf{CBC}^T+\mathbf{A}\right]^{-1}a\end{aligned}$$

where

$$\begin{aligned}\mathbf{A} &= (1-\eta)\mathbf{I}_n \\ \mathbf{B} &= \mathbf{\Sigma}_1 \\ \mathbf{C} &= \sqrt{\eta}\mathbf{U}_1 \\ \mathbf{C}^T &= \sqrt{\eta}\mathbf{U}_1^T\end{aligned}$$

Assuming $\mathbf{CBC}^T+\mathbf{A}$ is non-singular, the inverse of $\left[\mathbf{CBC}^T+\mathbf{A}\right]$ is given explicitly by the Woodbury formula (34)

$$\left(\mathbf{A}+\mathbf{CBC}^T\right)^{-1} = \mathbf{A}^{-1}-\mathbf{A}^{-1}\mathbf{C}\left(\mathbf{B}^{-1}+\mathbf{C}^T\mathbf{A}^{-1}\mathbf{C}\right)^{-1}\mathbf{C}^T\mathbf{A}^{-1} \tag{38}$$

Substituting the values for $\mathbf{A},\mathbf{B}$ and $\mathbf{C}$ into (38) we get

$$\begin{aligned}\left(\mathbf{A}+\mathbf{CBC}^T\right)^{-1} &= \frac{1}{1-\eta}\mathbf{I}_{N_T}-\frac{\sqrt{\eta}}{1-\eta}\mathbf{I}_{N_T}\mathbf{U}_1\left(\mathbf{\Sigma}_1^{-1}+\frac{\eta}{1-\eta}\mathbf{U}_1^T\mathbf{I}_{N_T}\mathbf{U}\right)^{-1}\frac{\sqrt{\eta}}{1-\eta}\mathbf{U}_1^T\mathbf{I}_{N_T} \\ &= \frac{1}{1-\eta}\left[\mathbf{I}_{N_T}-\frac{\eta}{1-\eta}\mathbf{U}_1\left(\mathbf{\Sigma}_1^{-1}+\frac{\eta}{1-\eta}\mathbf{I}_k\right)^{-1}\mathbf{U}_1^T\right] \\ &= \frac{1}{1-\eta}\left[\mathbf{I}_{N_T}-\frac{\eta}{1-\eta}\mathbf{U}_1\left(\frac{\eta}{1-\eta}\left(\frac{1-\eta}{\eta}\mathbf{\Sigma}_1^{-1}+\mathbf{I}_k\right)\right)^{-1}\mathbf{U}_1^T\right] \\ &= \frac{1}{1-\eta}\left[\mathbf{I}_{N_T}-\mathbf{U}_1\left(\frac{1-\eta}{\eta}\mathbf{\Sigma}_1^{-1}+\mathbf{I}_k\right)^{-1}\mathbf{U}_1^T\right]\end{aligned} \tag{39}$$

where we have used the following identities: $\mathbf{I}_k=\mathbf{U}_1^T\mathbf{U}_1$, $\mathbf{I}_{N_T-k}=\mathbf{U}_2^T\mathbf{U}_2$.

Substituting (37) and (39) in (12) we obtain

$$-\ell(\boldsymbol{\Theta}) \propto \frac{N_T}{2}\log(\sigma^2) + \frac{1}{2}\left(\sum_{i=1}^{k}\log(1+\eta(\Sigma_i-1)) + (N_T-k)\log(1-\eta)\right) + \frac{1}{2}\left\{(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})^T\left[\frac{1}{\sigma^2(1-\eta)}\left(\mathbf{I}_{N_T} - \mathbf{U}_1\left(\frac{1-\eta}{\eta}\boldsymbol{\Sigma}_1^{-1} + \mathbf{I}_k\right)^{-1}\mathbf{U}_1^T\right)\right](\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})\right\} \tag{40}$$

5 Group Lasso with Low-rank Similarity Matrix

This section focuses on the part of the log-likelihood (40) that depends on $\boldsymbol{\beta}$.

5.1 Model

Only the third term of the log-likelihood (40) depends on $\boldsymbol{\beta}$:

$$\frac{1}{2}\left\{(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})^T\left[\frac{1}{\sigma^2(1-\eta)}\left(\mathbf{I}_{N_T} - \mathbf{U}_1\left(\frac{1-\eta}{\eta}\boldsymbol{\Sigma}_1^{-1} + \mathbf{I}_k\right)^{-1}\mathbf{U}_1^T\right)\right](\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})\right\} \tag{41}$$

Equation (41) can be written more generally as

$$L(\boldsymbol{\beta} \mid \mathbf{D}) = \frac{1}{2}\left[\mathbf{Y}-\widehat{\mathbf{Y}}\right]^\top \mathbf{W}\left[\mathbf{Y}-\widehat{\mathbf{Y}}\right]$$

where $\widehat{\mathbf{Y}} = \sum_{j=1}^{p}\beta_j X_j$, $\mathbf{D}$ is the working data $\{\mathbf{Y}, \mathbf{X}\}$, and $\mathbf{W}$ is an $N_T \times N_T$ weight matrix given by

$$\mathbf{W} = \frac{1}{\sigma^2(1-\eta)}\left(\mathbf{I}_{N_T} - \mathbf{U}_1\left(\frac{1-\eta}{\eta}\boldsymbol{\Sigma}_1^{-1} + \mathbf{I}_k\right)^{-1}\mathbf{U}_1^T\right) \tag{42}$$

Assume that we the predictors in the design matrix $\mathbf{X} \in \mathbb{R}^{N_T \times p}$ belong to K groups and that the group membership is already defined such that $(1, 2, \ldots, p) = \bigcup_{k=1}^{K} I_k$ and the cardinality of index set I_k is p_k, $I_k \bigcap I_{k'} = \emptyset$ for $k \neq k', 1 \leq k, k' \leq K$. Thus group k contains p_k

predictors, which are x_j's for $j \in I_k$, and $1 \leq k \leq K$. If an intercept is included, then $I_1 = \{1\}$. Given the group partition, we use $\boldsymbol{\beta}_{(k)}$ to denote the segment of $\boldsymbol{\beta}$ corresponding to group k. This notation is used for any p-dimensional vector. We consider the group lasso penalized estimator

$$\min_{\boldsymbol{\beta}} L(\boldsymbol{\beta} \mid \mathbf{D}) + \lambda \sum_{k=1}^{K} w_k \|\boldsymbol{\beta}_{(k)}\|_2, \tag{43}$$

