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 (13, 22). Another issue with the aforementioned methods is that they first require computing the covariance matrix with a computation time of $\mathcal{O}(n^2k)$ followed by a spectral decomposition of this matrix in $\mathcal{O}(n^3)$ time where k is the number of SNP genotypes used to construct the covariance matrix. 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 called `ggmix` 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

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. Denote the stacked vectors $\mathbf{Y} = (\boldsymbol{y}_i, \ldots, \boldsymbol{y}_N)^T \in \mathbb{R}^{N_T \times 1}$, $\boldsymbol{b} = (\boldsymbol{b}_i, \ldots, \boldsymbol{b}_N)^T \in \mathbb{R}^{N_T \times 1}$, $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}_i, \ldots, \boldsymbol{\varepsilon}_N)^T \in \mathbb{R}^{N_T \times 1}$, and the stacked matrix $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_N)^T \in \mathbb{R}^{N_T \times (p+1)}$. Furthermore, let $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T \in \mathbb{R}^{(p+1)\times 1}$ be a vector of fixed effects regression coefficients corresponding to $\mathbf{X}$. We consider the following linear mixed model with a single random effect (28):

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

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

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

Here, $\Phi_{N_T \times N_T}$ is a known positive semi-definite and symmetric covariance or 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 $\boldsymbol{\varepsilon}$. Furthermore, η is also the narrow-sense heritability (h^2), defined as the proportion of phenotypic variance attributable to the additive genetic factors (1). The joint density of $\mathbf{Y}$ is multivariate normal:

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

The LMM-Lasso method (15) considers an alternative parameterization given by:

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

where $\delta = \sigma_e^2/\sigma_g^2$, σ_g^2 is the genetic variance and σ_e^2 is the residual variance. We instead consider the parameterization in (3) since maximization is easier over the compact set $\eta \in [0, 1]$ than over the unbounded interval $\delta \in [0, \infty)$ (28). We define the complete parameter

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:

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

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

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

function (13). The penalty term is a necessary constraint because in our applications, the sample size is much smaller than the number of predictors. We define the following objective function:

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

where $f(\Theta) := -\ell(\Theta)$ 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. An estimate of the regression parameters $\widehat{\Theta}_\lambda$ is obtained by

Since this would invalidate mixed model assumptions

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

This is the general set-up for our model. In Section 3 we provide more specific details on how we solve (15).

3 Computational Algorithm

All through here can you highlight the most novel parts?

We use a general purpose block coordinate gradient descent algorithm (CGD) (29) to solve (15). At each iteration, we cycle through the coordinates and minimize the objective function with respect to one coordinate only. For continuously differentiable $f(\cdot)$ and convex and block-separable $P(\cdot)$ (i.e. $P(\beta) = \sum_i P_i(\beta_i)$), Tseng and Yun (29) 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 has been successfully applied in fixed effects models (e.g. (30), (20)) and linear mixed models with an ℓ_1 penalty (31). In the next section we provide some brief details about Algorithm 1. A more thorough treatment of the algorithm is given in Appendix A.