

¹ Simultaneous SNP selection and adjustment for
² population structure in high dimensional prediction
³ models

⁴ Sahir R Bhatnagar^{1,2}, Yi Yang⁴, Tianyuan Lu², Erwin Schurr⁶,
⁵ JC Loredo-Osti⁷, Marie Forest², Karim Oualkacha³, and
⁶ Celia MT Greenwood^{1,2,5}

⁷ ¹Department of Epidemiology, Biostatistics and Occupational Health, McGill
⁸ University

⁹ ²Lady Davis Institute, Jewish General Hospital, Montréal, QC

¹⁰ ³Département de Mathématiques, Université de Québec à Montréal

¹¹ ⁴Department of Mathematics and Statistics, McGill University

¹² ⁵Departments of Oncology and Human Genetics, McGill University

¹³ ⁶Department of Medicine, McGill University

¹⁴ ⁷Department of Mathematics and Statistics, Memorial University

¹⁵ March 6, 2020

¹⁶ **Abstract**

¹⁷ Complex traits are known to be influenced by a combination of environmental fac-

tors 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 effects models (LMM) can account for correlations due to relatedness but have not been applicable in high-dimensional (HD) settings where the number of fixed effect predictors greatly exceeds the number of samples. False positives or 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 with a single random effect called `ggmix` for simultaneous SNP selection and adjustment for population structure in high dimensional prediction models. **We develop a blockwise coordinate descent algorithm with automatic tuning parameter selection which is highly scalable, computationally efficient and has theoretical guarantees of convergence.** Through simulations and three real data examples, we show that `ggmix` leads to more parsimonious models compared to the two-stage approach or principal component adjustment with better prediction accuracy. Our method performs well even in the presence of highly correlated markers, and when the causal SNPs are included in the kinship matrix. `ggmix` can be used to construct polygenic risk scores and select instrumental variables in Mendelian randomization studies. Our algorithms are available in an R package (<https://github.com/greenwoodlab/ggmx>).

39 1 Author Summary

40 This work addresses a recurring challenge in the analysis and interpretation of genetic as-
41 sociation studies: which genetic variants can best predict and are independently associated
42 with a given phenotype in the presence of population structure ? Not controlling confound-
43 ing due to geographic population structure, family and/or cryptic relatedness can lead to
44 spurious associations. Much of the existing research has therefore focused on modeling the

45 association between a phenotype and a single genetic variant in a linear mixed model with
46 a random effect. However, this univariate approach may miss true associations due to the
47 stringent significance thresholds required to reduce the number of false positives and also
48 ignores the correlations between markers. We propose an alternative method for fitting
49 high-dimensional multivariable models, which selects SNPs that are independently associ-
50 ated with the phenotype while also accounting for population structure. We provide an
51 efficient implementation of our algorithm and show through simulation studies and real data
52 examples that our method outperforms existing methods in terms of prediction accuracy and
53 controlling the false discovery rate.

54 2 Introduction

55 Genome-wide association studies (GWAS) have become the standard method for analyzing
56 genetic datasets owing to their success in identifying thousands of genetic variants associated
57 with complex diseases (<https://www.genome.gov/gwastudies/>). Despite these impressive
58 findings, the discovered markers have only been able to explain a small proportion of the
59 phenotypic variance; this is known as the missing heritability problem [1]. One plausible
60 reason is that there are many causal variants that each explain a small amount of variation
61 with small effect sizes [2]. Methods such GWAS, which test each variant or single nucleotide
62 polymorphism (SNP) independently, may miss these true associations due to the stringent
63 significance thresholds required to reduce the number of false positives [1]. Another major
64 issue to overcome is that of confounding due to geographic population structure, family
65 and/or cryptic relatedness which can lead to spurious associations [3]. For example, there
66 may be subpopulations within a study that differ with respect to their genotype frequencies
67 at a particular locus due to geographical location or their ancestry. This heterogeneity in
68 genotype frequency can cause correlations with other loci and consequently mimic the signal
69 of association even though there is no biological association [4, 5]. Studies that separate

70 their sample by ethnicity to address this confounding suffer from a loss in statistical power
71 due to the drop in sample size.

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

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

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

92 There has been recent interest in using penalized linear mixed models, which place a con-
93 straint on the magnitude of the effect sizes while controlling for confounding factors such as
94 population structure. For example, the LMM-lasso [15] places a Laplace prior on all main
95 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]. However, methods such as the LMM-lasso are normally performed in two steps. First, the variance components are estimated once from a LMM with a single random effect. These LMMs normally use the estimated covariance matrix from the individuals' genotypes to account for the relatedness but assumes no SNP main effects (i.e. a null model). The residuals from this null model with a single random effect can be treated as independent observations because the relatedness has been effectively removed from the original response. In the second step, these residuals are used as the response in any high-dimensional model that assumes uncorrelated errors. 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].

In this paper we develop a general penalized LMM framework called `ggmix` that simultaneously selects variables and estimates their effects, accounting for between-individual correlations. We develop a blockwise coordinate descent algorithm with automatic tuning parameter selection which is highly scalable, computationally efficient and has theoretical guarantees of convergence. Our method can handle several sparsity inducing penalties such as the lasso [17] and elastic net [23]. Through simulations and three real data examples, we show that `ggmix` leads to more parsimonious models compared to the two-stage approach or principal component adjustment with better prediction accuracy. Our method performs well even in the presence of highly correlated markers, and when the causal SNPs are included in the kinship matrix. All of our algorithms are implemented in the `ggmix` R package hosted on GitHub with extensive documentation (<https://sahirbhatnagar.com/ggmix>). We provide a brief demonstration of the `ggmix` package in Appendix C.

123 The rest of the paper is organized as follows. In Section 3, we compare the performance
124 of our proposed approach and demonstrate the scenarios where it can be advantageous to
125 use over existing methods through simulation studies and three real data analyses. This is
126 followed by a discussion of our results, some limitations and future directions in Section 4.
127 Section 5 describes the `gmmix` model, the optimization procedure and the algorithm used to
128 fit it.

129

3 Results

130 In this section we demonstrate the performance of `gmmix` in a simulation study and three
131 real data applications.

132

3.1 Simulation Study

133 We evaluated the performance of `gmmix` in a variety of simulated scenarios. For each simu-
134 lation scenario we compared `gmmix` to the `lasso` and the `twostep` method. For the `lasso`,
135 we included the top 10 principal components from the simulated genotypes used to calcu-
136 late the kinship matrix as unpenalized predictors in the design matrix. For the `twostep`
137 method, we first fitted an intercept only model with a single random effect using the average
138 information restricted maximum likelihood (AIREML) algorithm [24] as implemented in the
139 `gaston` R package [25]. The residuals from this model were then used as the response in a
140 regular `lasso` model. Note that in the `twostep` method, we removed the kinship effect in
141 the first step and therefore did not need to make any further adjustments when fitting the
142 penalized model. We fitted the `lasso` using the default settings and `standardize=FALSE`
143 in the `glmmnet` package [20], **with 10-fold cross-validation (CV) to select the optimal**
144 **tuning parameter. For other parameters in our simulation study, we defined the**
145 **following quantities:**

- 146 • n : sample size

- 147 • c : percentage of causal SNPs
- 148 • β : true effect size vector of length p
- 149 • $S_0 = \{j; (\beta)_j \neq 0\}$ the index of the true active set with cardinality $|S_0| = c \times p$
- 150 • ***causal*: the list of causal SNP indices**
- 151 • ***kinship*: the list of SNP indices for the kinship matrix**
- 152 • **X : $n \times p$ matrix of SNPs that were included as covariates in the model**

153 We simulated data from the model

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{P} + \boldsymbol{\varepsilon} \quad (1)$$

154 where $\mathbf{P} \sim \mathcal{N}(0, \eta\sigma^2\Phi)$ is the polygenic effect and $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, (1 - \eta)\sigma^2\mathbf{I})$ is the error term.
 155 **Here, $\Phi_{n \times n}$ is the covariance matrix based on the *kinship* SNPs from n individu-**
 156 **als, $\mathbf{I}_{n \times n}$ is the identity matrix and parameters σ^2 and $\eta \in [0, 1]$ determine how the variance**
 157 **is divided between \mathbf{P} and $\boldsymbol{\varepsilon}$. The values of the parameters that we used were as follows:**
 158 **narrow sense heritability $\eta = \{0.1, 0.3\}$, number of covariates $p = 5,000$, number of *kinship***
 159 **SNPs $k = 10,000$, percentage of *causal* SNPs $c = \{0\%, 1\%\}$ and $\sigma^2 = 1$. In addition to**
 160 **these parameters, we also varied the amount of overlap between the *causal* list**
 161 **and the *kinship* list. We considered two main scenarios:**

- 162 1. **None of the *causal* SNPs are included in *kinship* set.**
- 163 2. **All of the *causal* SNPs are included in the *kinship* set.**

164 Both kinship matrices were meant to contrast the model behavior when the causal SNPs are
 165 included in both the main effects and random effects (referred to as proximal contamina-
 166 tion [8]) versus when the causal SNPs are only included in the main effects. These scenarios
 167 are motivated by the current standard of practice in GWAS where the candidate marker
 168 is excluded from the calculation of the kinship matrix [8]. This approach becomes much

more difficult to apply in large-scale multivariable models where there is likely to be overlap between the variables in the design matrix and kinship matrix. We simulated random genotypes from the BN-PSD admixture model with 1D geography and 10 subpopulations using the `bnpssd` package [26, 27]. In Figure 1, we plot the estimated kinship matrix from a single simulated dataset in the form of a heatmap where a darker color indicates a closer genetic relationship.



Figure 1: Example of an empirical kinship matrix used in simulation studies. This scenario models a 1D geography with extensive admixture.

In Figure 2 we plot the first two principal component scores calculated from the simulated genotypes used to calculate the kinship matrix in Figure 1, and color each point by subpopulation membership. We can see that the PCs can identify the subpopulations which is why including them as additional covariates in a regression model has been considered a reasonable approach to control for confounding.

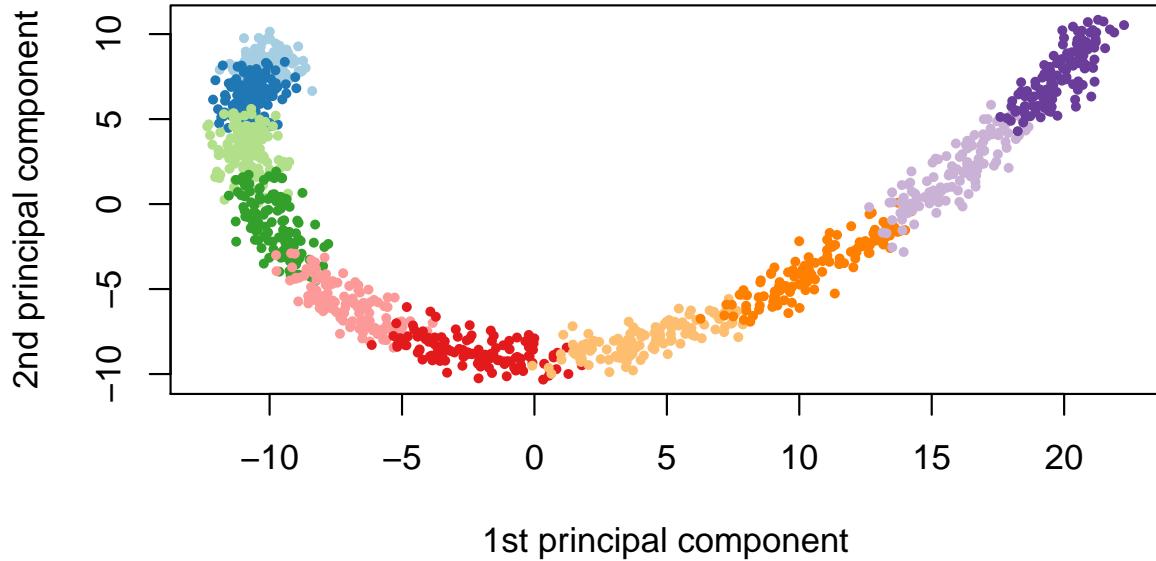


Figure 2: First two principal component scores of the genotype data used to estimate the kinship matrix where each color represents one of the 10 simulated subpopulations.

Using this set-up, we randomly partitioned 1000 simulated observations into 80% for training and 20% for testing. The training set was used to fit the model and select the optimal tuning parameter only, and the resulting model was evaluated on the test set. Let $\hat{\lambda}$ be the estimated value of the optimal regularization parameter, $\hat{\beta}_{\hat{\lambda}}$ the estimate of β at regularization parameter $\hat{\lambda}$, and $\hat{S}_{\hat{\lambda}} = \{j; (\hat{\beta}_{\hat{\lambda}})_j \neq 0\}$ the index of the set of non-zero estimated coefficients. To compare the methods in the context of true positive rate (TPR), we selected the largest tuning parameter that would result in a false positive rate (FPR) closest to 5%, but not more. We also compared the model size ($|\hat{S}_{\hat{\lambda}}|$), test set prediction error based on the refitted unpenalized estimates for each selected model, the estimation error ($\|\hat{\beta} - \beta\|_2^2$), and the variance components (η, σ^2) for the polygenic random effect and error term.

The results are summarized in Table 1. We see that ggmmix outperformed the

192 **twostep** in terms of TPR, and was comparable to the **lasso**. This was the case,
193 regardless of true heritability and whether the causal SNPs were included in the
194 calculation of the kinship matrix. For the **twostep** however, the TPR at a FPR
195 of 5%, drops, on average, from 0.84 (when causal SNPs are not in the kinship)
196 to 0.76 (when causal SNPs are in the kinship). Across all simulation scenarios, **gmmix**
197 had the smallest estimation error, and smallest root mean squared prediction error (RMSE)
198 on the test set while also producing the most parsimonious models. Both the **lasso** and
199 **twostep** selected more false positives, even in the null model scenario. Both the **twostep**
200 and **gmmix** overestimated the heritability though **gmmix** was closer to the true value. When
201 none of the causal SNPs were in the kinship, both methods tended to overestimate the truth
202 when $\eta = 10\%$ and underestimate when $\eta = 30\%$. Across all simulation scenarios **gmmix** was
203 able to (on average) correctly estimate the error variance. The **lasso** tended to overestimate
204 σ^2 in the null model while the **twostep** overestimated σ^2 when none of the causal SNPs were
205 in the kinship matrix.

206 Overall, we observed that variable selection results and RMSE for **gmmix** were similar regard-
207 less of whether the causal SNPs were in the kinship matrix or not. This result is encouraging
208 since in practice the kinship matrix is constructed from a random sample of SNPs across the
209 genome, some of which are likely to be causal, particularly in polygenic traits.

210 In particular, our simulation results show that the principal component adjustment method
211 may not be the best approach to control for confounding by population structure, particularly
212 when variable selection is of interest.

213 3.2 Real Data Applications

214 Three datasets with different features were used to illustrate the potential advantages of
215 **gmmix** over existing approaches such as PC adjustment in a **lasso** regression. In the first
216 two datasets, family structure induced low levels of correlation and sparsity in signals. In

Table 1: Mean (standard deviation) from 200 simulations stratified by the number of causal SNPs (null, 1%), the overlap between causal SNPs and kinship matrix (no overlap, all causal SNPs in kinship), and true heritability (10%, 30%). For all simulations, sample size is $n = 1000$, the number of covariates is $p = 5000$, and the number of SNPs used to estimate the kinship matrix is $k = 10000$. TPR at FPR=5% is the true positive rate at a fixed false positive rate of 5%. Model Size ($|\widehat{S}_{\lambda}|$) is the number of selected variables in the training set using the high-dimensional BIC for `gmmix` and 10-fold cross validation for `lasso` and `twostep`. RMSE is the root mean squared error on the test set. Estimation error is the squared distance between the estimated and true effect sizes. Error variance (σ^2) for `twostep` is estimated from an intercept only LMM with a single random effect and is modeled explicitly in `gmmix`. For the `lasso` we use $\frac{1}{n-|\widehat{S}_{\lambda}|} \|\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{\lambda}\|_2^2$ [28] as an estimator for σ^2 . Heritability (η) for `twostep` is estimated as $\sigma_g^2 / (\sigma_g^2 + \sigma_e^2)$ from an intercept only LMM with a single random effect where σ_g^2 and σ_e^2 are the variance components for the random effect and error term, respectively. η is explicitly modeled in `gmmix`. There is no positive way to calculate η for the `lasso` since we are using a PC adjustment.