The loss function L satisfies the quadratic majorization (QM) condition, since there exists a $p \times p$ matrix $\mathbf{H} = \mathbf{X}^\top \mathbf{W} \mathbf{X}$, and $\nabla L(\boldsymbol{\beta}|\mathbf{D}) = -\left(Y - \hat{Y}\right)^\top \mathbf{W}\mathbf{X}$, which may only depend on the data $\mathbf{D}$, such that for all $\boldsymbol{\beta}, \boldsymbol{\beta}^*$,

$$L(\boldsymbol{\beta} \mid \mathbf{D}) \leq L(\boldsymbol{\beta}^* \mid \mathbf{D}) + (\boldsymbol{\beta} - \boldsymbol{\beta}^*)^\top \nabla L(\boldsymbol{\beta}^*|\mathbf{D}) + \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)^\top \mathbf{H} (\boldsymbol{\beta} - \boldsymbol{\beta}^*). \tag{44}$$

5.2 Algorithm

Noticing that the penalty term $\sum_{k=1}^{K} w_k ||\boldsymbol{\beta}_{(k)}||_2$ is separable with respect to the indices of the features $k = 1, \ldots, K$, we can derive the *groupwise-majorization-descent* (GMD) algorithm for computing the solution of (43) when the loss function satisfies the QM condition. Let $\widetilde{\boldsymbol{\beta}}$ denote the current solution of $\boldsymbol{\beta}$. Without loss of generality, let us derive the GMD update of $\widetilde{\boldsymbol{\beta}}_{(k)}$, the coefficients of group k. Define $\mathbf{H}_k$ as the sub-matrix of $\mathbf{H}$ corresponding to group k. For example, if group 2 is $\{2, 4\}$ then $\mathbf{H}_{(2)}$ is a 2×2 matrix with

$$\mathbf{H}_{(2)} = \begin{bmatrix} h_{2,2} & h_{2,4} \\ h_{4,2} & h_{4,4} \end{bmatrix},$$

where $h_{i,j}$ is the i,jth entry of the $\mathbf{H}$ matrix. Write $\boldsymbol{\beta}$ such that $\boldsymbol{\beta}_{(k')} = \widetilde{\boldsymbol{\beta}}_{(k')}$ for $k' \neq k$. Given $\boldsymbol{\beta}_{(k')} = \widetilde{\boldsymbol{\beta}}_{(k')}$ for $k' \neq k$, the optimal $\boldsymbol{\beta}_{(k)}$ is defined as

$$\arg\min_{\boldsymbol{\beta}^{(k)}} L(\boldsymbol{\beta} \mid \mathbf{D}) + \lambda w_k \|\boldsymbol{\beta}_{(k)}\|_2. \tag{45}$$

Unfortunately, there is no closed form solution to (45) for a general loss function with general design matrix. We overcome the computational obstacle by taking advantage of the QM condition. From (44) we have

$$L(\boldsymbol{\beta} \mid \mathbf{D}) \leq L(\widetilde{\boldsymbol{\beta}} \mid \mathbf{D}) + (\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}})^\top \nabla L(\widetilde{\boldsymbol{\beta}}|\mathbf{D}) + \frac{1}{2}(\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}})^\top \mathbf{H}(\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}).$$

Write $U(\widetilde{\boldsymbol{\beta}}) = -\nabla L(\widetilde{\boldsymbol{\beta}}|\mathbf{D})$. Using

$$\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}} = (\underbrace{0, \ldots, 0}_{k-1}, \boldsymbol{\beta}_{(k)} - \widetilde{\boldsymbol{\beta}}_{(k)}, \underbrace{0, \ldots, 0}_{K-k}),$$

we can write

$$L(\boldsymbol{\beta} \mid \mathbf{D}) \leq L(\widetilde{\boldsymbol{\beta}} \mid \mathbf{D}) - (\boldsymbol{\beta}_{(k)} - \widetilde{\boldsymbol{\beta}}_{(k)})^\top U_{(k)} + \frac{1}{2}(\boldsymbol{\beta}_{(k)} - \widetilde{\boldsymbol{\beta}}_{(k)})^\top \mathbf{H}_{(k)}(\boldsymbol{\beta}_{(k)} - \widetilde{\boldsymbol{\beta}}_{(k)}). \tag{46}$$

where

$$U_{(k)} = \frac{\partial}{\partial \boldsymbol{\beta}_{(k)}} L_Q(\boldsymbol{\beta} \mid \mathbf{D}) = -\left(Y - \hat{Y}\right)^\top \mathbf{W}\mathbf{X}_{(k)}, \tag{47}$$

$$\mathbf{H}_{(k)} = \frac{\partial^2}{\partial \boldsymbol{\beta}_{(k)} \partial \boldsymbol{\beta}_{(k)}^\top} L_Q(\boldsymbol{\beta} \mid \mathbf{D}) = \mathbf{X}_{(k)}^\top \mathbf{W}\mathbf{X}_{(k)}. \tag{48}$$

Let η_k be the largest eigenvalue of $\mathbf{H}_{(k)}$. We set $\gamma_k = (1 + \varepsilon^*)\eta_k$, where $\varepsilon^* = 10^{-6}$. Then we can further relax the upper bound in (46) as

$$L(\boldsymbol{\beta} \mid \mathbf{D}) \leq L(\widetilde{\boldsymbol{\beta}} \mid \mathbf{D}) - (\boldsymbol{\beta}^{(k)} - \widetilde{\boldsymbol{\beta}}^{(k)})^\top U_{(k)} + \frac{1}{2}\gamma_k(\boldsymbol{\beta}^{(k)} - \widetilde{\boldsymbol{\beta}}^{(k)})^\top(\boldsymbol{\beta}^{(k)} - \widetilde{\boldsymbol{\beta}}^{(k)}). \tag{49}$$

