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.

lasso (17), elastic net (24) and group lasso (25). `ggmix` also readily handles prior annotation information in the form of a penalty factor, which can be useful for example when dealing with rare variants. We develop a blockwise coordinate descent algorithm which is highly scalable and has theoretical guarantees of convergence to a stationary point. When the matrix of genotypes used to construct the covariance matrix is low rank, there are additional computational speedups that can be implemented. While this has been developed for the univariate case (8), to our knowledge, this hasn't been explored in the multivariable case. The LMM-lasso paper mentions that this is possible but does not provide further details on how this can be implemented in a penalized mixed model framework. In the sequel, we develop a low rank version of the blockwise coordinate descent algorithm which reduces the time complexity from $\mathcal{O}(n^2k)$ to $\mathcal{O}(nk^2)$. All of our algorithms are implemented in the `ggmix` R package hosted on GitHub with extensive documentation (`http://sahirbhatnagar.com/ggmix/`). We provide a brief demonstration of the `ggmix` package in Appendix B.

The rest of the paper is organized as follows. Section 2 describes the `ggmix` model. Section 3 contains the optimization procedure and the algorithm used to fit the `ggmix` model. In Section 4, we compare the performance of our proposed approach and demonstrate the scenarios where it can be advantageous to use over existing methods through simulation studies. Section 5 contains some real data examples and Section 6 discusses some limitations and future directions.

2 Penalized Linear Mixed Models

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 or phenotypes, $\mathbf{X}_i$ an $n_i \times (p+1)$ design matrix (with

vector as $\Theta := (\beta, \eta, \sigma^2)$. The negative log-likelihood for (3) is given by

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

where $\mathbf{V} = \eta\Phi + (1-\eta)\mathbf{I}$ and $\det(\mathbf{V})$ is the determinant of $\mathbf{V}$.

Let $\Phi = \mathbf{U}\mathbf{D}\mathbf{U}^T$ be the eigen (spectral) decomposition of the kinship matrix Φ, 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 (28)

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

This is KEY RIGHT.
i.e. equations (6) - (13)
include the elements
that make sgmix feasible
+ scalable.

If so — play it up!
Say — why no one did it before?
perhaps — "through innovative ~~deconstructed~~ eigen decomposition
we are able to create a scalable
algorithm" ?

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 $\boldsymbol{\Theta} := (\Theta_0, \Theta_1, \ldots, \Theta_{p+1}, \Theta_{p+2}, \Theta_{p+3}) = (\boldsymbol{\beta}, \eta, \sigma^2)$ 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. In light of our goals to select variables associated with the response in high-dimensional data, we consider placing a constraint on the magnitude of the regression coefficients. This can be achieved by adding a penalty term to the likelihood

odd choice

"propose To" ?