Metric	Method	Null model				1% Causal SNPs			
		No overlap		All causal SNPs in kinship		No overlap		All causal SNPs in kinship	
		10%	30%	10%	30%	10%	30%	10%	30%
TPR at FPR=5%	twostep	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.84 (0.05)	0.84 (0.05)	0.76 (0.09)	0.77 (0.08)
	lasso	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.86 (0.05)	0.85 (0.05)	0.86 (0.05)	0.86 (0.05)
	gmmix	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.86 (0.05)	0.86 (0.05)	0.85 (0.05)	0.86 (0.05)
	twostep	0 (0, 5) (289)	0 (0, 2) (287)	0 (0, 5) (388)	0 (0, 2) (385)	328 (250)	332 (329)	284 (319)	284 (253)
	lasso	0 (0, 6) (246)	0 (0, 5) (317)	0 (0, 6) (314)	0 (0, 5) (245)	278 (252)	276 (321)	279 (244)	285 (319)
	gmmix	0 (0, 0) (43)	0 (0, 0) (43)	0 (0, 0) (39)	0 (0, 0) (48)	43 (39, 49) (43)	43 (39, 48) (44)	44 (38, 49) (43)	43 (38, 48) (43)
	twostep	1.02 (0.07)	1.02 (0.06)	1.02 (0.07)	1.02 (0.06)	1.42 (0.10)	1.41 (0.10)	1.44 (0.33)	1.40 (0.22)
	lasso	1.02 (0.06)	1.02 (0.06)	1.02 (0.06)	1.02 (0.06)	1.39 (0.09)	1.38 (0.09)	1.40 (0.08)	1.38 (0.08)
	gmmix	1.00 (0.05)	1.00 (0.05)	1.00 (0.05)	1.00 (0.05)	1.22 (0.10)	1.20 (0.10)	1.23 (0.11)	1.23 (0.12)
Model Size	twostep	0.12 (0.22)	0.09 (0.19)	0.12 (0.22)	0.09 (0.19)	2.97 (0.60)	2.92 (0.60)	3.60 (5.41)	3.21 (3.46)
	lasso	0.13 (0.21)	0.12 (0.22)	0.13 (0.21)	0.12 (0.22)	2.76 (0.46)	2.69 (0.47)	2.82 (0.48)	2.75 (0.48)
	gmmix	0.00 (0.01)	0.01 (0.02)	0.00 (0.01)	0.01 (0.02)	2.11 (1.28)	2.04 (1.22)	2.21 (1.24)	2.28 (1.34)
	twostep	0.87 (0.11)	0.69 (0.15)	0.87 (0.11)	0.69 (0.15)	14.23 (3.53)	14.13 (3.52)	1.42 (1.71)	1.28 (1.66)
	lasso	0.98 (0.05)	0.96 (0.05)	0.98 (0.05)	0.96 (0.05)	1.04 (0.13)	1.02 (0.13)	1.03 (0.14)	1.01 (0.14)
	gmmix	0.85 (0.18)	0.64 (0.20)	0.85 (0.18)	0.64 (0.20)	2.00 (0.49)	1.86 (0.51)	1.06 (0.46)	0.83 (0.45)
	twostep	0.13 (0.11)	0.31 (0.15)	0.13 (0.11)	0.31 (0.15)	0.26 (0.14)	0.26 (0.14)	0.92 (0.08)	0.93 (0.08)
	lasso	—	—	—	—	—	—	—	—
	gmmix	0.15 (0.18)	0.37 (0.21)	0.15 (0.18)	0.37 (0.21)	0.18 (0.16)	0.23 (0.17)	0.59 (0.20)	0.68 (0.19)

Note:

Median (Inter-quartile range) is given for Model Size.

217 the last, a dataset involving mouse crosses, correlations were extremely strong and could
218 confound signals.

219 **3.2.1 UK Biobank**

220 With more than 500,000 participants, the UK Biobank is one of the largest geno-
221 typed health care registries in the world. Among these participants, 147,731
222 have been inferred to be related to at least one individual in this cohort [29].
223 Such a widespread genetic relatedness may confound association studies and
224 bias trait predictions if not properly accounted for. Among these related indi-
225 viduals, 18,150 have a documented familial relationship (parent-offspring, full
226 siblings, second degree or third degree) that was previously inferred in [30]. We
227 attempted to derive a polygenic risk score for height among these individuals.
228 As suggested by a reviewer, the goal of this analysis was to see how the different
229 methods performed for a highly polygenic trait in a set of related individuals.
230 We compared the ggmmix-derived polygenic risk score to those derived by the
231 two step and lasso methods.

232 We first estimated the pairwise kinship coefficient among the 18,150 reportedly
233 related individuals based on 784,256 genotyped SNPs using KING [31]. We
234 grouped related individuals with a kinship coefficient > 0.044 [31] into 8,300
235 pedigrees. We then randomly split the dataset into a training set, a model
236 selection set and a test set of roughly equal sample size, ensuring all individuals
237 in the same pedigree were assigned into the same set. We inverse normalized the
238 standing height after adjusting for age, sex, genotyping array, and assessment
239 center following Yengo et al. [32].

240 To reduce computational complexity, we selected 10,000 SNPs with the largest
241 effect sizes associated with height from a recent large meta-analysis [32]. Among
242 these 10,000 SNPs, 1,233 were genotyped and used for estimating the kinship

whereas the other 8,767 SNPs were imputed based on the Haplotype Reference Consortium reference panel [33]. The distribution of the 10,000 SNPs by chromosome and whether or not the SNP was imputed is shown in Figure B.1 in Supplemental Section B. We see that every chromosome contributed SNPs to the model with 15% coming from chromosome 6. The markers we used are theoretically independent since Yengo et al. performed a COJO analysis which should have tuned down signals due to linkage disequilibrium [32]. We used ggmix, twostep and lasso to select SNPs most predictive of the inverse normalized height on the training set, and chose the λ with the lowest prediction RMSE on the model selection set for each method. We then examined the performance of each derived polygenic risk score on the test set. Similar to Section 3.1, we adjusted for the top 10 genetic PCs as unpenalized predictors when fitting the lasso models, and supplied the kinship matrix based on 784,256 genotyped SNPs to ggmix and twostep.

We found that with a kinship matrix estimated using all genotyped SNPs, ggmix had the possibility to achieve a lower RMSE on the model selection set compared to the twostep and lasso methods (Figure 3A). An optimized ggmix-derived polygenic risk score that utilized the least number of SNPs was also able to better predict the trait with lower RMSE on the test set (Figure 3B).

We additionally applied a Bayesian Sparse Linear Mixed Model (BSLMM) [34] implemented in the GEMMA package [35] to derive a polygenic risk score on the training set. A posterior probability of inclusion of each SNP was provided and prediction was based on all SNPs with a positive posterior probability. We found that although the BSLMM-based polygenic risk score leveraged the most SNPs, it did not achieve a comparable prediction accuracy as the other three methods (Figure 3B). Likely due to the small effect sizes of these SNPs, only

²⁶⁹ 94, 35 and 1 SNPs had a posterior inclusion probability above 0.05, 0.10 and 0.50,
²⁷⁰ respectively. The model would have further reduced prediction accuracy if the
²⁷¹ prediction was based only on these SNPs.

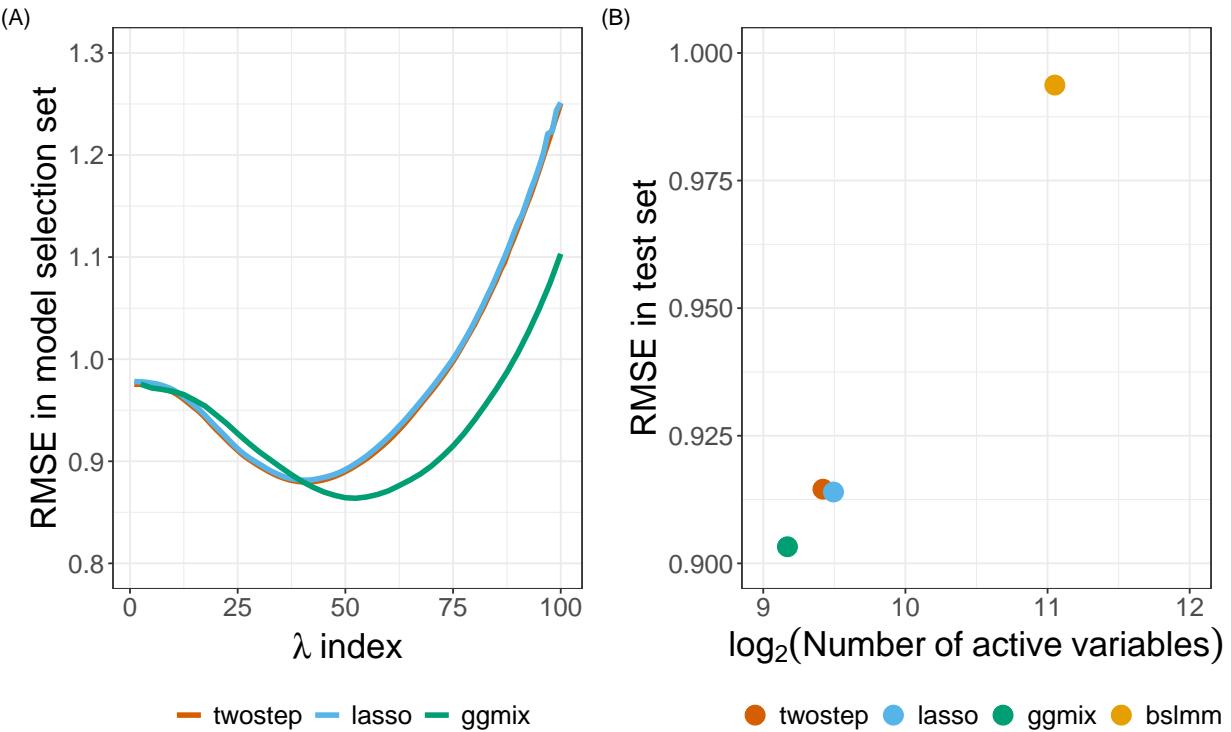


Figure 3: Model selection and testing in the UK Biobank. (A) Root-mean-square error of three methods on the model selection set with respect to a grid search of penalty factor used on the training set. (B) Performance of four methods on the test set with penalty factor optimized on the model selection set. The x-axis has a logarithmic scale. The BSLLMM method optimized coefficients of each SNP through an MCMC process on the training set and was directly evaluated on the test set.

²⁷² 3.2.2 GAW20

²⁷³ In the most recent Genetic Analysis Workshop 20 (GAW20), the causal modeling group in-
²⁷⁴ vestigated causal relationships between DNA methylation (exposure) within some genes and
²⁷⁵ the change in high-density lipoproteins Δ HDL (outcome) using Mendelian Randomization
²⁷⁶ (MR) [36]. Penalized regression methods were used to select SNPs strongly associated with
²⁷⁷ the exposure in order to be used as an instrumental variable (IV) [37, 38]. However, since

278 GAW20 data consisted of families, `twostep` methods were used which could have resulted
279 in a large number of false positives or false negatives. `gmmix` now provides an alternative
280 approach that could be used for selecting the IV while accounting for the family structure
281 of the data.

282 We applied `gmmix` to all 200 GAW20 simulation datasets, each of 679 observations, and
283 compared its performance to the `twostep` and `lasso` methods. Using a Factored Spectrally
284 Transformed Linear Mixed Model (FaST-LMM) [39] adjusted for age and sex, we validated
285 the effect of rs9661059 on blood lipid trait to be significant (genome-wide $p = 6.29 \times 10^{-9}$).
286 Though several other SNPs were also associated with the phenotype, these associations were
287 probably mediated by CpG-SNP interaction pairs and did not reach statistical significance.
288 Therefore, to avoid ambiguity, we only focused on chromosome 1 containing 51,104 SNPs,
289 including rs9661059. Given that population admixture in the GAW20 data was likely, we
290 estimated the population kinship using REAP [40] after decomposing population composi-
291 tions using ADMIXTURE [41]. We used 100,276 LD-pruned whole-genome genotyped SNPs
292 for estimating the kinship. Among these, 8100 were included as covariates in our models
293 based on chromosome 1. The causal SNP was also among the 100,276 SNPs. All methods
294 were fit according to the same settings described in our simulation study in Section 3.1,
295 and adjusting for age and sex. We calculated the median (inter-quartile range) number of
296 active variables, and RMSE (standard deviation) based on five-fold CV on each simulated
297 dataset.

298 On each simulated replicate, we calibrated the methods so that they could be easily compared
299 by fixing the true positive rate to 1 and then minimizing the false positive rate. Hence, the
300 selected SNP, rs9661059, was likely to be the true positive for each method, and non-causal
301 SNPs were excluded to the greatest extent. All three methods precisely chose the correct
302 predictor without any false positives in more than half of the replicates, as the causal signal
303 was strong. However, when some false positives were selected (i.e. when the number of active

variables > 1), `gmmix` performed comparably to `twostep`, while the `lasso` was inclined to select more false positives as suggested by the larger third quartile number of active variables (Table 2). We also observed that `gmmix` outperformed the `twostep` method with lower CV RMSE using the same number of SNPs. Meanwhile, it achieved roughly the same prediction accuracy as `lasso` but with fewer non-causal SNPs (Table 2). It is also worth mentioning that there was very little correlation between the causal SNP and SNPs within a 1Mb-window around it (Figure B.2 in Supplemental Section B.2), making it an ideal scenario for the `lasso` and related methods.

We also applied the BSLMM method by performing five-fold CV on each of the 200 simulated replicates. We found that while BSLMM achieved a lower CV RMSE compared to the other methods (Table 2), this higher prediction accuracy relied on approximately 80% of the 51,104 SNPs with a positive posterior probability. This may suggest overfitting in this dataset. On the other hand, we found that imposing a posterior inclusion probability threshold might improve feature selection, but the results were sensitive to the arbitrary threshold as most SNPs had a low posterior probability. It is also noteworthy that we did not adjust for age and sex in the BSLMM model, as the current implementation of the method in the GEMMA package does not allow adjustment for covariates.

Table 2: Summary of model performance based on 200 GAW20 simulations. Five-fold cross-validation root-mean-square error was reported for each simulation replicate. Prediction accuracy for BSLMM model with a positive posterior inclusion probability threshold was not reported because the model could contain no SNP.

Method	Median number of active variables (Inter-quartile range)	RMSE (SD)
twostep	1 (1 - 11)	0.3604 (0.0242)
lasso	1 (1 - 15)	0.3105 (0.0199)
gmmix	1 (1 - 12)	0.3146 (0.0210)
BSLMM	40,737 (39,901 - 41,539)	0.2503 (0.0099)
BSLMM, 322 posterior probability	2 (1 - 4) > 0.05	
BSLMM, posterior probability	0 (0 - 1) > 0.10	
BSLMM, posterior probability	0 (0 - 0) > 0.50	

323 3.2.3 Mouse Crosses and Sensitivity to Mycobacterial Infection

- 324 Mouse inbred strains of genetically identical individuals are extensively used in research.
 325 Crosses of different inbred strains are useful for various studies of heritability focusing on
 326 either observable phenotypes or molecular mechanisms, and in particular, recombinant con-

327 genic strains have been an extremely useful resource for many years [42]. However, ignor-
328 ing complex genetic relationships in association studies can lead to inflated false positives
329 in genetic association studies when different inbred strains and their crosses are investi-
330 gated [43, 44, 45]. Therefore, a previous study developed and implemented a mixed model
331 to find loci associated with mouse sensitivity to mycobacterial infection [46]. The random
332 effects in the model captured complex correlations between the recombinant congenic mouse
333 strains based on the proportion of the DNA shared identical by descent. Through a se-
334 ries of mixed model fits at each marker, new loci that impact growth of mycobacteria on
335 chromosome 1 and chromosome 11 were identified.

336 Here we show that `gmmix` can identify these loci, as well as potentially others, in a single
337 analysis. We reanalyzed the growth permissiveness in the spleen, as measured by colony
338 forming units (CFUs), 6 weeks after infection from *Mycobacterium bovis* Bacille Calmette-
339 Guerin (BCG) Russia strain as reported in [46].

340 By taking the consensus between the “main model” and the “conditional model” of the original
341 study, we regarded markers D1Mit435 on chromosome 1 and D11Mit119 on chromosome 11
342 as two true positive loci. We directly estimated the kinship between mice using genotypes at
343 625 microsatellite markers. The estimated kinship entered directly into `gmmix` and `twostep`.
344 For the `lasso`, we calculated and included the first 10 principal components of the estimated
345 kinship. To evaluate the robustness of different models, we bootstrapped the 189-sample
346 dataset and repeated the analysis 200 times. **We then conceived a two-fold criteria**
347 **to evaluate performance of each model.** We first examined whether a model
348 could pick up both true positive loci using some λ . If the model failed to pick
349 up both loci simultaneously with any λ , we counted as modeling failure on the
350 corresponding bootstrap replicate; otherwise, we counted as modeling success
351 and recorded which other loci were picked up given the largest λ . Consequently,
352 similar to the strategy used in the GAW20 analysis, we optimized the models by

353 tuning the penalty factor such that these two true positive loci were picked up,
354 while the number of other active loci was minimized. Significant markers were defined
355 as those captured in at least half of the successful bootstrap replicates (Figure 4).