It is important to note that the inequality strictly holds unless for $\boldsymbol{\beta}^{(k)} = \widetilde{\boldsymbol{\beta}}^{(k)}$. Instead of

minimizing (45) we solve

$$\arg\min_{\boldsymbol{\beta}^{(k)}} L(\widetilde{\boldsymbol{\beta}} \mid \mathbf{D}) - (\boldsymbol{\beta}^{(k)} - \widetilde{\boldsymbol{\beta}}^{(k)})^\top U_{(k)} + \frac{1}{2}\gamma_k(\boldsymbol{\beta}^{(k)} - \widetilde{\boldsymbol{\beta}}^{(k)})^\top(\boldsymbol{\beta}^{(k)} - \widetilde{\boldsymbol{\beta}}^{(k)}) + \lambda w_k \|\boldsymbol{\beta}^{(k)}\|_2. \quad (50)$$

Denote by $\widetilde{\boldsymbol{\beta}}^{(k)}(\text{new})$ the solution to (50). It is straightforward to see that $\widetilde{\boldsymbol{\beta}}^{(k)}(\text{new})$ has a simple closed-from expression

$$\widetilde{\boldsymbol{\beta}}^{(k)}(\text{new}) = \frac{1}{\gamma_k}\left(U^{(k)} + \gamma_k \widetilde{\boldsymbol{\beta}}^{(k)}\right)\left(1 - \frac{\lambda w_k}{\|U^{(k)} + \gamma_k \widetilde{\boldsymbol{\beta}}^{(k)}\|_2}\right)_+. \quad (51)$$

Algorithm 2 summarizes the details of GMD.

Algorithm 2: The GMD algorithm for general group-lasso learning.

1. For $k = 1, \ldots, K$, compute γ_k, the largest eigenvalue of $\mathbf{H}^{(k)}$.
2. Initialize $\widetilde{\boldsymbol{\beta}}$.
3. Repeat the following cyclic groupwise updates until convergence:
   — for $k = 1, \ldots, K$, do step (3.1)–(3.3)
   3.1 Compute $U(\widetilde{\boldsymbol{\beta}}) = -\nabla L(\widetilde{\boldsymbol{\beta}}|\mathbf{D})$.
   3.2 Compute $\widetilde{\boldsymbol{\beta}}^{(k)}(\text{new}) = \frac{1}{\gamma_k}\left(U^{(k)} + \gamma_k \widetilde{\boldsymbol{\beta}}^{(k)}\right)\left(1 - \frac{\lambda w_k}{\|U^{(k)} + \gamma_k \widetilde{\boldsymbol{\beta}}^{(k)}\|_2}\right)_+$.
   3.3 Set $\widetilde{\boldsymbol{\beta}}^{(k)} = \widetilde{\boldsymbol{\beta}}^{(k)}(\text{new})$.

5.3 Convergence

We can prove the strict descent property of GMD by using the MM principle (35, 36, 37). Define

$$Q(\boldsymbol{\beta} \mid \mathbf{D}) = L(\widetilde{\boldsymbol{\beta}} \mid \mathbf{D}) - (\boldsymbol{\beta}^{(k)} - \widetilde{\boldsymbol{\beta}}^{(k)})^\top U^{(k)} + \frac{1}{2}\gamma_k(\boldsymbol{\beta}^{(k)} - \widetilde{\boldsymbol{\beta}}^{(k)})^\top(\boldsymbol{\beta}^{(k)} - \widetilde{\boldsymbol{\beta}}^{(k)}) + \lambda w_k \|\boldsymbol{\beta}^{(k)}\|_2. \quad (52)$$

Obviously, $Q(\boldsymbol{\beta} \mid \mathbf{D}) = L(\boldsymbol{\beta} \mid \mathbf{D}) + \lambda w_k \|\boldsymbol{\beta}^{(k)}\|_2$ when $\boldsymbol{\beta}^{(k)} = \widetilde{\boldsymbol{\beta}}^{(k)}$ and (??) shows that $Q(\boldsymbol{\beta} \mid \mathbf{D}) > L(\boldsymbol{\beta} \mid \mathbf{D}) + \lambda w_k \|\boldsymbol{\beta}^{(k)}\|_2$ when $\boldsymbol{\beta}^{(k)} \neq \widetilde{\boldsymbol{\beta}}^{(k)}$. After updating $\widetilde{\boldsymbol{\beta}}^{(k)}$ using (??), we

have

$$\begin{aligned} L(\widetilde{\boldsymbol{\beta}}^{(k)}(\text{new}) \mid \mathbf{D}) + \lambda w_k \|\widetilde{\boldsymbol{\beta}}^{(k)}(\text{new})\|_2 &\leq Q(\widetilde{\boldsymbol{\beta}}^{(k)}(\text{new}) \mid \mathbf{D}) \\ &\leq Q(\widetilde{\boldsymbol{\beta}} \mid \mathbf{D}) \\ &= L(\widetilde{\boldsymbol{\beta}} \mid \mathbf{D}) + \lambda w_k \|\widetilde{\boldsymbol{\beta}}^{(k)}\|_2. \end{aligned}$$

Moreover, if $\widetilde{\boldsymbol{\beta}}^{(k)}(\text{new}) \neq \widetilde{\boldsymbol{\beta}}^{(k)}$, then the first inequality becomes

$$L(\widetilde{\boldsymbol{\beta}}^{(k)}(\text{new}) \mid \mathbf{D}) + \lambda w_k \|\widetilde{\boldsymbol{\beta}}^{(k)}(\text{new})\|_2 < Q(\widetilde{\boldsymbol{\beta}}^{(k)}(\text{new}) \mid \mathbf{D}).$$