356 We demonstrated that `gmmix` recognized the true associations more robustly than `twostep`
357 and `lasso`. In almost all (99%) bootstrap replicates, `gmmix` was able to capture both true
358 positives, while the `twostep` failed in 19% of the replicates and the `lasso` failed in 56% of
359 the replicates by missing at least one of the two true positives (Figure 4). **The robustness**
360 **of gmmix is particularly noteworthy due to the strong correlations between all**
361 **microsatellite markers in this dataset (Figure B.3 in Supplemental Section B.2).**
362 **These strong correlations with the causal markers, partially explain the poor**
363 **performance of the lasso as it suffers from unstable selections in the presence of**
364 **correlated variables (e.g. [47]).**

365 We also identified several other loci that might also be associated with susceptibility to my-
366 cobacterial infection (Table 3). Among these new potentially-associated markers, D2Mit156
367 was found to play a role in control of parasite numbers of *Leishmania tropica* in lymph
368 nodes [48]. An earlier study identified a parent-of-origin effect at D17Mit221 on CD4M
369 levels [49]. This effect was more visible in crosses than in parental strains. In addition,
370 D14Mit131, selected only by `gmmix`, was found to have a 9% loss of heterozygosity in hy-
371 brids of two inbred mouse strains [50], indicating the potential presence of putative suppressor
372 genes pertaining to immune surveillance and tumor progression [51]. This result might also
373 suggest association with anti-bacterial responses yet to be discovered.

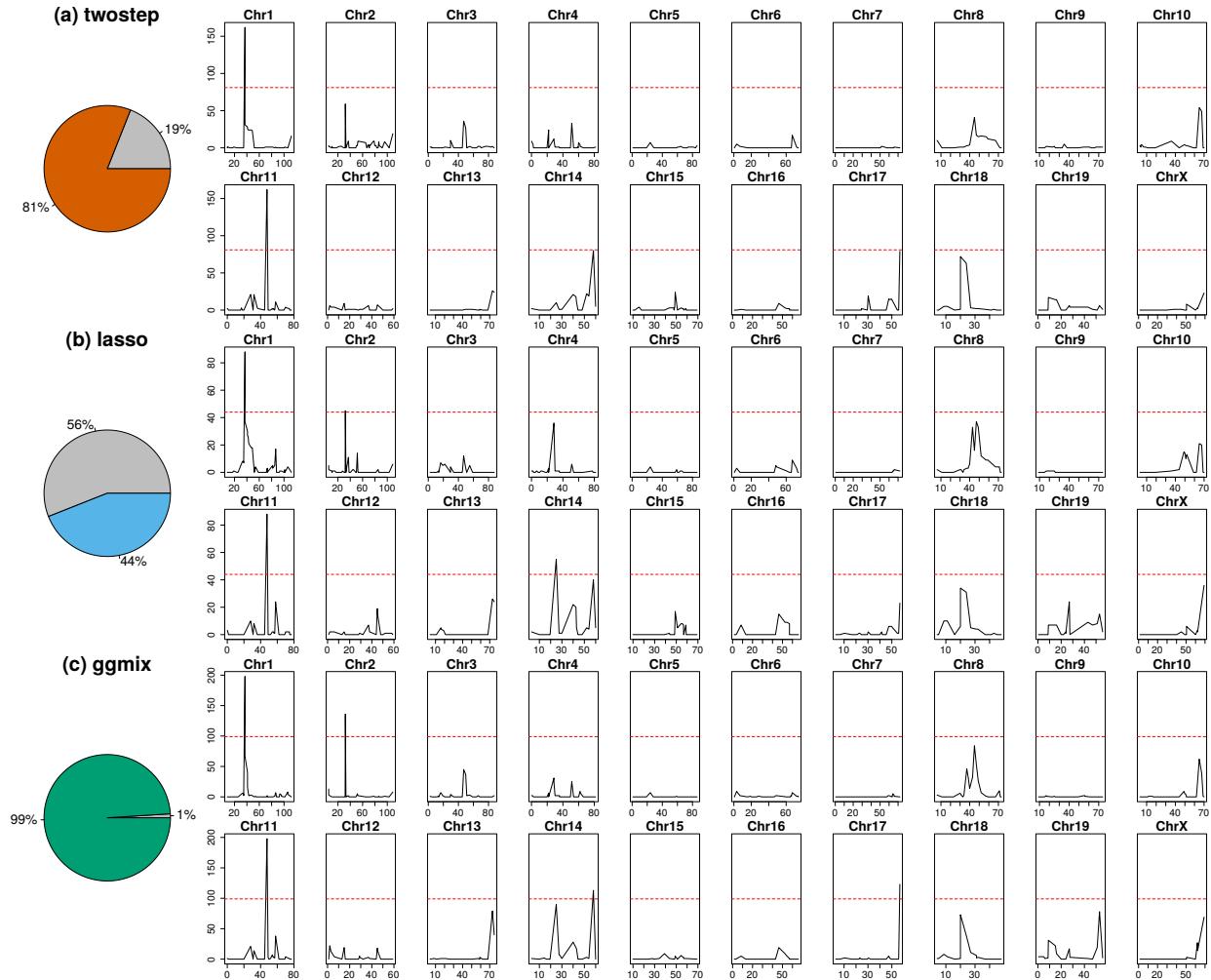


Figure 4: Comparison of model performance on the mouse cross data. Pie charts depict model robustness where grey areas denote bootstrap replicates on which the corresponding model is unable to capture both true positives using any penalty factor, whereas colored areas denote successful replicates. Chromosome-based signals record in how many successful replicates the corresponding loci are picked up by the corresponding optimized model. Red dashed lines delineate significance thresholds.

Table 3: Additional loci significantly associated with mouse susceptibility to myobacterial infection, after excluding two true positives. Loci needed to be identified in at least 50% of the successful bootstrap replicates that captured both true positive loci.

Method	Marker	Position in cM	Position in bp
twostep	N/A	N/A	N/A
³⁷⁴ lasso	D2Mit156	Chr2:31.66	Chr2:57081653-57081799
	D14Mit155	Chr14:31.52	Chr14:59828398-59828596
³⁷⁵ gmmix	D2Mit156	Chr2:31.66	Chr2:57081653-57081799
	D14Mit131	Chr14:63.59	Chr14:120006565-120006669
	D17Mit221	Chr17:59.77	Chr17:90087704-90087842

4 Discussion

³⁷⁶ We have developed a general penalized LMM framework called **gmmix** which simultaneously
³⁷⁷ selects SNPs and adjusts for population structure in high dimensional prediction models.
³⁷⁸ **We compared our method to the twostage procedure, where in the first stage,**
³⁷⁹ **the dependence between observations is adjusted for in a LMM with a single**
³⁸⁰ **random effect and no covariates (i.e. null model).** The residuals from this null
³⁸¹ **model can then be used in any model for independent observations because the**
³⁸² **relatedness has been effectively removed from the original response.** We also
³⁸³ **compared our method to the lasso and BSLMM which are closely related to gmmix**
³⁸⁴ **since they also jointly model the relatedness and SNPs in a single step.** The key
³⁸⁵ **differences are that the lasso uses a principal component adjustment and BSLMM**
³⁸⁶ **is a Bayesian method focused on phenotype prediction.**

³⁸⁷ Through an extensive simulation study and three real data analyses that mimic many ex-
³⁸⁸ perimental designs in genetics, we show that the current approaches of PC adjustment and
³⁸⁹ two-stage procedures are not necessarily sufficient to control for confounding by population

390 structure leading to a high number of false positives. Our simulation results show that `gmmix`
391 outperforms existing methods in terms of sparsity and prediction error even when the causal
392 variants are included in the kinship matrix (Table 1). Many methods for single-SNP analyses
393 avoid this proximal contamination [8] by using a leave-one-chromosome-out scheme [52], i.e.,
394 construct the kinship matrix using all chromosomes except the one on which the marker
395 being tested is located. **However, this approach is not possible if we want to model**
396 **many SNPs (across many chromosomes) jointly to create, for example, a poly-**
397 **genic risk score. For the purposes of variable selection, we would also want to**
398 **model all chromosomes together since the power to detect an association for a**
399 **given SNP may be increased when other causal SNPs have been accounted for.**
400 Conversely, a stronger signal from a causal SNP may weaken false signals when
401 modeled jointly [6], particularly when the markers are highly correlated as in
402 the mouse crosses example.

403 In the UK Biobank, we found that with a kinship matrix estimated using all
404 genotyped SNPs, `gmmix` had achieved a lower RMSE on the model selection set
405 compared to the `twostep` and `lasso` methods. Furthermore, an optimized `gmmix`-
406 derived polygenic risk score that utilized the least number of SNPs was also able
407 to better predict the trait with lower RMSE on the test set. In the GAW20 example,
408 we showed that while all methods were able to select the strongest causal SNP, `gmmix` did
409 so with the least amount of false positives while also maintaining good predictive ability.
410 In the mouse crosses example, we showed that `gmmix` is robust to perturbations in the data
411 using a bootstrap analysis. Indeed, `gmmix` was able to consistently select the true positives
412 across bootstrap replicates, while `twostep` failed in 19% of the replicates and `lasso` failed
413 in 56% of the replicates by missing of at least one of the two true positives. Our re-analysis
414 of the data also lead to some potentially new findings, not found by existing methods, that
415 may warrant further study. This particular example had many markers that were
416 strongly correlated with each other (Figure B.3 of Supplemental Section B.2).

417 Nevertheless, we observed that the two true positive loci were the most often
418 selected while none of the nearby markers were picked up in more than 50% of
419 the 200 bootstrap replicates. This shows that our method does recognize the
420 true positives in the presence of highly correlated markers. Nevertheless, we
421 think the issue of variable selection for correlated SNPs warrants further study.
422 The recently proposed Precision Lasso [47] seeks to address this problem in the
423 high-dimensional fixed effects model.

424 We emphasize here that previously developed methods such as the LMM-lasso [15] use a two-
425 stage fitting procedure without any convergence details. From a practical point of view, there
426 is currently no implementation that provides a principled way of determining the sequence
427 of tuning parameters to fit, nor a procedure that automatically selects the optimal value of
428 the tuning parameter. To our knowledge, we are the first to develop a coordinate gradient
429 descent (CGD) algorithm in the specific context of fitting a penalized LMM for population
430 structure correction with theoretical guarantees of convergence. Furthermore, we develop
431 a principled method for automatic tuning parameter selection and provide an easy-to-use
432 software implementation in order to promote wider uptake of these more complex methods
433 by applied practitioners.

434 Although we derive a CGD algorithm for the ℓ_1 penalty, our approach can also be easily
435 extended to other penalties such as the elastic net and group lasso with the same guarantees
436 of convergence. A limitation of `ggmix` is that it first requires computing the covariance ma-
437 trix with a computation time of $\mathcal{O}(n^2k)$ followed by a spectral decomposition of this matrix
438 in $\mathcal{O}(n^3)$ time where k is the number of SNP genotypes used to construct the covariance
439 matrix. This computation becomes prohibitive for large cohorts such as the UK Biobank [53]
440 which have collected genetic information on half a million individuals. When the matrix of
441 genotypes used to construct the covariance matrix is low rank, there are additional computa-
442 tional speedups that can be implemented. While this has been developed for the univariate

443 case [8], to our knowledge, this has not been explored in the multivariable case. We are cur-
444 rently developing a low rank version of the penalized LMM developed here, which reduces
445 the time complexity from $\mathcal{O}(n^2k)$ to $\mathcal{O}(nk^2)$. There is also the issue of how our model
446 scales with an increasing number of covariates (p). Due to the coordinate-wise
447 optimization procedure, we expect this to be less of an issue, but still prohibitive
448 for $p > 1e5$. The `biglasso` package [54] uses memory mapping strategies for large
449 p , and this is something we are exploring for `gmmix`.

450 As was brought up by a reviewer, the simulations and real data analyses pre-
451 sented here contained many more markers used to estimate the kinship than
452 the sample size ($n/k \leq 0.1$). In the single locus association test, Yang et al. [22]
453 found that proximal contamination was an issue when $n/k \approx 1$. We believe fur-
454 ther theoretical study is needed to see if these results can be generalized to the
455 multivariable models being fit here. Once the computational limitations of sam-
456 ple size mentioned above have been addressed, these theoretical results can be
457 supported by simulation studies.

458 There are other applications in which our method could be used as well. For example, there
459 has been a renewed interest in polygenic risk scores (PRS) which aim to predict complex
460 diseases from genotypes. `gmmix` could be used to build a PRS with the distinct advantage
461 of modeling SNPs jointly, allowing for main effects as well as interactions to be accounted
462 for. Based on our results, `gmmix` has the potential to produce more robust and parsimonious
463 models than the `lasso` with better predictive accuracy. Our method is also suitable for fine
464 mapping SNP association signals in genomic regions, where the goal is to pinpoint individual
465 variants most likely to impact the underlying biological mechanisms of disease [55].

466 5 Materials and Methods

467 5.1 Model Set-up

468 Let $i = 1, \dots, N$ be a grouping index, $j = 1, \dots, n_i$ the observation index within a group
 469 and $N_T = \sum_{i=1}^N n_i$ the total number of observations. For each group let $\mathbf{y}_i = (y_1, \dots, y_{n_i})$ be
 470 the observed vector of responses or phenotypes, \mathbf{X}_i an $n_i \times (p + 1)$ design matrix (with
 471 the column of 1s for the intercept), \mathbf{b}_i a group-specific random effect vector of length
 472 n_i and $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \dots, \varepsilon_{in_i})$ the individual error terms. Denote the stacked vectors $\mathbf{Y} =$
 473 $(\mathbf{y}_1, \dots, \mathbf{y}_N)^T \in \mathbb{R}^{N_T \times 1}$, $\mathbf{b} = (\mathbf{b}_1, \dots, \mathbf{b}_N)^T \in \mathbb{R}^{N_T \times 1}$, $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}_1, \dots, \boldsymbol{\varepsilon}_N)^T \in \mathbb{R}^{N_T \times 1}$, and the
 474 stacked matrix
 475 $\mathbf{X} = (\mathbf{X}_1^T, \dots, \mathbf{X}_N^T) \in \mathbb{R}^{N_T \times (p+1)}$. Furthermore, let $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)^T \in \mathbb{R}^{(p+1) \times 1}$ be a vec-
 476 tor of fixed effects regression coefficients corresponding to \mathbf{X} . We consider the following
 477 linear mixed model with a single random effect [56]:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{b} + \boldsymbol{\varepsilon} \quad (2)$$

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

$$\mathbf{b} \sim \mathcal{N}(0, \eta\sigma^2\boldsymbol{\Phi}) \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(0, (1 - \eta)\sigma^2\mathbf{I}) \quad (3)$$

479 Here, $\boldsymbol{\Phi}_{N_T \times N_T}$ is a known positive semi-definite and symmetric covariance or kinship ma-
 480 trix calculated from SNPs sampled across the genome, $\mathbf{I}_{N_T \times N_T}$ is the identity matrix and
 481 parameters σ^2 and $\eta \in [0, 1]$ determine how the variance is divided between \mathbf{b} and $\boldsymbol{\varepsilon}$. Note
 482 that η is also the narrow-sense heritability (h^2), defined as the proportion of phenotypic
 483 variance attributable to the additive genetic factors [1]. The joint density of \mathbf{Y} is therefore
 484 multivariate normal:

$$\mathbf{Y}|(\boldsymbol{\beta}, \eta, \sigma^2) \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \eta\sigma^2\boldsymbol{\Phi} + (1 - \eta)\sigma^2\mathbf{I}) \quad (4)$$

485 The LMM-Lasso method [15] considers an alternative but equivalent parameterization given
486 by:

$$\mathbf{Y}|(\boldsymbol{\beta}, \delta, \sigma_g^2) \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma_g^2(\boldsymbol{\Phi} + \delta\mathbf{I})) \quad (5)$$

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 (4) since maximization is easier over the compact set $\eta \in [0, 1]$ than over the unbounded interval $\delta \in [0, \infty)$ [56]. We define the complete parameter vector as $\boldsymbol{\Theta} := (\boldsymbol{\beta}, \eta, \sigma^2)$. The negative log-likelihood for (4) is given by

$$-\ell(\boldsymbol{\Theta}) \propto \frac{N_T}{2} \log(\sigma^2) + \frac{1}{2} \log(\det(\mathbf{V})) + \frac{1}{2\sigma^2} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T \mathbf{V}^{-1} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \quad (6)$$

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

Let $\boldsymbol{\Phi} = \mathbf{U}\mathbf{D}\mathbf{U}^T$ be the eigen (spectral) decomposition of the kinship matrix $\boldsymbol{\Phi}$, 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 [56]

$$\begin{aligned} \mathbf{V} &= \eta\boldsymbol{\Phi} + (1 - \eta)\mathbf{I} \\ &= \eta\mathbf{U}\mathbf{D}\mathbf{U}^T + (1 - \eta)\mathbf{U}\mathbf{I}\mathbf{U}^T \\ &= \mathbf{U}\eta\mathbf{D}\mathbf{U}^T + \mathbf{U}(1 - \eta)\mathbf{I}\mathbf{U}^T \\ &= \mathbf{U}(\eta\mathbf{D} + (1 - \eta)\mathbf{I})\mathbf{U}^T \\ &= \mathbf{U}\tilde{\mathbf{D}}\mathbf{U}^T \end{aligned} \quad (7)$$

where

$$\tilde{\mathbf{D}} = \eta \mathbf{D} + (1 - \eta) \mathbf{I} \quad (8)$$

$$\begin{aligned} &= \eta \begin{bmatrix} \Lambda_1 & & & \\ & \Lambda_2 & & \\ & & \ddots & \\ & & & \Lambda_{N_T} \end{bmatrix} + (1 - \eta) \begin{bmatrix} 1 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{bmatrix} \\ &= \begin{bmatrix} 1 + \eta(\Lambda_1 - 1) & & & \\ & 1 + \eta(\Lambda_2 - 1) & & \\ & & \ddots & \\ & & & 1 + \eta(\Lambda_{N_T} - 1) \end{bmatrix} \\ &= \text{diag}\{1 + \eta(\Lambda_1 - 1), 1 + \eta(\Lambda_2 - 1), \dots, 1 + \eta(\Lambda_{N_T} - 1)\} \end{aligned} \quad (9)$$

Since (8) is a diagonal matrix, its inverse is also a diagonal matrix:

$$\tilde{\mathbf{D}}^{-1} = \text{diag} \left\{ \frac{1}{1 + \eta(\Lambda_1 - 1)}, \frac{1}{1 + \eta(\Lambda_2 - 1)}, \dots, \frac{1}{1 + \eta(\Lambda_{N_T} - 1)} \right\} \quad (10)$$

From (7) and (9), $\log(\det(\mathbf{V}))$ simplifies to

$$\begin{aligned} \log(\det(\mathbf{V})) &= \log \left(\det(\mathbf{U}) \det(\tilde{\mathbf{D}}) \det(\mathbf{U}^T) \right) \\ &= \log \left\{ \prod_{i=1}^{N_T} (1 + \eta(\Lambda_i - 1)) \right\} \\ &= \sum_{i=1}^{N_T} \log(1 + \eta(\Lambda_i - 1)) \end{aligned} \quad (11)$$

since $\det(\mathbf{U}) = 1$. It also follows from (7) that

$$\begin{aligned}\mathbf{V}^{-1} &= \left(\mathbf{U} \tilde{\mathbf{D}} \mathbf{U}^T \right)^{-1} \\ &= (\mathbf{U}^T)^{-1} \left(\tilde{\mathbf{D}} \right)^{-1} \mathbf{U}^{-1} \\ &= \mathbf{U} \tilde{\mathbf{D}}^{-1} \mathbf{U}^T\end{aligned}\tag{12}$$

since for an orthonormal matrix $\mathbf{U}^{-1} = \mathbf{U}^T$. Substituting (10), (11) and (12) into (6) the negative log-likelihood becomes

$$\begin{aligned}-\ell(\boldsymbol{\Theta}) &\propto \frac{N_T}{2} \log(\sigma^2) + \frac{1}{2} \sum_{i=1}^{N_T} \log(1 + \eta(\Lambda_i - 1)) + \frac{1}{2\sigma^2} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T \mathbf{U} \tilde{\mathbf{D}}^{-1} \mathbf{U}^T (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \\ &\quad (13)\end{aligned}$$

$$\begin{aligned}&= \frac{N_T}{2} \log(\sigma^2) + \frac{1}{2} \sum_{i=1}^{N_T} \log(1 + \eta(\Lambda_i - 1)) + \frac{1}{2\sigma^2} (\mathbf{U}^T \mathbf{Y} - \mathbf{U}^T \mathbf{X}\boldsymbol{\beta})^T \tilde{\mathbf{D}}^{-1} (\mathbf{U}^T \mathbf{Y} - \mathbf{U}^T \mathbf{X}\boldsymbol{\beta}) \\ &= \frac{N_T}{2} \log(\sigma^2) + \frac{1}{2} \sum_{i=1}^{N_T} \log(1 + \eta(\Lambda_i - 1)) + \frac{1}{2\sigma^2} (\tilde{\mathbf{Y}} - \tilde{\mathbf{X}}\boldsymbol{\beta})^T \tilde{\mathbf{D}}^{-1} (\tilde{\mathbf{Y}} - \tilde{\mathbf{X}}\boldsymbol{\beta}) \\ &= \frac{N_T}{2} \log(\sigma^2) + \frac{1}{2} \sum_{i=1}^{N_T} \log(1 + \eta(\Lambda_i - 1)) + \frac{1}{2\sigma^2} \sum_{i=1}^{N_T} \frac{\left(\tilde{Y}_i - \sum_{j=0}^p \tilde{X}_{ij+1} \beta_j \right)^2}{1 + \eta(\Lambda_i - 1)}\end{aligned}\tag{14}$$

where $\tilde{\mathbf{Y}} = \mathbf{U}^T \mathbf{Y}$, $\tilde{\mathbf{X}} = \mathbf{U}^T \mathbf{X}$, \tilde{Y}_i denotes the i^{th} element of $\tilde{\mathbf{Y}}$, \tilde{X}_{ij} is the i, j^{th} entry of $\tilde{\mathbf{X}}$

and $\mathbf{1}$ is a column vector of N_T ones.

5.2 Penalized Maximum Likelihood Estimator

We define the $p + 3$ length vector of parameters $\boldsymbol{\Theta} := (\Theta_0, \Theta_1, \dots, \Theta_{p+1}, \Theta_{p+2}, \Theta_{p+3}) = (\boldsymbol{\beta}, \eta, \sigma^2)$ 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 propose to place a constraint on the magnitude of the regression coefficients. This can be achieved by adding a penalty term to the likelihood function (14). The penalty term is a necessary constraint because in our applications, the

497 sample size is much smaller than the number of predictors. We define the following objective
498 function:

$$Q_\lambda(\Theta) = f(\Theta) + \lambda \sum_{j \neq 0} v_j P_j(\beta_j) \quad (15)$$

499 where $f(\Theta) := -\ell(\Theta)$ is defined in (14), $P_j(\cdot)$ is a penalty term on the fixed regression
500 coefficients $\beta_1, \dots, \beta_{p+1}$ (we do not penalize the intercept) controlled by the nonnegative
501 regularization parameter λ , and v_j is the penalty factor for j th covariate. These penalty
502 factors serve as a way of allowing parameters to be penalized differently. Note that we do
503 not penalize η or σ^2 . An estimate of the regression parameters $\widehat{\Theta}_\lambda$ is obtained by

$$\widehat{\Theta}_\lambda = \arg \min_{\Theta} Q_\lambda(\Theta) \quad (16)$$

504 This is the general set-up for our model. In Section 5.3 we provide more specific details on
505 how we solve (16). **We note here that the main difference between the proposed**
506 **model, and the `lmmlasso` [57], is that we are limiting ourselves to a single unpenal-**
507 **ized random effect.** Another key difference is that we rotate the response vector
508 Y and the design matrix X by the eigen vectors of the kinship matrix, resulting
509 in a diagonal covariance matrix which greatly speeds up the computation.

510 5.3 Computational Algorithm

511 We use a general purpose block coordinate gradient descent algorithm (CGD) [58] to solve (16).
512 At each iteration, we cycle through the coordinates and minimize the objective function with
513 respect to one coordinate only. For continuously differentiable $f(\cdot)$ and convex and block-
514 separable $P(\cdot)$ (i.e. $P(\beta) = \sum_i P_i(\beta_i)$), Tseng and Yun [58] show that the solution generated
515 by the CGD method is a stationary point of $Q_\lambda(\cdot)$ if the coordinates are updated in a
516 Gauss-Seidel manner i.e. $Q_\lambda(\cdot)$ is minimized with respect to one parameter while holding
517 all others fixed. The CGD algorithm has been successfully applied in fixed effects models
518 (e.g. [59], [20]) and linear mixed models with an ℓ_1 penalty [57]. In the next section we

519 provide some brief details about Algorithm 1. A more thorough treatment of the algorithm
 520 is given in Appendix A.

Algorithm 1: Block Coordinate Gradient Descent

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

for $\lambda \in \{\lambda_{\max}, \dots, \lambda_{\min}\}$ **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: $\|\Theta^{(k+1)} - \Theta^{(k)}\|_2 < \epsilon$;

end

521 **5.3.1 Updates for the β parameter**

522 Recall that the part of the objective function that depends on β has the form

$$Q_\lambda(\Theta) = \frac{1}{2} \sum_{i=1}^{N_T} w_i \left(\tilde{Y}_i - \sum_{j=0}^p \tilde{X}_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^p v_j |\beta_j| \quad (17)$$

523 where

$$w_i := \frac{1}{\sigma^2 (1 + \eta(\Lambda_i - 1))} \quad (18)$$

Conditional on $\eta^{(k)}$ and $\sigma^2^{(k)}$, it can be shown that the solution for β_j , $j = 1, \dots, p$ is given by

$$\beta_j^{(k+1)} \leftarrow \frac{\mathcal{S}_\lambda \left(\sum_{i=1}^{N_T} w_i \tilde{X}_{ij} \left(\tilde{Y}_i - \sum_{\ell \neq j} \tilde{X}_{i\ell} \beta_\ell^{(k)} \right) \right)}{\sum_{i=1}^{N_T} w_i \tilde{X}_{ij}^2} \quad (19)$$

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

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

524 $\text{sign}(x)$ is the signum function

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

525 and $(x)_+ = \max(x, 0)$. We provide the full derivation in Appendix A.1.2.

526 5.3.2 Updates for the η parameter

527 Given $\beta^{(k+1)}$ and $\sigma^{2(k)}$, solving for $\eta^{(k+1)}$ becomes a univariate optimization problem:

$$\eta^{(k+1)} \leftarrow \arg \min_{\eta} \frac{1}{2} \sum_{i=1}^{N_T} \log(1 + \eta(\Lambda_i - 1)) + \frac{1}{2\sigma^{2(k)}} \sum_{i=1}^{N_T} \frac{\left(\tilde{Y}_i - \sum_{j=0}^p \tilde{X}_{ij+1} \beta_j^{(k+1)}\right)^2}{1 + \eta(\Lambda_i - 1)} \quad (20)$$

528 We use a bound constrained optimization algorithm [60] implemented in the `optim` function

529 in R and set the lower and upper bounds to be 0.01 and 0.99, respectively.

530 5.3.3 Updates for the σ^2 parameter

531 Conditional on $\beta^{(k+1)}$ and $\eta^{(k+1)}$, $\sigma^{2(k+1)}$ can be solved for using the following equation:

$$\sigma^{2(k+1)} \leftarrow \arg \min_{\sigma^2} \frac{N_T}{2} \log(\sigma^2) + \frac{1}{2\sigma^2} \sum_{i=1}^{N_T} \frac{\left(\tilde{Y}_i - \sum_{j=0}^p \tilde{X}_{ij+1} \beta_j\right)^2}{1 + \eta(\Lambda_i - 1)} \quad (21)$$

There exists an analytic solution for (21) given by:

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

5.3.4 Regularization path

In this section we describe how determine the sequence of tuning parameters λ at which to fit the model. Recall that our objective function has the form

$$Q_\lambda(\Theta) = \frac{N_T}{2} \log(\sigma^2) + \frac{1}{2} \sum_{i=1}^{N_T} \log(1 + \eta(\Lambda_i - 1)) + \frac{1}{2} \sum_{i=1}^{N_T} w_i \left(\tilde{Y}_i - \sum_{j=0}^p \tilde{X}_{ij+1} \beta_j \right)^2 + \lambda \sum_{j=1}^p v_j |\beta_j| \quad (23)$$

The Karush-Kuhn-Tucker (KKT) optimality conditions for (23) are given by:

$$\begin{aligned} \frac{\partial}{\partial \beta_1, \dots, \beta_p} Q_\lambda(\Theta) &= \mathbf{0}_p \\ \frac{\partial}{\partial \beta_0} Q_\lambda(\Theta) &= 0 \\ \frac{\partial}{\partial \eta} Q_\lambda(\Theta) &= 0 \\ \frac{\partial}{\partial \sigma^2} Q_\lambda(\Theta) &= 0 \end{aligned} \quad (24)$$

536 The equations in (24) are equivalent to

$$\begin{aligned}
 & \sum_{i=1}^{N_T} w_i \tilde{X}_{i1} \left(\tilde{Y}_i - \sum_{j=0}^p \tilde{X}_{ij+1} \beta_j \right) = 0 \\
 & \frac{1}{v_j} \sum_{i=1}^{N_T} w_i \tilde{X}_{ij} \left(\tilde{Y}_i - \sum_{j=0}^p \tilde{X}_{ij+1} \beta_j \right) = \lambda \gamma_j, \\
 & \gamma_j \in \begin{cases} \text{sign}(\hat{\beta}_j) & \text{if } \hat{\beta}_j \neq 0 \\ [-1, 1] & \text{if } \hat{\beta}_j = 0 \end{cases}, \quad \text{for } j = 1, \dots, p \\
 & \frac{1}{2} \sum_{i=1}^{N_T} \frac{\Lambda_i - 1}{1 + \eta(\Lambda_i - 1)} \left(1 - \frac{\left(\tilde{Y}_i - \sum_{j=0}^p \tilde{X}_{ij+1} \beta_j \right)^2}{\sigma^2 (1 + \eta(\Lambda_i - 1))} \right) = 0 \\
 & \sigma^2 - \frac{1}{N_T} \sum_{i=1}^{N_T} \frac{\left(\tilde{Y}_i - \sum_{j=0}^p \tilde{X}_{ij+1} \beta_j \right)^2}{1 + \eta(\Lambda_i - 1)} = 0
 \end{aligned} \tag{25}$$

537 where w_i is given by (18), $\tilde{\mathbf{X}}_{-1}^T$ is $\tilde{\mathbf{X}}^T$ with the first column removed, $\tilde{\mathbf{X}}_1^T$ is the first column
 538 of $\tilde{\mathbf{X}}^T$, and $\gamma \in \mathbb{R}^p$ is the subgradient function of the ℓ_1 norm evaluated at $(\hat{\beta}_1, \dots, \hat{\beta}_p)$.

539 Therefore $\widehat{\Theta}$ is a solution in (16) if and only if $\widehat{\Theta}$ satisfies (25) for some γ . We can determine
 540 a decreasing sequence of tuning parameters by starting at a maximal value for $\lambda = \lambda_{max}$
 541 for which $\hat{\beta}_j = 0$ for $j = 1, \dots, p$. In this case, the KKT conditions in (25) are equivalent
 542 to

$$\begin{aligned}
 & \frac{1}{v_j} \sum_{i=1}^{N_T} \left| w_i \tilde{X}_{ij} \left(\tilde{Y}_i - \tilde{X}_{i1} \beta_0 \right) \right| \leq \lambda, \quad \forall j = 1, \dots, p \\
 & \beta_0 = \frac{\sum_{i=1}^{N_T} w_i \tilde{X}_{i1} \tilde{Y}_i}{\sum_{i=1}^{N_T} w_i \tilde{X}_{i1}^2} \\
 & \frac{1}{2} \sum_{i=1}^{N_T} \frac{\Lambda_i - 1}{1 + \eta(\Lambda_i - 1)} \left(1 - \frac{\left(\tilde{Y}_i - \tilde{X}_{i1} \beta_0 \right)^2}{\sigma^2 (1 + \eta(\Lambda_i - 1))} \right) = 0 \\
 & \sigma^2 = \frac{1}{N_T} \sum_{i=1}^{N_T} \frac{\left(\tilde{Y}_i - \tilde{X}_{i1} \beta_0 \right)^2}{1 + \eta(\Lambda_i - 1)}
 \end{aligned} \tag{26}$$

543 We can solve the KKT system of equations in (26) (with a numerical solution for η) in order

544 to have an explicit form of the stationary point $\widehat{\Theta}_0 = \left\{ \widehat{\beta}_0, \mathbf{0}_p, \widehat{\eta}, \widehat{\sigma}^2 \right\}$. Once we have $\widehat{\Theta}_0$, we
545 can solve for the smallest value of λ such that the entire vector $(\widehat{\beta}_1, \dots, \widehat{\beta}_p)$ is 0:

$$\lambda_{max} = \max_j \left\{ \left| \frac{1}{v_j} \sum_{i=1}^{N_T} \widehat{w}_i \widetilde{X}_{ij} (\widetilde{Y}_i - \widetilde{X}_{i1} \widehat{\beta}_0) \right| \right\}, \quad j = 1, \dots, p \quad (27)$$

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

549 **5.3.5 Warm Starts**

550 The way in which we have derived the sequence of tuning parameters using the KKT con-
551 ditions, allows us to implement warm starts. That is, the solution $\widehat{\Theta}$ for λ_k is used as the
552 initial value $\Theta^{(0)}$ for λ_{k+1} . This strategy leads to computational speedups and has been
553 implemented in the `ggmix` R package.