Therefore, the objective function is strictly decreased after updating all groups in a cycle, unless the solution does not change after each groupwise update. If this is the case, we can show that the solution must satisfy the KKT conditions, which means that the algorithm converges and finds the right answer. To see this, if $\widetilde{\boldsymbol{\beta}}^{(k)}(\text{new}) = \widetilde{\boldsymbol{\beta}}^{(k)}$ for all k, then by the update formula (51) we have that for all k

$$\widetilde{\boldsymbol{\beta}}^{(k)} = \frac{1}{\gamma_k}\left(U^{(k)} + \gamma_k \widetilde{\boldsymbol{\beta}}^{(k)}\right)\left(1 - \frac{\lambda w_k}{\|U^{(k)} + \gamma_k \widetilde{\boldsymbol{\beta}}^{(k)}\|_2}\right) \qquad \text{if } \|U^{(k)} + \gamma_k \widetilde{\boldsymbol{\beta}}^{(k)}\|_2 > \lambda w_k, \quad (53)$$

$$\widetilde{\boldsymbol{\beta}}^{(k)} = \mathbf{0} \qquad \text{if } \|U^{(k)} + \gamma_k \widetilde{\boldsymbol{\beta}}^{(k)}\|_2 \leq \lambda w_k. \quad (54)$$

By straightforward algebra we obtain the KKT conditions:

$$\begin{aligned} -U^{(k)} + \lambda w_k \cdot \frac{\widetilde{\boldsymbol{\beta}}^{(k)}}{\|\widetilde{\boldsymbol{\beta}}^{(k)}\|_2} = \mathbf{0} \qquad & \text{if } \widetilde{\boldsymbol{\beta}}^{(k)} \neq \mathbf{0}, \\ \left\|U^{(k)}\right\|_2 \leq \lambda w_k \qquad & \text{if } \widetilde{\boldsymbol{\beta}}^{(k)} = \mathbf{0}, \end{aligned}$$

where $k = 1, 2, \ldots, K$. Therefore, if the objective function stays unchanged after a cycle, the algorithm necessarily converges to the right answer.

5.4 Fitting Options and Algorithms

Recall $\mathbf{K} \in \mathbb{R}^{N_T \times k}$ is the matrix containing the k SNPs used to compute the factored kinship matrix $\mathbf{\Phi}$. The dimension of this matrix will determine the algorithm used as shown in the table below.

Table 1: Algorithm used based on dimension of $\mathbf{K}$.

Dimension of $\mathbf{K}$	lasso	group lasso
$N_T > k$	gcdnet (or degenerate gglasso)	gglasso (GMD Algorithm with weight matrix)
$N_T < k$	glmnet (Coordinate descent with observation weights)	gglasso (GMD Algorithm with observation weights)

6 Simulation Study

To assess the performance of penfam we used genotyped data from the UK Biobank cohort to maintain LD structure. We restricted our simulation study to 1st degree relatives defined by the KING estimate for kinship coefficients. We define the following quantities:

- c: percentage of causal SNPs
- ρ: linkage disequilibrium between two SNPs
- $\mathbf{X}^{(test)}$: $n \times 1000$ matrix of SNPs that have been randomly sampled across the genome, with sampling weights proportional to the size of each chromosome. These are the SNPs that will be included as fixed effects in our model.

- $\mathbf{X}^{(causal)}$: $n \times (c*1000)$ matrix of SNPs out of the SNPs included in the fixed effect model that will be truly associated with the simulated phenotype, where $\mathbf{X}^{(causal)} \subseteq \mathbf{X}^{(test)}$
- $\mathbf{X}^{(other)}$: $n \times 4000$ matrix of SNPs that have been randomly sampled across the genome, with sampling weights proportional to the size of each chromosome. This matrix will be used in the construction of the kinship matrix. Some of these $\mathbf{X}^{(other)}$ SNPs, in conjunction with some of the SNPs in $\mathbf{X}^{(test)}$ will be used in construction of the kinship matrix. We will alter the balance between these two contributors and with the proportion of causal SNPs used to calculate kinship. The maximum LD between any two SNPs in $\mathbf{X}^{(test)}$ and $\mathbf{X}^{(other)}$ will be ρ.
- $\mathbf{X}^{(kinship)}$: $n \times k$ matrix of SNPs used to construct the kinship matrix.
- β_j: effect size for the j^{th} SNP, simulated from a standard normal distribution for $j = 1, \ldots, (c * 1000)$
- $Y^* = \sum_{j=1}^{c*1000} \beta_j \mathbf{X}_j^{(causal)}$
- $Y = Y^* + k \cdot \varepsilon$, where the error term ε is generated from a standard normal distribution, and k is chosen such that the signal-to-noise ratio $SNR = (Var(Y^*)/Var(\varepsilon))$ is 1

We will consider the following simulation scenarios. In each scenario we consider $c = \{0.1, 0.5\}$ and $\rho = \{0.1, 0.5, 0.9\}$:

Scenario 1

All the causal SNPs are included in the calculation of the kinship matrix.

$\mathbf{X}^{(kinship)} = \left[\mathbf{X}^{(other)}; \mathbf{X}^{(causal)}\right]$

Scenario 3

None of the causal SNPs are included in the calculation of the kinship matrix.

$$\mathbf{X}^{(kinship)} = \left[\mathbf{X}^{(other)}\right]$$

6.1 Results

7 Computational Algorithm version 2

We use a general purpose block coordinate descent algorithm (CGD) (38) to solve (15). At each iteration, the algorithm approximates the negative log-likelihood $f(\cdot)$ in $Q_\lambda(\cdot)$ by a strictly convex quadratic function and then applies block coordinate decent to generate a decent direction followed by an inexact line search along this direction (38). For continuously differentiable $f(\cdot)$ and convex and block-separable $P(\cdot)$ (i.e. $P(\boldsymbol{\beta}) = \sum_i P_i(\beta_i)$), (38) show that the solution generated by the CGD method is a stationary point of $Q_\lambda(\cdot)$ if the coordinates are updated in a Gauss-Seidel manner i.e. $Q_\lambda(\cdot)$ is minimized with respect to one parameter while holding all others fixed. The CGD algorithm can thus be run in parallel and therefore suited for large p settings. It has been successfully applied in fixed effects models (e.g. (39), (20)) and (30) for mixed models with an ℓ_1 penalty.

Following (38), the CGD algorithm is given in Algorithm 3.

Algorithm 3: Coordinate Gradient Descent Algorithm

Set the iteration counter $k \leftarrow 0$ and choose initial values for the parameter vector $\boldsymbol{\Theta}^{(0)}$;

repeat

Approximate the Hessian $\nabla^2 f(\boldsymbol{\Theta}^{(k)})$ by a symmetric matix $H^{(k)}$:

$$H^{(k)} = \text{diag}\left[\min\left\{\max\left\{\left[\nabla^2 f(\boldsymbol{\Theta}^{(k)})\right]_{jj}, c_{min}\right\} c_{max}\right\}\right]_{j=1,\ldots,p+1} \tag{55}$$

for $j = 1, \ldots, p+1$ **do**

Solve the descent direction $d^{(k)} := d_{H^{(k)}}(\Theta_j^{(k)})$;

if $\Theta_j^{(k)} \in \{\beta_1, \ldots, \beta_p\}$ **then**

$$d_{H^{(k)}}(\Theta_j^{(k)}) \leftarrow \arg\min_d \left\{\nabla f(\Theta_j^{(k)})d + \frac{1}{2}d^2 H_{jj}^{(k)} + \lambda P(\Theta_j^{(k)} + d)\right\} \tag{56}$$

end

if $\Theta_j^{(k)} \in \{\eta\}$ **then**

$$d_{H^{(k)}}(\Theta_j^{(k)}) \leftarrow -\nabla f(\Theta_j^{(k)})/H_{jj}^{(k)} \tag{57}$$

end

Choose a stepsize;

$$\alpha_j^{(k)} \leftarrow \text{line search given by the Armijo rule}$$

Update;

$$\widehat{\Theta}_j^{(k+1)} \leftarrow \widehat{\Theta}_j^{(k)} + \alpha_j^{(k)} d^{(k)}$$

end

Update;

$$\widehat{\sigma^2}^{(k+1)} \leftarrow \frac{1}{N_T}\sum_{i=1}^{N_T} \frac{([\widetilde{\mathbf{Y}} - \widetilde{\mathbf{X}}\widehat{\boldsymbol{\beta}}^{(k+1)}]_i)^2}{1 + \widehat{\eta}^{(k+1)}(\Lambda_i - 1)} \tag{58}$$

$k \leftarrow k + 1$

until *convergence criterion is satisfied*;

We note that conditional on $\widehat{\boldsymbol{\beta}}$ and $\widehat{\eta}$, there exists an analytic solution for $\widehat{\sigma^2}$:

$$\begin{aligned}\frac{\partial}{\partial\sigma^2}f(\boldsymbol{\Theta}) &= \frac{N_T}{2\sigma^2} - \frac{1}{2\sigma^4}\sum_{i=1}^{N_T}\frac{([\widetilde{\mathbf{Y}} - \widetilde{\mathbf{X}}\boldsymbol{\beta}]_i)^2}{1+\eta(\Lambda_i - 1)} = 0 \\ \widehat{\sigma^2} &= \frac{1}{N_T}\sum_{i=1}^{N_T}\frac{([\widetilde{\mathbf{Y}} - \widetilde{\mathbf{X}}\widehat{\boldsymbol{\beta}}]_i)^2}{1+\widehat{\eta}(\Lambda_i - 1)}\end{aligned} \tag{59}$$

The Armijo rule is defined as follows (38):

Choose $\alpha_{init}^{(k)} > 0$ and let $\alpha^{(k)}$ be the largest element of $\left\{\alpha_{init}^{k}\delta^r\right\}_{r=0,1,2,\ldots}$ satisfying

$$Q_\lambda(\Theta_j^{(k)} + \alpha^{(k)}d^{(k)}) \leq Q_\lambda(\Theta_j^{(k)}) + \alpha^{(k)}\varrho\Delta^{(k)} \tag{60}$$

where $0 < \delta < 1$, $0 < \varrho < 1$, $0 \leq \gamma < 1$ and

$$\Delta^{(k)} := \nabla f(\Theta_j^{(k)})d^{(k)} + \gamma(d^{(k)})^2 H_{jj}^{(k)} + \lambda P(\Theta_j^{(k)} + d^{(k)}) - \lambda P(\Theta^{(k)}) \tag{61}$$

Common choices for the constants are $\delta = 0.1$, $\varrho = 0.001$, $\gamma = 0$, $\alpha_{init}^{(k)} = 1$ for all k (30).

Below we detail the specifics of Algorithm 3 for different penalty functions $P(\boldsymbol{\beta})$.

7.1 ℓ_1 penalty

The objective function is given by

$$Q_\lambda(\boldsymbol{\Theta}) = f(\boldsymbol{\Theta}) + \lambda|\boldsymbol{\beta}| \tag{62}$$

7.1.1 Descent Direction

For simplicity, we remove the iteration counter (k) from the derivation below.