554 **5.3.6 Prediction of the random effects**

555 We use an empirical Bayes approach (e.g. [61]) to predict the random effects \mathbf{b} . Let the
556 maximum a posteriori (MAP) estimate be defined as

$$\widehat{\mathbf{b}} = \arg \max_{\mathbf{b}} f(\mathbf{b} | \mathbf{Y}, \boldsymbol{\beta}, \eta, \sigma^2) \quad (28)$$

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

$$\begin{aligned} f(\mathbf{b}|\mathbf{Y}, \boldsymbol{\beta}, \eta, \sigma^2) &= \frac{f(\mathbf{Y}|\mathbf{b}, \boldsymbol{\beta}, \eta, \sigma^2)\pi(\mathbf{b}|\eta, \sigma^2)}{f(\mathbf{Y}|\boldsymbol{\beta}, \eta, \sigma^2)} \\ &\propto f(\mathbf{Y}|\mathbf{b}, \boldsymbol{\beta}, \eta, \sigma^2)\pi(\mathbf{b}|\eta, \sigma^2) \\ &\propto \exp \left\{ -\frac{1}{2\sigma^2}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{b})^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{b}) - \frac{1}{2\eta\sigma^2}\mathbf{b}^T\boldsymbol{\Phi}^{-1}\mathbf{b} \right\} \\ &= \exp \left\{ -\frac{1}{2\sigma^2} \left[(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{b})^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{b}) + \frac{1}{\eta}\mathbf{b}^T\boldsymbol{\Phi}^{-1}\mathbf{b} \right] \right\} \end{aligned} \quad (29)$$

Solving for (28) is equivalent to minimizing the exponent in (29):

$$\hat{\mathbf{b}} = \arg \min_{\mathbf{b}} \left\{ (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{b})^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{b}) + \frac{1}{\eta}\mathbf{b}^T\boldsymbol{\Phi}^{-1}\mathbf{b} \right\} \quad (30)$$

Taking the derivative of (30) with respect to \mathbf{b} and setting it to 0 we get:

$$\begin{aligned} 0 &= -2(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{b}) + \frac{2}{\eta}\boldsymbol{\Phi}^{-1}\mathbf{b} \\ &= -(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + \mathbf{b} + \left(\frac{1}{\eta}\boldsymbol{\Phi}^{-1} \right) \mathbf{b} \\ (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) &= \left(\mathbf{I}_{N_T \times N_T} + \frac{1}{\eta}\boldsymbol{\Phi}^{-1} \right) \mathbf{b} \\ \hat{\mathbf{b}} &= \left(\mathbf{I}_{N_T \times N_T} + \frac{1}{\hat{\eta}}\boldsymbol{\Phi}^{-1} \right)^{-1} (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\ &= \left(\mathbf{I}_{N_T \times N_T} + \frac{1}{\hat{\eta}}\mathbf{U}\mathbf{D}^{-1}\mathbf{U}^T \right)^{-1} (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \end{aligned} \quad (31)$$

557 where $(\hat{\boldsymbol{\beta}}, \hat{\eta})$ are the estimates obtained from Algorithm 1.

558 5.3.7 Phenotype prediction

559 Here we describe the method used for predicting the unobserved phenotype \mathbf{Y}^* in a set of
 560 individuals with predictor set \mathbf{X}^* that were not used in the model training e.g. a testing
 561 set. Let q denote the number of observations in the testing set and $N - q$ the number of
 562 observations in the training set. We assume that a `gmmix` model has been fit on a set of

563 training individuals with observed phenotype \mathbf{Y} and predictor set \mathbf{X} . We further assume
564 that \mathbf{Y} and \mathbf{Y}^* are jointly multivariate Normal:

$$\begin{bmatrix} \mathbf{Y}^* \\ \mathbf{Y} \end{bmatrix} \sim \mathcal{N} \left(\begin{bmatrix} \boldsymbol{\mu}_{1(q \times 1)} \\ \boldsymbol{\mu}_{2(N-q) \times 1} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{11(q \times q)} & \boldsymbol{\Sigma}_{12_{q \times (N-q)}} \\ \boldsymbol{\Sigma}_{21_{(N-q) \times q}} & \boldsymbol{\Sigma}_{22_{(N-q) \times (N-q)}} \end{bmatrix} \right) \quad (32)$$

565 Then, from standard multivariate Normal theory, the conditional distribution $\mathbf{Y}^* | \mathbf{Y}, \eta, \sigma^2, \boldsymbol{\beta}, \mathbf{X}, \mathbf{X}^*$
566 is $\mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$ where

$$\boldsymbol{\mu}^* = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_{22}^{-1} (\mathbf{Y} - \boldsymbol{\mu}_2) \quad (33)$$

$$\boldsymbol{\Sigma}^* = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\Sigma}_{21} \quad (34)$$

567 The phenotype prediction is thus given by:

$$\boldsymbol{\mu}_{q \times 1}^* = \mathbf{X}^* \boldsymbol{\beta} + \frac{1}{\sigma^2} \boldsymbol{\Sigma}_{12} \mathbf{V}^{-1} (\mathbf{Y} - \mathbf{X} \boldsymbol{\beta}) \quad (35)$$

$$= \mathbf{X}^* \boldsymbol{\beta} + \frac{1}{\sigma^2} \boldsymbol{\Sigma}_{12} \mathbf{U} \tilde{\mathbf{D}}^{-1} \mathbf{U}^T (\mathbf{Y} - \mathbf{X} \boldsymbol{\beta}) \quad (36)$$

$$= \mathbf{X}^* \boldsymbol{\beta} + \frac{1}{\sigma^2} \boldsymbol{\Sigma}_{12} \mathbf{U} \tilde{\mathbf{D}}^{-1} (\tilde{\mathbf{Y}} - \tilde{\mathbf{X}} \boldsymbol{\beta}) \quad (37)$$

$$= \mathbf{X}^* \boldsymbol{\beta} + \frac{1}{\sigma^2} \eta \sigma^2 \boldsymbol{\Phi}^* \mathbf{U} \tilde{\mathbf{D}}^{-1} (\tilde{\mathbf{Y}} - \tilde{\mathbf{X}} \boldsymbol{\beta}) \quad (38)$$

$$= \mathbf{X}^* \boldsymbol{\beta} + \eta \boldsymbol{\Phi}^* \mathbf{U} \tilde{\mathbf{D}}^{-1} (\tilde{\mathbf{Y}} - \tilde{\mathbf{X}} \boldsymbol{\beta}) \quad (39)$$

568 where $\boldsymbol{\Phi}^*$ is the $q \times (N - q)$ covariance matrix between the testing and training individu-
569 als.

570 5.3.8 Choice of the optimal tuning parameter

571 In order to choose the optimal value of the tuning parameter λ , we use the generalized
572 information criterion [62] (GIC):

$$GIC_\lambda = -2\ell(\hat{\beta}, \hat{\sigma}^2, \hat{\eta}) + a_n \cdot \hat{df}_\lambda \quad (40)$$

573 where \hat{df}_λ is the number of non-zero elements in $\hat{\beta}_\lambda$ [63] plus two (representing the variance
574 parameters η and σ^2). Several authors have used this criterion for variable selection in mixed
575 models with $a_n = \log N_T$ [57, 64], which corresponds to the BIC. We instead choose the high-
576 dimensional BIC [65] given by $a_n = \log(\log(N_T)) * \log(p)$. This is the default choice in our
577 `ggmix` R package, though the interface is flexible to allow the user to select their choice of
578 a_n .

579 **Availability of data and material**

- 580 1. The UK Biobank data is available upon successful project application.
- 581 2. The GAW20 data is freely available upon request from <https://www.gaworkshop.org/data-sets>.
- 583 3. Mouse cross data is available from GitHub at <https://github.com/sahirbhatnagar/ggmix/blob/pgen/RealData/mice.RData>.
- 585 4. The entire simulation study is reproducible. Source code available at <https://github.com/sahirbhatnagar/ggmix/tree/pgen/simulation>. This includes scripts for ggmix, lasso and twostep methods.
- 588 5. The R package ggmix is freely available from GitHub at <https://github.com/greenwoodlab/ggmix>.
- 590 6. A website describing how to use the package is available at <https://sahirbhatnagar.com/ggmix/>.
- 591

592 **Competing interests**

593 The authors declare that they have no competing interests.

594 **Author's contributions**

595 SRB, KO, YY and CMTG conceived the idea. SRB developed the algorithms, software
596 and simulation study. TL completed the real data analysis. ES and JCLO provided data
597 and interpretations. SRB, TL and CMTG wrote a draft of the manuscript then all authors
598 edited, read and approved the final manuscript.

599 **Acknowledgements**

600 SRB was supported by the Ludmer Centre for Neuroinformatics and Mental Health and
601 the Canadian Institutes for Health Research PJT 148620. This research was enabled in
602 part by support provided by Calcul Québec (www.calculquebec.ca) and Compute Canada
603 (www.computecanada.ca). The funders had no role in study design, data collection and
604 analysis, decision to publish, or preparation of the manuscript.

605 **Supporting Information**

606 Contains the following sections:

607 **A Block Coordinate Descent Algorithm** - a detailed description of the algorithm
608 used to fit our `gmmix` model.

609 **B Additional Real Data Analysis Results** - supporting information for the GAW20
610 and UK Biobank analyses

611 **C gmmix Package Showcase** - a vignette describing how to use our `gmmix` R package

612 **References**

613 [1] Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding
614 the missing heritability of complex diseases. *Nature*. 2009;461(7265):747. [3](#), [25](#)

615 [2] Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, et al. Common
616 SNPs explain a large proportion of the heritability for human height. *Nature genetics*.
617 2010;42(7):565. [3](#)

618 [3] Astle W, Balding DJ, et al. Population structure and cryptic relatedness in genetic
619 association studies. *Statistical Science*. 2009;24(4):451–471. [3](#), [4](#)

- [4] Song M, Hao W, Storey JD. Testing for genetic associations in arbitrarily structured populations. *Nature genetics*. 2015;47(5):550–554. [3](#)
- [5] Marchini J, Cardon LR, Phillips MS, Donnelly P. The effects of human population structure on large genetic association studies. *Nature genetics*. 2004;36(5):512. [3](#)
- [6] Hoggart CJ, Whittaker JC, De Iorio M, Balding DJ. Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. *PLoS genetics*. 2008;4(7):e1000130. [4, 22](#)
- [7] Li J, Das K, Fu G, Li R, Wu R. The Bayesian lasso for genome-wide association studies. *Bioinformatics*. 2010;27(4):516–523. [4](#)
- [8] Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, Heckerman D. FaST linear mixed models for genome-wide association studies. *Nature methods*. 2011;8(10):833–835. [4, 7, 22, 24](#)
- [9] Kang HM, Sul JH, Zaitlen NA, Kong Sy, Freimer NB, Sabatti C, et al. Variance component model to account for sample structure in genome-wide association studies. *Nature genetics*. 2010;42(4):348. [4](#)
- [10] Yu J, Pressoir G, Briggs WH, Bi IV, Yamasaki M, Doebley JF, et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature genetics*. 2006;38(2):203. [4](#)
- [11] Eu-Ahsunthornwattana J, Miller EN, Fakiola M, Jeronimo SM, Blackwell JM, Cordell HJ, et al. Comparison of methods to account for relatedness in genome-wide association studies with family-based data. *PLoS Genet*. 2014;10(7):e1004445. [4](#)
- [12] Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*. 2006;38(8):904. [4](#)

- 644 [13] Oualkacha K, Dastani Z, Li R, Cingolani PE, Spector TD, Hammond CJ, et al. Ad-
645 justed sequence kernel association test for rare variants controlling for cryptic and family
646 relatedness. *Genetic epidemiology*. 2013;37(4):366–376. 4, 5
- 647 [14] Cordell HJ, Clayton DG. A unified stepwise regression procedure for evaluating the
648 relative effects of polymorphisms within a gene using case/control or family data:
649 application to HLA in type 1 diabetes. *The American Journal of Human Genetics*.
650 2002;70(1):124–141. 4
- 651 [15] Rakitsch B, Lippert C, Stegle O, Borgwardt K. A Lasso multi-marker mixed
652 model for association mapping with population structure correction. *Bioinformatics*.
653 2013;29(2):206–214. 4, 23, 26
- 654 [16] Wang D, Eskridge KM, Crossa J. Identifying QTLs and epistasis in structured plant
655 populations using adaptive mixed LASSO. *Journal of agricultural, biological, and en-*
656 *vironmental statistics*. 2011;16(2):170–184. 4
- 657 [17] Tibshirani R. Regression shrinkage and selection via the lasso. *Journal of the Royal*
658 *Statistical Society Series B (Methodological)*. 1996;p. 267–288. 4, 5
- 659 [18] Zou H. The adaptive lasso and its oracle properties. *Journal of the American statistical*
660 *association*. 2006;101(476):1418–1429. 5
- 661 [19] Ding X, Su S, Nandakumar K, Wang X, Fardo DW. A 2-step penalized regression
662 method for family-based next-generation sequencing association studies. In: *BMC pro-*
663 *ceedings*. vol. 8. BioMed Central; 2014. p. S25. 5
- 664 [20] Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models
665 via coordinate descent. *Journal of statistical software*. 2010;33(1):1. 5, 6, 29, 34, 47
- 666 [21] Yang Y, Zou H. A fast unified algorithm for solving group-lasso penalize learning
667 problems. *Statistics and Computing*. 2015;25(6):1129–1141. 5

- 668 [22] Yang J, Zaitlen NA, Goddard ME, Visscher PM, Price AL. Advantages and pitfalls in
669 the application of mixed-model association methods. *Nature genetics*. 2014;46(2):100.
670 5, 24
- 671 [23] Zou H, Hastie T. Regularization and variable selection via the elastic net. *Journal of*
672 the Royal Statistical Society: Series B (Statistical Methodology). 2005;67(2):301–320.
673 5
- 674 [24] Gilmour AR, Thompson R, Cullis BR. Average information REML: an efficient algo-
675 rithm for variance parameter estimation in linear mixed models. *Biometrics*. 1995;p.
676 1440–1450. 6
- 677 [25] Dandine-Roulland C. *gaston: Genetic Data Handling (QC, GRM, LD, PCA) and*
678 *Linear Mixed Models*; 2018. R package version 1.5.3. Available from: <https://CRAN.R-project.org/package=gaston>. 6
- 680 [26] Ochoa A, Storey JD. FST and kinship for arbitrary population structures I: Generalized
681 definitions. *bioRxiv*. 2016;. 8
- 682 [27] Ochoa A, Storey JD. FST and kinship for arbitrary population structures II: Method
683 of moments estimators. *bioRxiv*. 2016;. 8
- 684 [28] Reid S, Tibshirani R, Friedman J. A study of error variance estimation in lasso regres-
685 sion. *Statistica Sinica*. 2016;p. 35–67. 11
- 686 [29] Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK Biobank
687 resource with deep phenotyping and genomic data. *Nature*. 2018;562(7726):203. 12
- 688 [30] Biobank U. Genotyping and quality control of UK Biobank, a large-scale, ex-
689 tensively phenotyped prospective resource. Available at biobank ctsu ox ac
690 uk/crystal/docs/genotyping_qc pdf Accessed April. 2015;1:2016. 12

- 691 [31] Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM. Robust relation-
692 ship inference in genome-wide association studies. *Bioinformatics*. 2010;26(22):2867–
693 2873. 12
- 694 [32] Yengo L, Sidorenko J, Kemper KE, Zheng Z, Wood AR, Weedon MN, et al. Meta-
695 analysis of genome-wide association studies for height and body mass index in 700000
696 individuals of European ancestry. *Human molecular genetics*. 2018;27(20):3641–3649.
697 12, 13
- 698 [33] McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, Teumer A, et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nature genetics*.
699 2016;48(10):1279. 13
- 700 701 [34] Zhou X, Carbonetto P, Stephens M. Polygenic modeling with Bayesian sparse linear
702 mixed models. *PLoS genetics*. 2013;9(2):e1003264. 13
- 703 704 [35] Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association stud-
705 ies. *Nature genetics*. 2012;44(7):821. 13
- 706 707 [36] Davey Smith G, Ebrahim S. ‘Mendelian randomization’: can genetic epidemiology con-
708 tribute to understanding environmental determinants of disease? *International journal
709 of epidemiology*. 2003;32(1):1–22. 14
- 710 711 [37] Cherlin S, Howey RA, Cordell HJ. Using penalized regression to predict phenotype
712 from SNP data. In: *BMC proceedings*. vol. 12. BioMed Central; 2018. p. 38. 14
- 713 714 [38] Zhou W, Lo SH. Analysis of genotype by methylation interactions through sparsity-
715 inducing regularized regression. In: *BMC proceedings*. vol. 12. BioMed Central; 2018.
716 p. 40. 14
- 717 [39] Howey RA, Cordell HJ. Application of Bayesian networks to GAW20 genetic and blood
718 lipid data. In: *BMC proceedings*. vol. 12. BioMed Central; 2018. p. 19. 15

- 715 [40] Thornton T, Tang H, Hoffmann TJ, Ochs-Balcom HM, Caan BJ, Risch N. Estimating kinship in admixed populations. *The American Journal of Human Genetics*.
716 2012;91(1):122–138. 15
- 717
- 718 [41] Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in
719 unrelated individuals. *Genome research*. 2009;19(9):1655–1664. 15
- 720
- 721 [42] Fortin A, Diez E, Rochefort D, Laroche L, Malo D, Rouleau GA, et al. Recombinant
722 congenic strains derived from A/J and C57BL/6J: a tool for genetic dissection of com-
plex traits. *Genomics*. 2001;74(1):21–35. 18
- 723
- 724 [43] Bennett BJ, Farber CR, Orozco L, Kang HM, Ghazalpour A, Siemers N, et al. A
725 high-resolution association mapping panel for the dissection of complex traits in mice.
Genome research. 2010;20(2):281–290. 18
- 726
- 727 [44] Flint J, Eskin E. Genome-wide association studies in mice. *Nature Reviews Genetics*.
2012;13(11):807. 18
- 728
- 729 [45] Cheng R, Lim JE, Samocha KE, Sokoloff G, Abney M, Skol AD, et al. Genome-wide
730 association studies and the problem of relatedness among advanced intercross lines and
other highly recombinant populations. *Genetics*. 2010;185(3):1033–1044. 18
- 731
- 732 [46] Di Pietrantonio T, Hernandez C, Girard M, Verville A, Orlova M, Belley A, et al.
Strain-specific differences in the genetic control of two closely related mycobacteria.
733 PLoS pathogens. 2010;6(10):e1001169. 18
- 734
- 735 [47] Wang H, Lengerich BJ, Aragam B, Xing EP. Precision Lasso: accounting for cor-
relations and linear dependencies in high-dimensional genomic data. *Bioinformatics*.
736 2018;35(7):1181–1187. 19, 23
- 737 [48] Sohrabi Y, Havelková H, Kobets T, Šíma M, Volkova V, Grekov I, et al. Mapping the