For $\Theta_j^{(k)} \in \{\beta_1, \ldots, \beta_p\}$, let

$$d_H(\Theta_j) = \arg\min_d G(d) \tag{63}$$

where

$$G(d) = \nabla f(\Theta_j)d + \frac{1}{2}d^2 H_{jj} + \lambda|\Theta_j + d|$$

Since $G(d)$ is not differentiable at $-\Theta_j$, we calculate the subdifferential $\partial G(d)$ and search for d with $0 \in \partial G(d)$:

$$\partial G(d) = \nabla f(\Theta_j) + dH_{jj} + \lambda u \tag{64}$$

where

$$u = \begin{cases} 1 & \text{if} \quad d > -\Theta_j \\ -1 & \text{if} \quad d < -\Theta_j \\ [-1, 1] & \text{if} \quad d = \Theta_j \end{cases} \tag{65}$$

We consider each of the three cases in (64) below

1. $d > -\Theta_j$

$$\partial G(d) = \nabla f(\Theta_j) + dH_{jj} + \lambda = 0$$
$$d = \frac{-(\nabla f(\Theta_j) + \lambda)}{H_{jj}}$$

Since $\lambda > 0$ and $H_{jj} > 0$, we have

$$\frac{-(\nabla f(\Theta_j) - \lambda)}{H_{jj}} > \frac{-(\nabla f(\Theta_j) + \lambda)}{H_{jj}} = d \overset{\text{def}}{>} -\Theta_j$$

The solution can be written compactly as

$$d = \text{mid}\left\{\frac{-(\nabla f(\Theta_j) - \lambda)}{H_{jj}}, -\Theta_j, \frac{-(\nabla f(\Theta_j) + \lambda)}{H_{jj}}\right\}$$

where mid $\{a, b, c\}$ denotes the median (mid-point) of a, b, c (38).

2. $d < -\Theta_j$

$$\partial G(d) = \nabla f(\Theta_j) + dH_{jj} - \lambda = 0$$
$$d = \frac{-(\nabla f(\Theta_j) - \lambda)}{H_{jj}}$$

Since $\lambda > 0$ and $H_{jj} > 0$, we have

$$\frac{-(\nabla f(\Theta_j) + \lambda)}{H_{jj}} < \frac{-(\nabla f(\Theta_j) - \lambda)}{H_{jj}} = d \overset{\text{def}}{<} -\Theta_j$$

Again, the solution can be written compactly as

$$d = \text{mid}\left\{\frac{-(\nabla f(\Theta_j) - \lambda)}{H_{jj}}, -\Theta_j, \frac{-(\nabla f(\Theta_j) + \lambda)}{H_{jj}}\right\}$$

3. $d_j = -\Theta_j$

There exists $u \in [-1, 1]$ such that

$$\partial G(d) = \nabla f(\Theta_j) + dH_{jj} + \lambda u = 0$$
$$d = \frac{-(\nabla f(\Theta_j) + \lambda u)}{H_{jj}}$$

For $-1 \leq u \leq 1$, $\lambda > 0$ and $H_{jj} > 0$ we have

$$\frac{-(\nabla f(\Theta_j) + \lambda)}{H_{jj}} \leq d \overset{\text{def}}{=} -\Theta_j \leq \frac{-(\nabla f(\Theta_j) - \lambda)}{H_{jj}}$$

The solution can again be written compactly as

$$d = \text{mid}\left\{\frac{-(\nabla f(\Theta_j) - \lambda)}{H_{jj}}, -\Theta_j, \frac{-(\nabla f(\Theta_j) + \lambda)}{H_{jj}}\right\}$$

We see all three cases lead to the same solution for (63). Therefore the descent direction for $\Theta_j^{(k)} \in \{\beta_1, \ldots, \beta_p\}$ for the ℓ_1 penalty is given by

$$d = \text{mid}\left\{\frac{-(\nabla f(\beta_j) - \lambda)}{H_{jj}}, -\beta_j, \frac{-(\nabla f(\beta_j) + \lambda)}{H_{jj}}\right\} \tag{66}$$

7.1.2 Solution for the β parameter

If the Hessian $\nabla^2 f(\boldsymbol{\Theta}^{(k)}) > 0$ then $H^{(k)}$ defined in (55) is equal to $\nabla^2 f(\boldsymbol{\Theta}^{(k)})$. Using $\alpha_{init} = 1$, the largest element of $\left\{\alpha_{init}^{(k)}\delta^r\right\}_{r=0,1,2,\ldots}$ satisfying the Armijo Rule inequality is reached for $\alpha^{(k)} = \alpha_{init}^{(k)}\delta^0 = 1$. The Armijo rule update for the $\boldsymbol{\beta}$ parameter is then given by

$$\beta_j^{(k+1)} \leftarrow \beta_j^{(k)} + d^{(k)}, \qquad j = 1, \ldots, p \tag{67}$$

Substituting the descent direction given by (66) into (67) we get

$$\beta_j^{(k+1)} = \text{mid}\left\{\beta_j^{(k)} + \frac{-(\nabla f(\beta_j^{(k)}) - \lambda)}{H_{jj}}, 0, \beta_j^{(k)} + \frac{-(\nabla f(\beta_j^{(k)}) + \lambda)}{H_{jj}}\right\} \tag{68}$$

We can further simplify this expression. Let

$$w_i \coloneqq \frac{1}{\sigma^2\left(1 + \eta(\Lambda_i - 1)\right)} \tag{69}$$

.

Re-write the part depending on $\boldsymbol{\beta}$ of the negative log-likelihood in (13) as

$$g(\boldsymbol{\beta}^{(k)}) = \frac{1}{2}\sum_{i=1}^{N_T} w_i \left(\widetilde{Y}_i - \sum_{\ell \neq j} \widetilde{X}_{i\ell}\beta_\ell^{(k)} - \widetilde{X}_{ij}\beta_j^{(k)}\right)^2 \tag{70}$$

The gradient and Hessian are given by

$$\nabla f(\beta_j^{(k)}) \coloneqq \frac{\partial}{\partial \beta_j^{(k)}} g(\boldsymbol{\beta}^{(k)}) = -\sum_{i=1}^{N_T} w_i \widetilde{X}_{ij}\left(\widetilde{Y}_i - \sum_{\ell \neq j} \widetilde{X}_{i\ell}\beta_\ell^{(k)} - \widetilde{X}_{ij}\beta_j^{(k)}\right) \tag{71}$$

$$H_{jj} \coloneqq \frac{\partial^2}{\partial {\beta_j^{(k)}}^2} g(\boldsymbol{\beta}^{(k)}) = \sum_{i=1}^{N_T} w_i \widetilde{X}_{ij}^2 \tag{72}$$

Substituting (71) and (72) into $\beta_j^{(k)} + \frac{-(\nabla f(\beta_j^{(k)}) - \lambda)}{H_{jj}}$

$$\begin{aligned}
&\beta_j^{(k)} + \frac{\sum_{i=1}^{N_T} w_i \widetilde{X}_{ij}\left(\widetilde{Y}_i - \sum_{\ell \neq j} \widetilde{X}_{i\ell}\beta_\ell^{(k)} - \widetilde{X}_{ij}\beta_j^{(k)}\right) + \lambda}{\sum_{i=1}^{N_T} w_i \widetilde{X}_{ij}^2} \\
&= \beta_j^{(k)} + \frac{\sum_{i=1}^{N_T} w_i \widetilde{X}_{ij}\left(\widetilde{Y}_i - \sum_{\ell \neq j} \widetilde{X}_{i\ell}\beta_\ell^{(k)}\right) + \lambda}{\sum_{i=1}^{N_T} w_i \widetilde{X}_{ij}^2} - \frac{\sum_{i=1}^{N_T} w_i \widetilde{X}_{ij}^2 \beta_j^{(k)}}{\sum_{i=1}^{N_T} w_i \widetilde{X}_{ij}^2} \\
&= \frac{\sum_{i=1}^{N_T} w_i \widetilde{X}_{ij}\left(\widetilde{Y}_i - \sum_{\ell \neq j} \widetilde{X}_{i\ell}\beta_\ell^{(k)}\right) + \lambda}{\sum_{i=1}^{N_T} w_i \widetilde{X}_{ij}^2}
\end{aligned} \tag{73}$$

Similarly, substituting (71) and (72) in $\beta_j^{(k)} + \frac{-(\nabla f(\beta_j^{(k)}) + \lambda)}{H_{jj}}$ we get

$$\frac{\sum_{i=1}^{N_T} w_i \widetilde{X}_{ij}\left(\widetilde{Y}_i - \sum_{\ell \neq j} \widetilde{X}_{i\ell}\beta_\ell^{(k)}\right) - \lambda}{\sum_{i=1}^{N_T} w_i \widetilde{X}_{ij}^2} \tag{74}$$

Finally, substituting (73) and (74) into (68) we get

$$\begin{aligned}\beta_j^{(k+1)} &= \text{mid}\left\{\frac{\sum_{i=1}^{N_T} w_i \widetilde{X}_{ij}\left(\widetilde{Y}_i - \sum_{\ell\neq j}\widetilde{X}_{i\ell}\beta_\ell^{(k)}\right) - \lambda}{\sum_{i=1}^{N_T} w_i \widetilde{X}_{ij}^2}, 0, \frac{\sum_{i=1}^{N_T} w_i \widetilde{X}_{ij}\left(\widetilde{Y}_i - \sum_{\ell\neq j}\widetilde{X}_{i\ell}\beta_\ell^{(k)}\right) + \lambda}{\sum_{i=1}^{N_T} w_i \widetilde{X}_{ij}^2}\right\} \\ &= \frac{\mathcal{S}_\lambda\left(\sum_{i=1}^{N_T} w_i \widetilde{X}_{ij}\left(\widetilde{Y}_i - \sum_{\ell\neq j}\widetilde{X}_{i\ell}\beta_\ell^{(k)}\right)\right)}{\sum_{i=1}^{N_T} w_i \widetilde{X}_{ij}^2} \end{aligned} \tag{75}$$

Where $\mathcal{S}_\lambda(x)$ is the soft-thresholding operator

$$\mathcal{S}_\lambda(x) = \text{sign}(x)(|x| - \lambda)_+$$

$\text{sign}(x)$ is the signum function

$$\text{sign}(x) = \begin{cases} -1 & x < 0 \\ 0 & x = 0 \\ 1 & x > 0 \end{cases}$$

and $(x)_+ = \max(x, 0)$.

We note that the parameter update for β_j given by (75) takes the same form as the weighted updates of the `glmnet` algorithm (20) (Section 2.4, equation (10)) with $\alpha = 1$.

References

[1] Teri A Manolio, Francis S Collins, Nancy J Cox, David B Goldstein, Lucia A Hindorff, David J Hunter, Mark I McCarthy, Erin M Ramos, Lon R Cardon, Aravinda Chakravarti, et al. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747, 2009. 2

[2] Jian Yang, Beben Benyamin, Brian P McEvoy, Scott Gordon, Anjali K Henders, Dale R Nyholt, Pamela A Madden, Andrew C Heath, Nicholas G Martin, Grant W Montgomery, et al. Common snps explain a large proportion of the heritability for human height. *Nature genetics*, 42(7):565, 2010. 2

[3] William Astle, David J Balding, et al. Population structure and cryptic relatedness in genetic association studies. *Statistical Science*, 24(4):451–471, 2009. 3

[4] Minsun Song, Wei Hao, and John D Storey. Testing for genetic associations in arbitrarily structured populations. *Nature genetics*, 47(5):550–554, 2015. 3

[5] Jonathan Marchini, Lon R Cardon, Michael S Phillips, and Peter Donnelly. The effects of human population structure on large genetic association studies. *Nature genetics*, 36(5):512, 2004. 3

[6] Clive J Hoggart, John C Whittaker, Maria De Iorio, and David J Balding. Simultaneous analysis of all snps in genome-wide and re-sequencing association studies. *PLoS genetics*, 4(7):e1000130, 2008. 3

[7] Jiahan Li, Kiranmoy Das, Guifang Fu, Runze Li, and Rongling Wu. The bayesian lasso for genome-wide association studies. *Bioinformatics*, 27(4):516–523, 2010. 3

[8] Christoph Lippert, Jennifer Listgarten, Ying Liu, Carl M Kadie, Robert I Davidson, and David Heckerman. Fast linear mixed models for genome-wide association studies. *Nature methods*, 8(10):833–835, 2011. 3, 6, 17