- 738 Genes for Susceptibility and Response to *Leishmania tropica* in Mouse. PLoS neglected
739 tropical diseases. 2013;7(7):e2282. 19
- 740 [49] Jackson AU, Fornés A, Galecki A, Miller RA, Burke DT. Multiple-trait quantitative
741 trait loci analysis using a large mouse sibship. Genetics. 1999;151(2):785–795. 19
- 742 [50] C Stern1 M, Benavides F, A Klingelberger E, J Conti2 C. Allelotype analysis of chemi-
743 cally induced squamous cell carcinomas in F1 hybrids of two inbred mouse strains with
744 different susceptibility to tumor progression. Carcinogenesis. 2000;21(7):1297–1301. 19
- 745 [51] Lasko D, Cavenee W, Nordenskjöld M. Loss of constitutional heterozygosity in human
746 cancer. Annual review of genetics. 1991;25(1):281–314. 19
- 747 [52] Loh PR, Tucker G, Bulik-Sullivan BK, Vilhjalmsson BJ, Finucane HK, Salem RM, et al.
748 Efficient Bayesian mixed-model analysis increases association power in large cohorts.
749 Nature genetics. 2015;47(3):284. 22
- 750 [53] Allen N, Sudlow C, Downey P, Peakman T, Danesh J, Elliott P, et al. UK Biobank:
751 Current status and what it means for epidemiology. Health Policy and Technology.
752 2012;1(3):123–126. 23
- 753 [54] Zeng Y, Breheny P. The biglasso package: a memory-and computation-efficient solver
754 for lasso model fitting with big data in R. arXiv preprint arXiv:170105936. 2017;. 24
- 755 [55] Spain SL, Barrett JC. Strategies for fine-mapping complex traits. Human molecular
756 genetics. 2015;24(R1):R111–R119. 24
- 757 [56] Pirinen M, Donnelly P, Spencer CC, et al. Efficient computation with a linear mixed
758 model on large-scale data sets with applications to genetic studies. The Annals of
759 Applied Statistics. 2013;7(1):369–390. 25, 26

- 760 [57] Schelldorfer J, Bühlmann P, DE G, VAN S. Estimation for High-Dimensional Lin-
761 ear Mixed-Effects Models Using L1-Penalization. Scandinavian Journal of Statistics.
762 2011;38(2):197–214. [29](#), [37](#), [47](#)
- 763 [58] Tseng P, Yun S. A coordinate gradient descent method for nonsmooth separable mini-
764 mization. Mathematical Programming. 2009;117(1):387–423. [29](#), [47](#), [50](#)
- 765 [59] Meier L, Van De Geer S, Bühlmann P. The group lasso for logistic regression. Journal
766 of the Royal Statistical Society: Series B (Statistical Methodology). 2008;70(1):53–71.
767 [29](#), [47](#)
- 768 [60] Byrd RH, Lu P, Nocedal J, Zhu C. A limited memory algorithm for bound constrained
769 optimization. SIAM Journal on Scientific Computing. 1995;16(5):1190–1208. [31](#)
- 770 [61] Wakefield J. Bayesian and frequentist regression methods. Springer Science & Business
771 Media; 2013. [34](#)
- 772 [62] Nishii R. Asymptotic properties of criteria for selection of variables in multiple regres-
773 sion. The Annals of Statistics. 1984;p. 758–765. [37](#)
- 774 [63] Zou H, Hastie T, Tibshirani R, et al. On the “degrees of freedom” of the lasso. The
775 Annals of Statistics. 2007;35(5):2173–2192. [37](#)
- 776 [64] Bondell HD, Krishna A, Ghosh SK. Joint Variable Selection for Fixed and Random
777 Effects in Linear Mixed-Effects Models. Biometrics. 2010;66(4):1069–1077. [37](#)
- 778 [65] Fan Y, Tang CY. Tuning parameter selection in high dimensional penalized likeli-
779 hood. Journal of the Royal Statistical Society: Series B (Statistical Methodology).
780 2013;75(3):531–552. [37](#)
- 781 [66] Xie Y. Dynamic Documents with R and knitr. vol. 29. CRC Press; 2015. [57](#)

782 A Block Coordinate Descent Algorithm

783 We use a general purpose block coordinate descent algorithm (CGD) [58] to solve (16). At
 784 each iteration, the algorithm approximates the negative log-likelihood $f(\cdot)$ in $Q_\lambda(\cdot)$ by a
 785 strictly convex quadratic function and then applies block coordinate decent to generate a
 786 decent direction followed by an inexact line search along this direction [58]. For continuously
 787 differentiable $f(\cdot)$ and convex and block-separable $P(\cdot)$ (i.e. $P(\beta) = \sum_i P_i(\beta_i)$), [58] show
 788 that the solution generated by the CGD method is a stationary point of $Q_\lambda(\cdot)$ if the coor-
 789 dinates are updated in a Gauss-Seidel manner i.e. $Q_\lambda(\cdot)$ is minimized with respect to one
 790 parameter while holding all others fixed. The CGD algorithm can thus be run in parallel and
 791 therefore suited for large p settings. It has been successfully applied in fixed effects models
 792 (e.g. [59], [20]) and [57] for mixed models with an ℓ_1 penalty. Following Tseng and Yun [58],
 793 the CGD algorithm is given by Algorithm 2.

794 The Armijo rule is defined as follows [58]:

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

$$Q_\lambda(\Theta_j^{(k)} + \alpha^{(k)} d^{(k)}) \leq Q_\lambda(\Theta_j^{(k)}) + \alpha^{(k)} \varrho \Delta^{(k)} \quad (45)$$

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

$$\Delta^{(k)} := \nabla f(\Theta_j^{(k)}) d^{(k)} + \gamma (d^{(k)})^2 H_{jj}^{(k)} + \lambda P(\Theta_j^{(k)} + d^{(k)}) - \lambda P(\Theta_j^{(k)}) \quad (46)$$

795 796 Common choices for the constants are $\delta = 0.1$, $\varrho = 0.001$, $\gamma = 0$, $\alpha_{init}^{(k)} = 1$ for all k [57].

797 Below we detail the specifics of Algorithm 2 for the ℓ_1 penalty.

Algorithm 2: Coordinate Gradient Descent Algorithm to solve (16)

Set the iteration counter $k \leftarrow 0$ and choose initial values for the parameter vector

$$\Theta^{(0)};$$

repeat

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

$$H^{(k)} = \text{diag} \left[\min \left\{ \max \left\{ \left[\nabla^2 f(\Theta^{(k)}) \right]_{jj}, c_{min} \right\} c_{max} \right\} \right]_{j=1,\dots,p} \quad (41)$$

for $j = 1, \dots, p$ **do**

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

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

$$d_{H^{(k)}}(\Theta_j^{(k)}) \leftarrow \arg \min_d \left\{ \nabla f(\Theta_j^{(k)})d + \frac{1}{2}d^2 H_{jj}^{(k)} + \lambda P(\Theta_j^{(k)} + d) \right\} \quad (42)$$

end
end

Choose a stepsize;

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

Update;

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

Update;

$$\widehat{\eta}^{(k+1)} \leftarrow \arg \min_{\eta} \frac{1}{2} \sum_{i=1}^{N_T} \log(1 + \eta(\Lambda_i - 1)) + \frac{1}{2\sigma^2(k)} \sum_{i=1}^{N_T} \frac{\left(\widetilde{Y}_i - \sum_{j=0}^p \widetilde{X}_{ij+1} \beta_j^{(k+1)} \right)^2}{1 + \eta(\Lambda_i - 1)} \quad (43)$$

Update;

$$\widehat{\sigma^2}^{(k+1)} \leftarrow \frac{1}{N_T} \sum_{i=1}^{N_T} \frac{\left(\widetilde{Y}_i - \sum_{j=0}^p \widetilde{X}_{ij+1} \beta_j^{(k+1)} \right)^2}{1 + \eta^{(k+1)}(\Lambda_i - 1)} \quad (44)$$

$$k \leftarrow k + 1$$

until convergence criterion is satisfied;

798 **A.1 ℓ_1 penalty**

799 The objective function is given by

$$Q_\lambda(\Theta) = f(\Theta) + \lambda|\beta| \quad (47)$$

800 **A.1.1 Descent Direction**

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

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

$$d_H(\Theta_j) = \arg \min_d G(d) \quad (48)$$

803 where

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

804 Since $G(d)$ is not differentiable at $-\Theta_j$, we calculate the subdifferential $\partial G(d)$ and search

805 for d with $0 \in \partial G(d)$:

$$\partial G(d) = \nabla f(\Theta_j) + dH_{jj} + \lambda u \quad (49)$$

806 where

$$u = \begin{cases} 1 & \text{if } d > -\Theta_j \\ -1 & \text{if } d < -\Theta_j \\ [-1, 1] & \text{if } d = \Theta_j \end{cases} \quad (50)$$

807 We consider each of the three cases in (49) 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

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

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 $\text{mid} \{a, b, c\}$ denotes the median (mid-point) of a, b, c [58].

2. $d < -\Theta_j$

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

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

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

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

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

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

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

The solution can again be written compactly as

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

809 We see all three cases lead to the same solution for (48). Therefore the descent direction for
810 $\Theta_j^{(k)} \in \{\beta_1, \dots, \beta_p\}$ for the ℓ_1 penalty is given by

$$d = \text{mid} \left\{ \frac{-(\nabla f(\beta_j) - \lambda)}{H_{jj}}, -\beta_j, \frac{-(\nabla f(\beta_j) + \lambda)}{H_{jj}} \right\} \quad (51)$$

811 **A.1.2 Solution for the β parameter**

812 If the Hessian $\nabla^2 f(\Theta^{(k)}) > 0$ then $H^{(k)}$ defined in (41) is equal to $\nabla^2 f(\Theta^{(k)})$. Using $\alpha_{init} = 1$,
813 the largest element of $\{\alpha_{init}^{(k)} \delta^r\}_{r=0,1,2,\dots}$ satisfying the Armijo Rule inequality is reached for
814 $\alpha^{(k)} = \alpha_{init}^{(k)} \delta^0 = 1$. The Armijo rule update for the β parameter is then given by

$$\beta_j^{(k+1)} \leftarrow \beta_j^{(k)} + d^{(k)}, \quad j = 1, \dots, p \quad (52)$$

815 Substituting the descent direction given by (51) into (52) we get

$$\beta_j^{(k+1)} = \text{mid} \left\{ \beta_j^{(k)} + \frac{-(\nabla f(\beta_j^{(k)}) - \lambda)}{H_{jj}}, 0, \beta_j^{(k)} + \frac{-(\nabla f(\beta_j^{(k)}) + \lambda)}{H_{jj}} \right\} \quad (53)$$

816 We can further simplify this expression. Let

$$w_i := \frac{1}{\sigma^2 (1 + \eta(\Lambda_i - 1))} \quad (54)$$

Re-write the part depending on β of the negative log-likelihood in (14) as

$$g(\boldsymbol{\beta}^{(k)}) = \frac{1}{2} \sum_{i=1}^{N_T} w_i \left(\tilde{Y}_i - \sum_{\ell \neq j} \tilde{X}_{i\ell} \beta_\ell^{(k)} - \tilde{X}_{ij} \beta_j^{(k)} \right)^2 \quad (55)$$

The gradient and Hessian are given by

$$\nabla f(\beta_j^{(k)}) := \frac{\partial}{\partial \beta_j^{(k)}} g(\boldsymbol{\beta}^{(k)}) = - \sum_{i=1}^{N_T} w_i \tilde{X}_{ij} \left(\tilde{Y}_i - \sum_{\ell \neq j} \tilde{X}_{i\ell} \beta_\ell^{(k)} - \tilde{X}_{ij} \beta_j^{(k)} \right) \quad (56)$$

$$H_{jj} := \frac{\partial^2}{\partial \beta_j^{(k)} \partial \beta_j^{(k)}} g(\boldsymbol{\beta}^{(k)}) = \sum_{i=1}^{N_T} w_i \tilde{X}_{ij}^2 \quad (57)$$

Substituting (56) and (57) into $\beta_j^{(k)} + \frac{-(\nabla f(\beta_j^{(k)}) - \lambda)}{H_{jj}}$

$$\begin{aligned} & \beta_j^{(k)} + \frac{\sum_{i=1}^{N_T} w_i \tilde{X}_{ij} \left(\tilde{Y}_i - \sum_{\ell \neq j} \tilde{X}_{i\ell} \beta_\ell^{(k)} - \tilde{X}_{ij} \beta_j^{(k)} \right) + \lambda}{\sum_{i=1}^{N_T} w_i \tilde{X}_{ij}^2} \\ &= \beta_j^{(k)} + \frac{\sum_{i=1}^{N_T} w_i \tilde{X}_{ij} \left(\tilde{Y}_i - \sum_{\ell \neq j} \tilde{X}_{i\ell} \beta_\ell^{(k)} \right) + \lambda}{\sum_{i=1}^{N_T} w_i \tilde{X}_{ij}^2} - \frac{\sum_{i=1}^{N_T} w_i \tilde{X}_{ij}^2 \beta_j^{(k)}}{\sum_{i=1}^{N_T} w_i \tilde{X}_{ij}^2} \\ &= \frac{\sum_{i=1}^{N_T} w_i \tilde{X}_{ij} \left(\tilde{Y}_i - \sum_{\ell \neq j} \tilde{X}_{i\ell} \beta_\ell^{(k)} \right) + \lambda}{\sum_{i=1}^{N_T} w_i \tilde{X}_{ij}^2} \end{aligned} \quad (58)$$

Similarly, substituting (56) and (57) in $\beta_j^{(k)} + \frac{-(\nabla f(\beta_j^{(k)}) + \lambda)}{H_{jj}}$ we get

$$\frac{\sum_{i=1}^{N_T} w_i \tilde{X}_{ij} \left(\tilde{Y}_i - \sum_{\ell \neq j} \tilde{X}_{i\ell} \beta_\ell^{(k)} \right) - \lambda}{\sum_{i=1}^{N_T} w_i \tilde{X}_{ij}^2} \quad (59)$$

Finally, substituting (58) and (59) into (53) we get

$$\begin{aligned}\beta_j^{(k+1)} &= \text{mid} \left\{ \frac{\sum_{i=1}^{N_T} w_i \tilde{X}_{ij} \left(\tilde{Y}_i - \sum_{\ell \neq j} \tilde{X}_{i\ell} \beta_\ell^{(k)} \right) - \lambda}{\sum_{i=1}^{N_T} w_i \tilde{X}_{ij}^2}, 0, \frac{\sum_{i=1}^{N_T} w_i \tilde{X}_{ij} \left(\tilde{Y}_i - \sum_{\ell \neq j} \tilde{X}_{i\ell} \beta_\ell^{(k)} \right) + \lambda}{\sum_{i=1}^{N_T} w_i \tilde{X}_{ij}^2} \right\} \\ &= \frac{\mathcal{S}_\lambda \left(\sum_{i=1}^{N_T} w_i \tilde{X}_{ij} \left(\tilde{Y}_i - \sum_{\ell \neq j} \tilde{X}_{i\ell} \beta_\ell^{(k)} \right) \right)}{\sum_{i=1}^{N_T} w_i \tilde{X}_{ij}^2}\end{aligned}\tag{60}$$

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

819 **B Additional Real Data Analysis Results**

820 **B.1 Distribution of SNPs used in UK Biobank analysis**

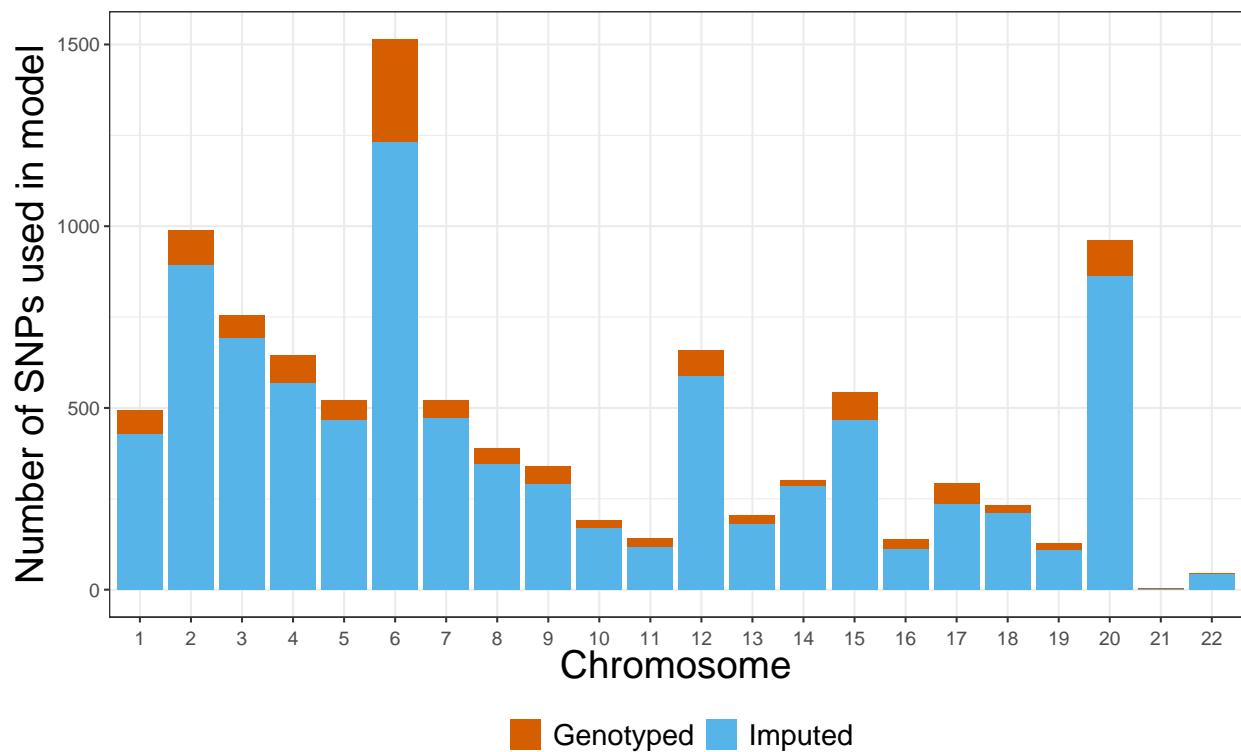


Figure B.1: Distribution of SNPs used in UK Biobank analysis by chromosome and whether or not the SNP was imputed.