[9] Hyun Min Kang, Jae Hoon Sul, Noah A Zaitlen, Sit-yee Kong, Nelson B Freimer, Chiara Sabatti, Eleazar Eskin, et al. Variance component model to account for sample structure in genome-wide association studies. *Nature genetics*, 42(4):348, 2010. 3

[10] Jianming Yu, Gael Pressoir, William H Briggs, Irie Vroh Bi, Masanori Yamasaki, John F Doebley, Michael D McMullen, Brandon S Gaut, Dahlia M Nielsen, James B Holland, et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature genetics*, 38(2):203, 2006. 3

[11] Jakris Eu-Ahsunthornwattana, E Nancy Miller, Michaela Fakiola, Selma MB Jeronimo, Jenefer M Blackwell, Heather J Cordell, Wellcome Trust Case Control Consortium 2, et al. Comparison of methods to account for relatedness in genome-wide association studies with family-based data. *PLoS Genet*, 10(7):e1004445, 2014. 3

[12] Alkes L Price, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, 38(8):904, 2006. 3

[13] Karim Oualkacha, Zari Dastani, Rui Li, Pablo E Cingolani, Timothy D Spector, Christopher J Hammond, J Brent Richards, Antonio Ciampi, and Celia MT Greenwood. Adjusted sequence kernel association test for rare variants controlling for cryptic and family relatedness. *Genetic epidemiology*, 37(4):366–376, 2013. 3

[14] Heather J Cordell and David G Clayton. A unified stepwise regression procedure for evaluating the relative effects of polymorphisms within a gene using case/control or family data: application to hla in type 1 diabetes. *The American Journal of Human Genetics*, 70(1):124–141, 2002. 3

[15] Barbara Rakitsch, Christoph Lippert, Oliver Stegle, and Karsten Borgwardt. A lasso multi-marker mixed model for association mapping with population structure correction. *Bioinformatics*, 29(2):206–214, 2013. 4

[16] Dong Wang, Kent M Eskridge, and Jose Crossa. Identifying qtls and epistasis in structured plant populations using adaptive mixed lasso. *Journal of agricultural, biological, and environmental statistics*, 16(2):170–184, 2011. 4

[17] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996. 4

[18] Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006. 4

[19] Xiuhua Ding, Shaoyong Su, Kannabiran Nandakumar, Xiaoling Wang, and David W Fardo. A 2-step penalized regression method for family-based next-generation sequencing association studies. In *BMC proceedings*, volume 8, page S25. BioMed Central, 2014. 4

[20] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1, 2010. 4, 12, 14, 27, 34

[21] Yi Yang and Hui Zou. A fast unified algorithm for solving group-lasso penalize learning problems. *Statistics and Computing*, 25(6):1129–1141, 2015. 4

[22] Jian Yang, Noah A Zaitlen, Michael E Goddard, Peter M Visscher, and Alkes L Price. Advantages and pitfalls in the application of mixed-model association methods. *Nature genetics*, 46(2):100, 2014. 4

[23] Naomi Allen, Cathie Sudlow, Paul Downey, Tim Peakman, John Danesh, Paul Elliott, John Gallacher, Jane Green, Paul Matthews, Jill Pell, et al. Uk biobank: Current status and what it means for epidemiology. *Health Policy and Technology*, 1(3):123–126, 2012. 4

[24] Matti Pirinen, Peter Donnelly, Chris CA Spencer, et al. Efficient computation with a linear mixed model on large-scale data sets with applications to genetic studies. *The Annals of Applied Statistics*, 7(1):369–390, 2013. 5, 6, 7

[25] Jan De Leeuw. Block-relaxation algorithms in statistics. In *Information systems and data analysis*, pages 308–324. Springer, 1994. 10

[26] Richard H Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208, 1995. 12

[27] Jon Wakefield. *Bayesian and frequentist regression methods*. Springer Science & Business Media, 2013. 15

[28] Hui Zou, Trevor Hastie, Robert Tibshirani, et al. On the degrees of freedom of the lasso. *The Annals of Statistics*, 35(5):2173–2192, 2007. 16

[29] Howard D Bondell, Arun Krishna, and Sujit K Ghosh. Joint variable selection for fixed and random effects in linear mixed-effects models. *Biometrics*, 66(4):1069–1077, 2010. 16

[30] Jürg Schelldorfer, Peter Bühlmann, GEER DE, and SARA VAN. Estimation for high-dimensional linear mixed-effects models using l1-penalization. *Scandinavian Journal of Statistics*, 38(2):197–214, 2011. 16, 27, 29

[31] Joseph G Ibrahim, Hongtu Zhu, Ramon I Garcia, and Ruixin Guo. Fixed and random effects selection in mixed effects models. *Biometrics*, 67(2):495–503, 2011. 16

[32] Hansheng Wang, Bo Li, and Chenlei Leng. Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3):671–683, 2009. 16

[33] Daniel P Berrar, Werner Dubitzky, Martin Granzow, et al. *A practical approach to microarray data analysis*. Springer, 2003. 17

[34] Gene H Golub and Charles F Van Loan. *Matrix computations*, volume 3. JHU Press, 2012. 18

[35] K. Lange, D. Hunter, and I. Yang. Optimization transfer using sur- rogate objective functions (with discussion). *Journal of Computational and Graphical Statistics*, 9:1–20, 2000. 22

[36] D.R. Hunter and K. Lange. A tutorial on MM algorithms. *The American Statistician*, 58(1):30–37, 2004. 22

[37] T. Wu and K. Lange. The MM alternative to EM. *Statistical Science*, 4:492–505, 2010. 22

[38] Paul Tseng and Sangwoon Yun. A coordinate gradient descent method for nonsmooth separable minimization. *Mathematical Programming*, 117(1):387–423, 2009. 27, 28, 29, 31

[39] Lukas Meier, Sara Van De Geer, and Peter Bühlmann. The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):53–71, 2008. 27

A Algorithm Details

In this section we provide more specific details about the algorithms used to solve th objective function.

A.1 title

B title