821 **B.2 LD structure among the markers in the GAW20 and the mouse**
 822 **dataset**

823 We illustrate the LD structure among the markers in the GAW20 dataset and the mouse
 824 dataset separately in Figures B.2 and B.3, respectively. In Figure B.2, we show the pairwise
 825 r^2 for 655 SNPs within a 1Mb-window around the causal SNP rs9661059 (indicated) that we
 826 focused on. The dotplot above the heatmap denotes r^2 between each SNP and the causal
 827 SNP. It is clear that although strong correlation does exist between some SNPs, none of these
 828 nearby SNPs is correlated with the causal SNP. The only dot denoting an $r^2 = 1$ represents
 829 the causal SNP itself.

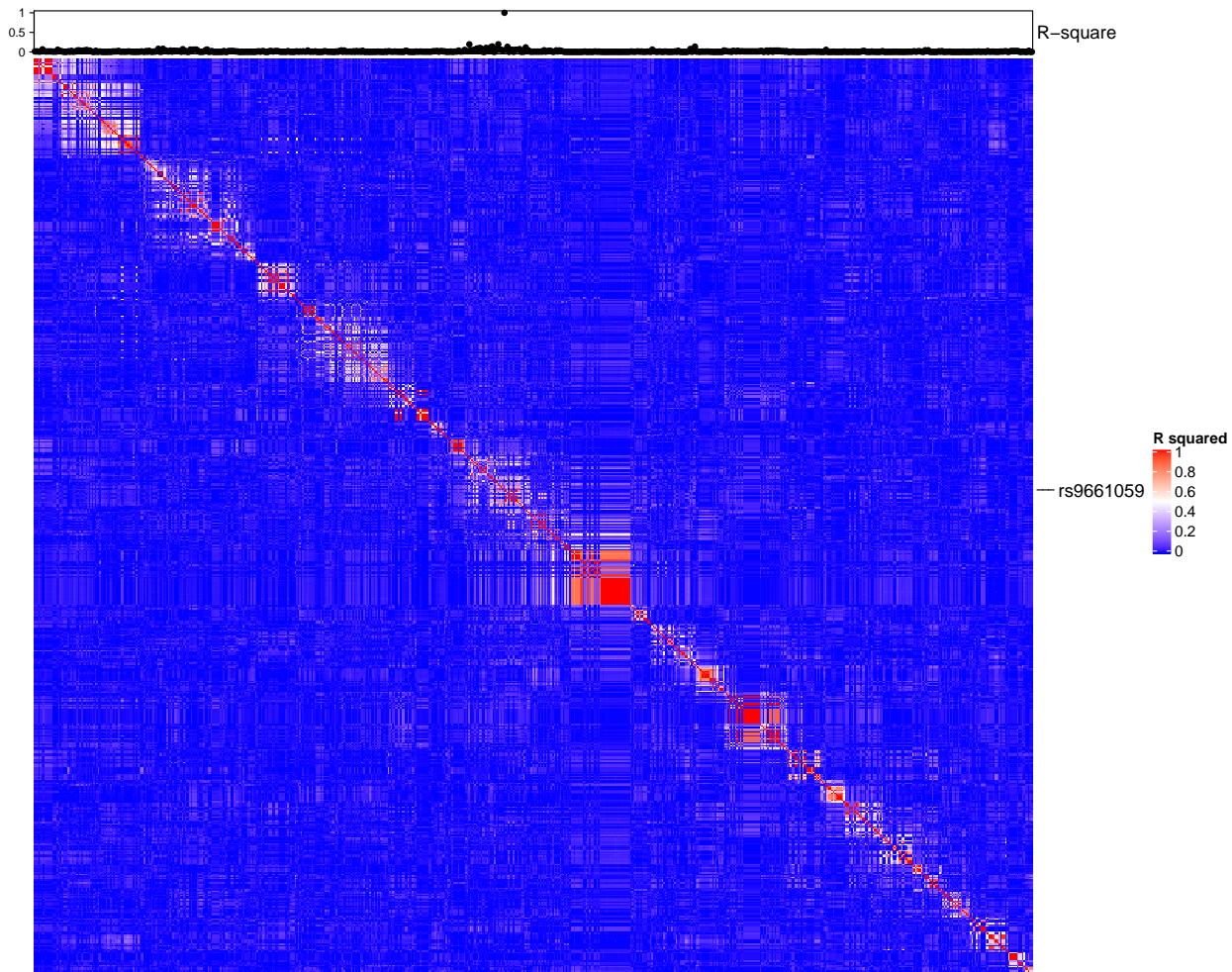


Figure B.2: LD structure among the markers in the GAW20 dataset

830 In Figure B.3, we show the pairwise r^2 for all microsatellite markers in the mouse dataset.
831 It is clear that many markers are considerably strongly correlated with each other, as we
832 expected.

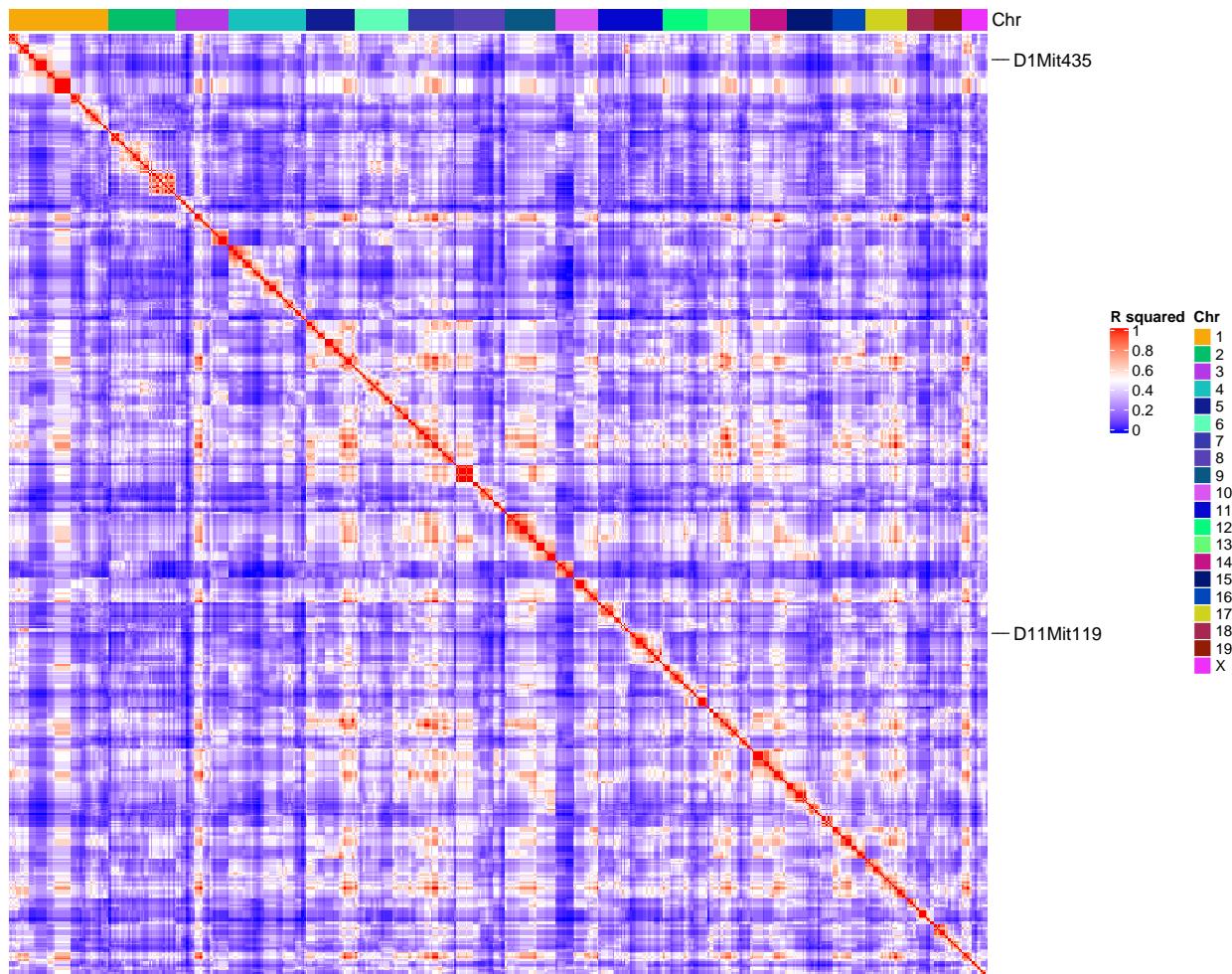


Figure B.3: LD structure among the markers in the mouse dataset

833 C ggmix Package Showcase

- 834 In this section we briefly introduce the freely available and open source `ggmix` package in R.
- 835 More comprehensive documentation is available at <https://sahirbhatnagar.com/ggmix>.
- 836 Note that this entire section is reproducible; the code and text are combined in an `.Rnw`¹ file
- 837 and compiled using `knitr` [66].

838 C.1 Installation

- 839 The package can be installed from [GitHub](#) via

```
install.packages("pacman")
pacman::p_load_gh('sahirbhatnagar/ggmix')
```

- 840 To showcase the main functions in `ggmix`, we will use the simulated data which ships with
 841 the package and can be loaded via:

```
## library(ggmix)
data("admixed")
names(admixed)

## [1] "ytrain"      "ytune"        "ytest"        "xtrain"
## [5] "xtune"        "xtest"        "xtrain_lasso" "xtune_lasso"
## [9] "xtest_lasso"  "Xkinship"     "kin_train"    "kin_tune_train"
## [13] "kin_test_train" "mu_train"     "causal"       "beta"
## [17] "not_causal"   "kinship"     "coancestry"  "PC"
## [21] "subpops"
```

- 842 For details on how this data was simulated, see `help(admixed)`.

- 843 There are three basic inputs that `ggmix` needs:

- 844 1. Y : a continuous response variable
- 845 2. X : a matrix of covariates of dimension $N \times p$ where N is the sample size and p is the
 846 number of covariates
- 847 3. Φ : a kinship matrix

¹scripts available at <https://github.com/sahirbhatnagar/ggmix/tree/pgen/manuscript>

848 We can visualize the kinship matrix in the `admixed` data using the `popkin` package:

```
# need to install the package if you don't have it
# pacman::p_load_gh('StoreyLab/popkin')
popkin::plot_popkin(admixed$kin)

## Error in popkin::plot_popkin(admixed$kin): Every element of list "kinship" is NULL!
```

849 C.2 Fit the linear mixed model with Lasso Penalty

850 We will use the most basic call to the main function of this package, which is called `ggmix`.

851 This function will by default fit a L_1 penalized linear mixed model (LMM) for 100 distinct

852 values of the tuning parameter λ . It will choose its own sequence:

```
fit <- ggmix(x = admixed$xtrain,
              y = admixed$ytrain,
              kinship = admixed$kin_train)
names(fit)

## [1] "result"      "ggmix_object"  "n_design"     "p_design"    "lambda"
## [6] "coef"         "b0"           "beta"        "df"          "eta"
## [11] "sigma2"       "nlambda"       "cov_names"   "call"

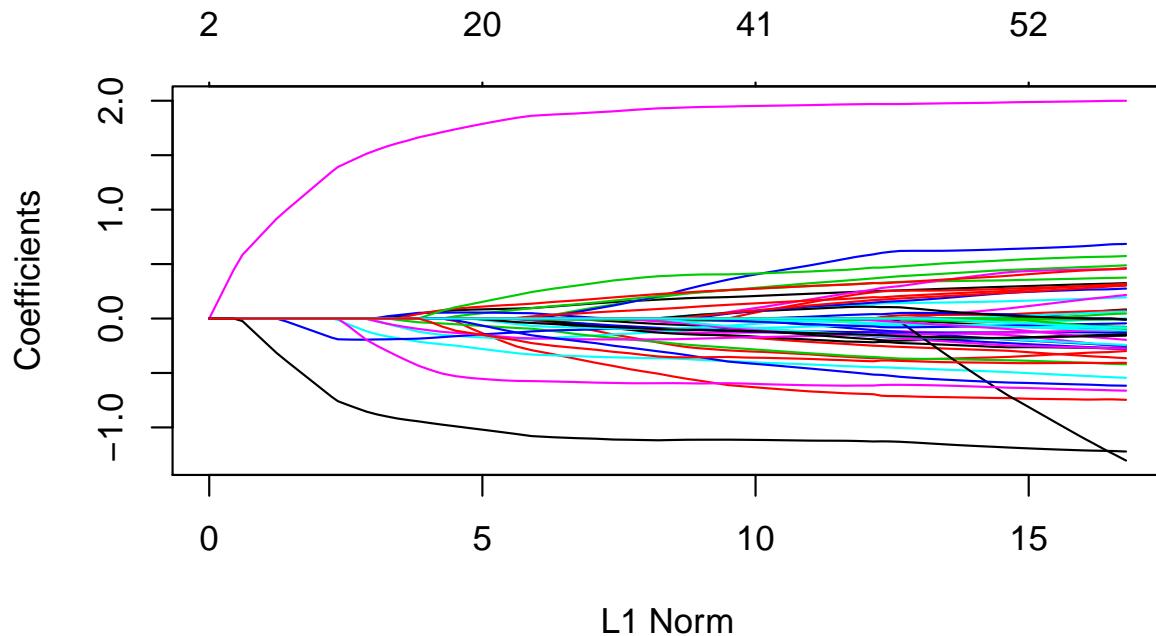
class(fit)

## [1] "lassofullrank" "ggmix_fit"
```

853 We can see the solution path for each variable by calling the `plot` method for objects of

854 class `ggmix_fit`:

```
plot(fit)
```



855

856 We can also get the coefficients for given value(s) of lambda using the `coef` method for
 857 objects of class `ggmix_fit`:

```
# only the first 5 coefficients printed here for brevity
coef(fit, s = c(0.1,0.02))[1:5, ]

## 5 x 2 Matrix of class "dgeMatrix"
##           1         2
## (Intercept) -0.03715135  0.247105426
## X23        0.00000000  0.098030248
## X36        0.00000000 -0.013022250
## X38        0.00000000  0.005378361
## X40        0.00000000  0.004028934
```

858 Here, `s` specifies the value(s) of λ at which the extraction is made. The function uses linear
 859 interpolation to make predictions for values of `s` that do not coincide with the lambda
 860 sequence used in the fitting algorithm.

861 We can also get predictions ($X\hat{\beta}$) using the `predict` method for objects of class `ggmix_fit`:

```
# need to provide x to the predict function
```

```
# predict for the first 5 subjects
predict(fit, s = c(0.1,0.02), newx = admixed$xtest[1:5,])

##          1          2
## id26  2.30208546  2.45597763
## id39  0.87334032  1.62931898
## id45 -0.12296837 -0.06075786
## id52 -0.03715135 -0.97519671
## id53 -0.21046107 -0.23151040
```

862 C.3 Find the Optimal Value of the Tuning Parameter

863 We use the Generalized Information Criterion (GIC) to select the optimal value for λ . The
 864 default is $a_n = \log(\log(n)) * \log(p)$ which corresponds to a high-dimensional BIC (HD-
 865 BIC):

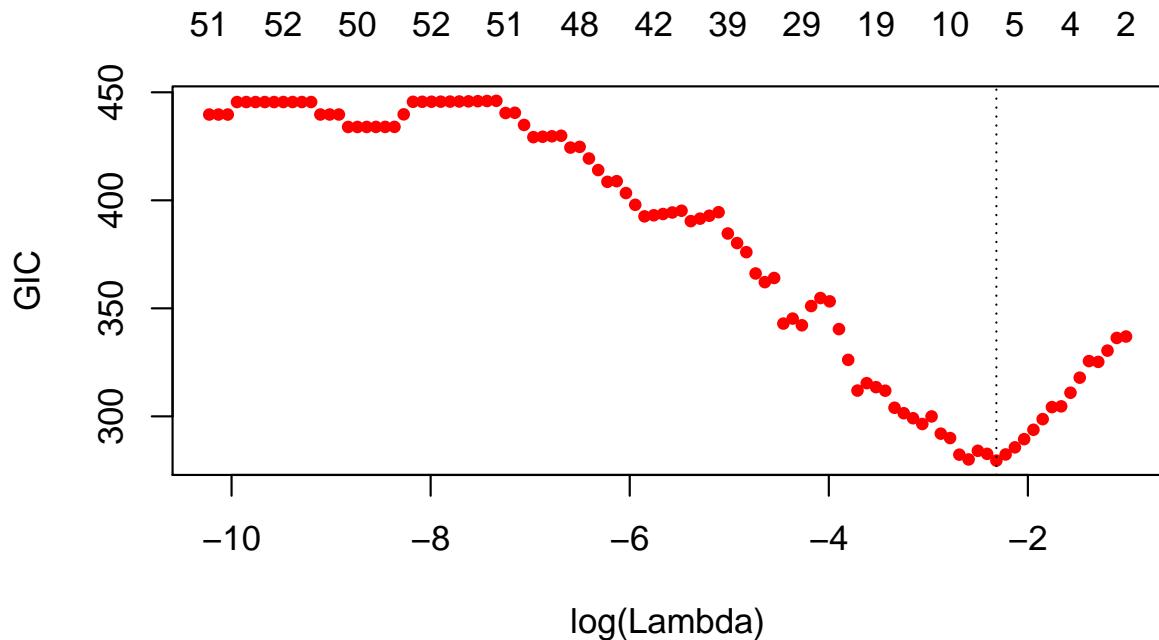
```
# pass the fitted object from ggmix to the gic function:
hdbic <- gic(fit)
class(hdbic)

## [1] "ggmix_gic"      "lassofullrank" "ggmix_fit"

# we can also fit the BIC by specifying the an argument
bicfit <- gic(fit, an = log(length(admixed$ytrain)))
```

866 We can plot the HDBIC values against $\log(\lambda)$ using the `plot` method for objects of class
 867 `ggmix_gic`:

```
plot(hdbic)
```



868

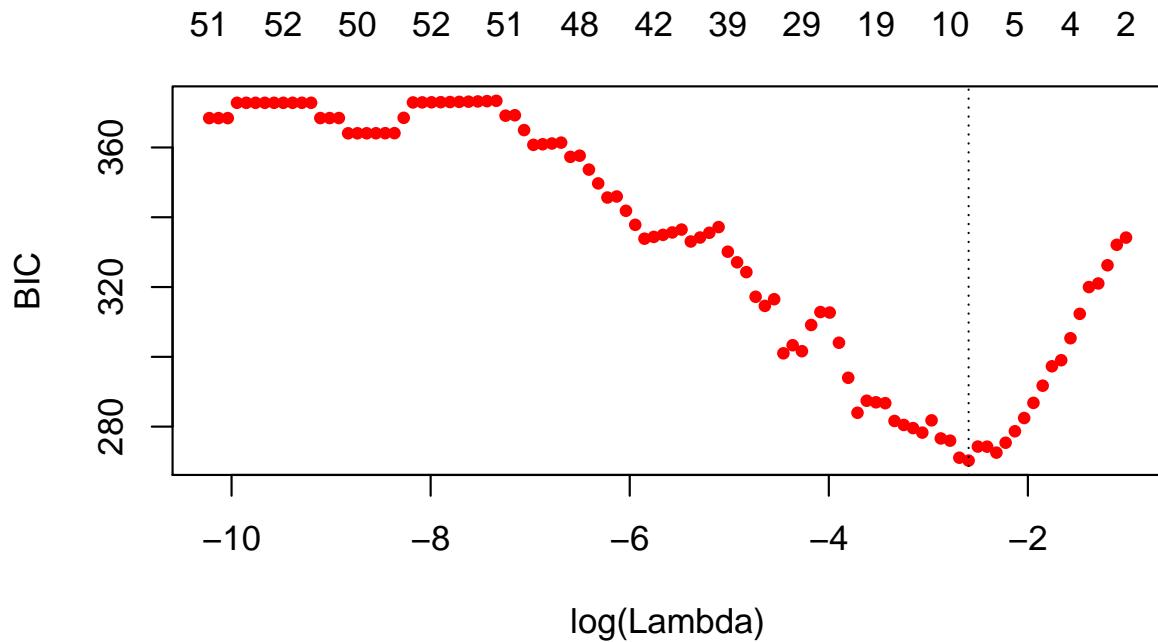
869 The optimal value for λ according to the HDBIC, i.e., the λ that leads to the minium HDBIC

870 is:

```
hdbic[["lambda.min"]]  
## [1] 0.09862269
```

871 We can also plot the BIC results:

```
plot(bicfit, ylab = "BIC")
```



872

```
bicfit[["lambda.min"]]
## [1] 0.07460445
```

873 C.4 Get Coefficients Corresponding to Optimal Model

874 We can use the object outputted by the `gic` function to extract the coefficients corresponding
 875 to the selected model using the `coef` method for objects of class `ggmix_gic`:

```
coef(hdbic)[1:5, , drop = FALSE]
## 5 x 1 sparse Matrix of class "dgCMatrix"
##           1
## (Intercept) -0.03660806
## X23         .
## X36         .
## X38         .
## X40         .
```

876 We can also extract just the nonzero coefficients which also provide the estimated variance
 877 components η and σ^2 :

```

coef(hdbic, type = "nonzero")

##          1
## (Intercept) -0.03660806
## X302       -0.17607392
## X524        1.34951500
## X538       -0.72052613
## eta         0.99000000
## sigma2      1.60476289

```

- 878 We can also make predictions from the `hdbic` object, which by default will use the model
 879 corresponding to the optimal tuning parameter:

```

predict(hdbic, newx = admixed$xtest[1:5,])

##          1
## id26   2.31027410
## id39   0.86922183
## id45  -0.12814532
## id52  -0.03660806
## id53  -0.21268198

```

880 C.5 Extracting Random Effects

- 881 The user can compute the random effects using the provided `ranef` method for objects of
 882 class `ggmix_gic`. This command will compute the estimated random effects for each subject
 883 using the parameters of the selected model:

```

ranef(hdbic)[1:5]

## [1] -2.4889655  1.1834200 -0.5641832 -0.9310334 -0.3458703

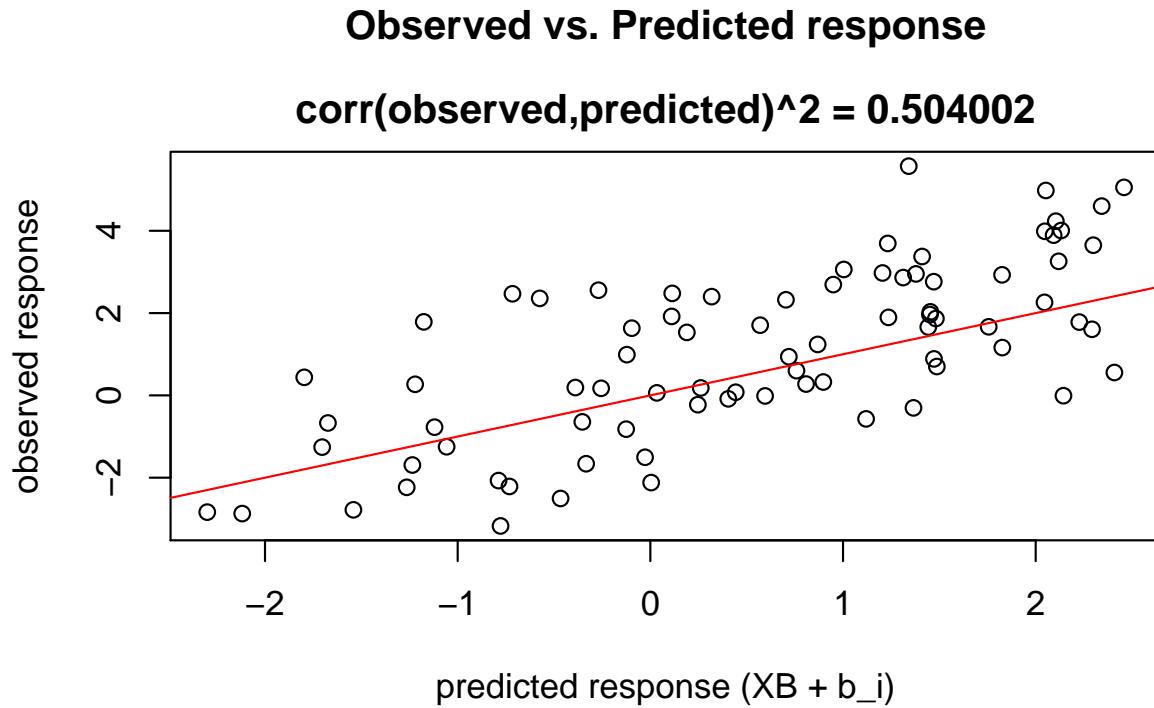
```

884 C.6 Diagnostic Plots

- 885 We can also plot some standard diagnostic plots such as the observed vs. predicted response,
 886 QQ-plots of the residuals and random effects and the Tukey-Anscombe plot. These can be
 887 plotted using the `plot` method on a `ggmix_gic` object as shown below.

888 C.6.1 Observed vs. Predicted Response

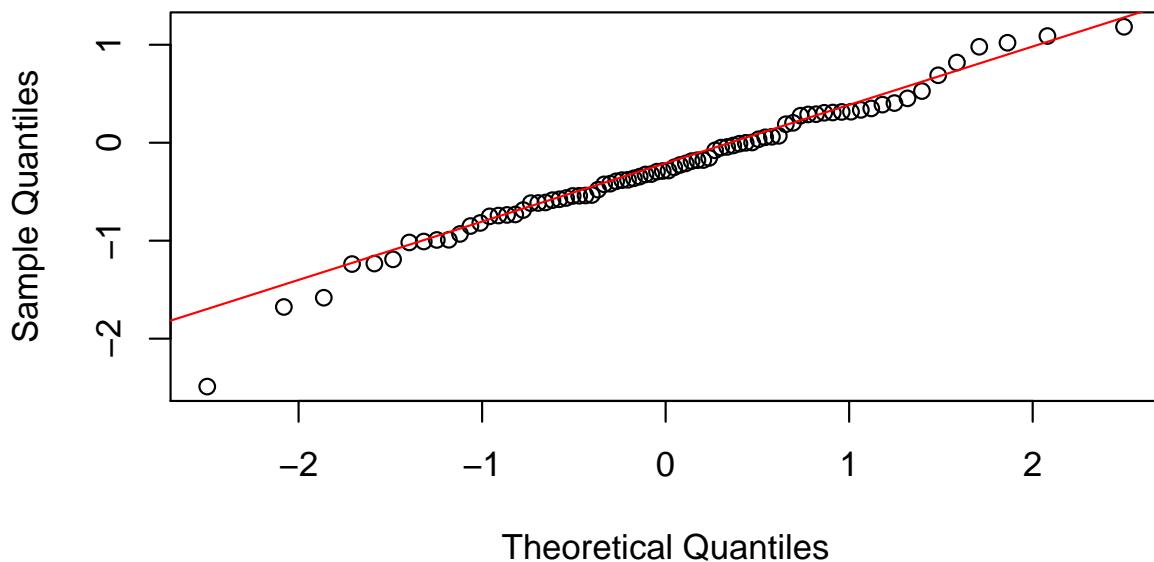
```
plot(hdbic, type = "predicted", newx = admixed$xtrain, newy = admixed$ytrain)
```



889

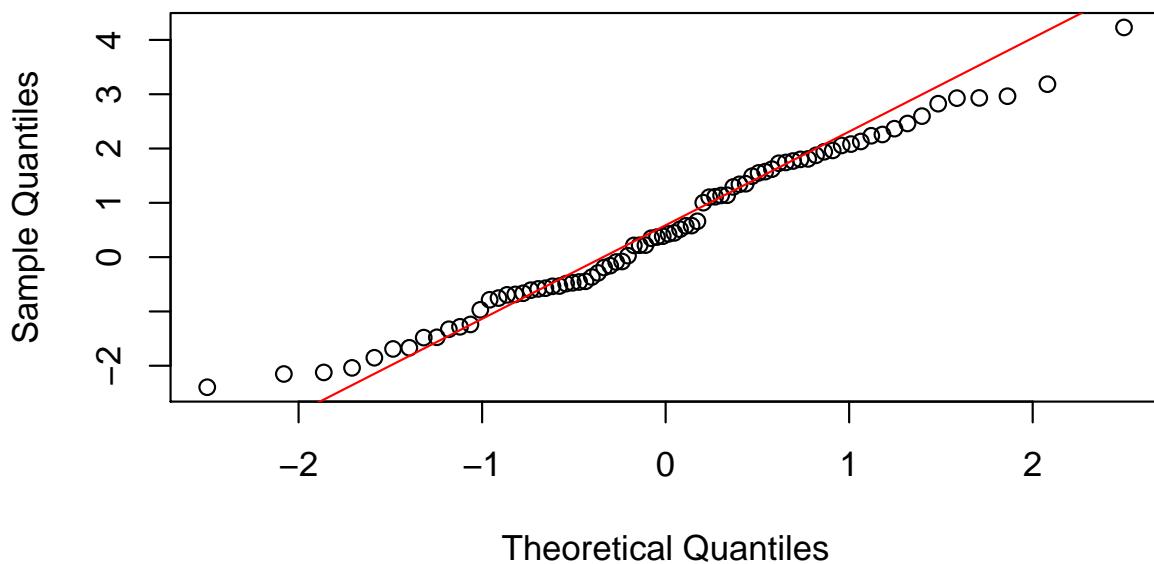
890 C.6.2 QQ-plots for Residuals and Random Effects

```
plot(hdbic, type = "QQranef", newx = admixed$xtrain, newy = admixed$ytrain)
```

QQ-Plot of the random effects at lambda = 0.10

891

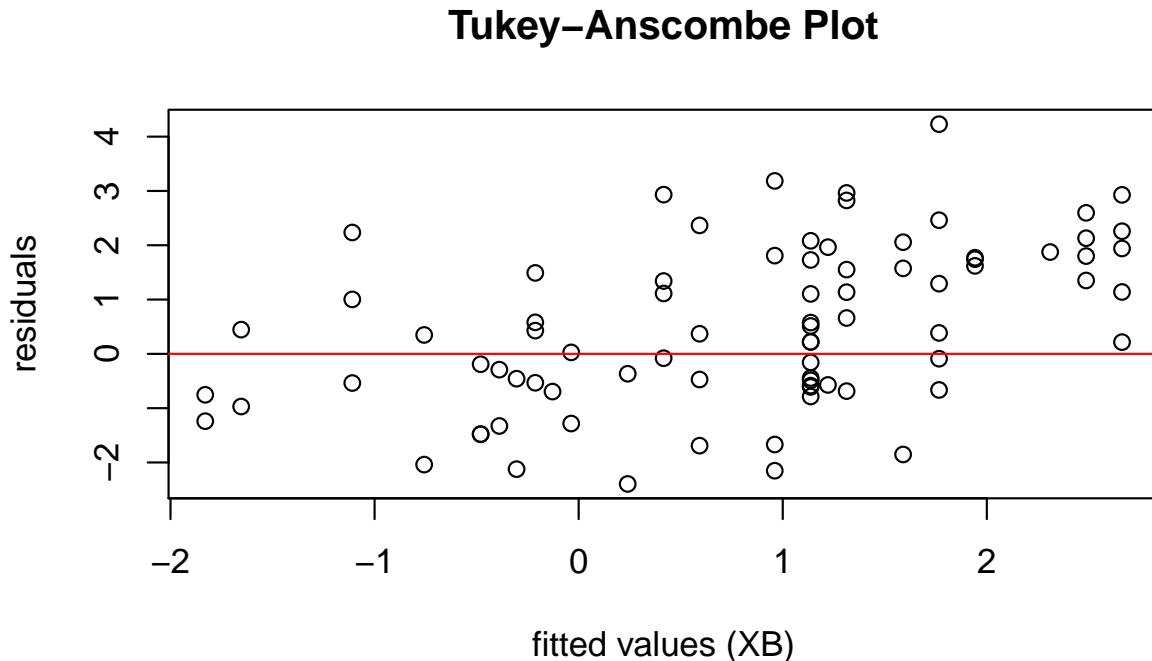
```
plot(hdbic, type = "QQresid", newx = admixed$xtrain, newy = admixed$ytrain)
```

QQ-Plot of the residuals at lambda = 0.10

892

893 C.6.3 Tukey-Anscombe Plot

```
plot(hdbic, type = "Tukey", newx = admixed$xtrain, newy = admixed$ytrain)
```



894