

Hierarchical selection of genetic and gene by environment interaction effects in high-dimensional mixed models

Statistical Methods in Medical Research
1–19

© The Author(s) 2024



Article reuse guidelines:

sagepub.com/journals-permissions

DOI: 10.1177/09622802241293768

journals.sagepub.com/home/smm

Julien St-Pierre¹ , Karim Oualkacha² and Sahir Rai Bhatnagar¹

Abstract

Interactions between genes and environmental factors may play a key role in the etiology of many common disorders. Several regularized generalized linear models have been proposed for hierarchical selection of gene by environment interaction effects, where a gene-environment interaction effect is selected only if the corresponding genetic main effect is also selected in the model. However, none of these methods allow to include random effects to account for population structure, subject relatedness and shared environmental exposure. In this article, we develop a unified approach based on regularized penalized quasi-likelihood estimation to perform hierarchical selection of gene-environment interaction effects in sparse regularized mixed models. We compare the selection and prediction accuracy of our proposed model with existing methods through simulations under the presence of population structure and shared environmental exposure. We show that for all simulation scenarios, including and additional random effect to account for the shared environmental exposure reduces the false positive rate and false discovery rate of our proposed method for selection of both gene-environment interaction and main effects. Using the F_1 score as a balanced measure of the false discovery rate and true positive rate, we further show that in the hierarchical simulation scenarios, our method outperforms other methods for retrieving important gene-environment interaction effects. Finally, we apply our method to a real data application using the Orofacial Pain: Prospective Evaluation and Risk Assessment (OPPERA) study, and found that our method retrieves previously reported significant loci.

Keywords

Gene by environment interaction, hierarchical variable selection, mixed effects model, high-dimensional data, statistical genetics

1 Introduction

Genome-wide association studies (GWASs) have led to the identification of hundreds of common genetic variants, or single nucleotide polymorphisms (SNPs), associated with complex traits¹ and are typically conducted by testing association on each SNP independently. However, these studies are plagued with the multiple testing burden that limits discovery of potentially important predictors, as genome-wide significance p -value threshold of 5×10^{-8} has become the standard. Moreover, GWASs have brought to light the problem of missing heritability, that is, identified variants only explain a low fraction of the total observed variability for traits under study.² Beyond the identified genetic variants, interactions between

¹Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, QC, Canada

²Département de Mathématiques, Faculté des Sciences, Université du Québec à Montréal, Montreal, QC, Canada

Corresponding author:

Julien St-Pierre, Department of Epidemiology, Biostatistics and Occupational Health, McGill University, 2001 Av. McGill College, Montreal, QC H3A 1Y7, Canada.

Email: julien.st-pierre@mail.mcgill.ca

genes and environmental factors may play a key role in the multifactorial etiology of many complex diseases that are subject to both genetic and environmental risk factors. For example, in assessing interactions between a polygenic risk score (PRS) and non-genetic risk factors for young-onset breast cancers (YOBC), Shi et al.³ showed a decreased association between the PRS and YOBC risk for women who had ever used hormonal birth control, suggesting that environmental exposure might result in risk stratification by interacting with genetic factors. Thus, there is a rising interest for discovering gene-environment interaction (GEI) effects as they are fundamental to better understand the effect of environmental factors in disease and to increase risk prediction accuracy.⁴

Several regularized generalized linear models (GLMs) have been proposed for selection of both genetic and GEI effects in genetic association studies,^{5–7} but currently no such method allows to include any random effect to account for genetic similarity between subjects. Indeed, one can control for population structure and/or closer relatedness by including in the model a polygenic random effect with variance-covariance structure proportional to a kinship or genetic similarity matrix (GSM).⁸ However, because kinship is a high-dimensional process, it cannot be fully captured by including only a few principal components (PCs) as fixed effects in the model.⁹ Hence, while both PC analysis (PCA) and mixed models (MMs) share the same underlying model, MMs are more robust in the sense that they do not require distinguishing between the different types of confounders.¹⁰ Moreover, MMs alleviate the need to evaluate the optimal number of PCs to retain in the model as fixed effects.

Except for normal responses, the joint estimation of variance components and fixed effects in regularized models is challenging both from a computational and analytical point of view, as the marginal likelihood for a generalized linear mixed model (GLMM) has no analytical form. To address these challenges, penalized quasi-likelihood (PQL) estimation is conceptually attractive as under this method, random effects can be treated as fixed effects, which allows to perform regularized estimation of both fixed and random effects as in the GLM framework. The computational efficiency of multivariable methods for high-dimensional MMs rely on performing a single spectral or Cholesky decomposition of the covariance matrix to rotate the phenotype and design matrix such that the transformed data become uncorrelated. For very large sample sizes, computing these decompositions can be very burdensome, with complexity of $O(n^3)$, where n is the sample size. Secondly, to obtain regularized estimates for the genetic predictors and GEI effects in linear mixed models (LMMs), we need to perform matrix multiplications with complexity of $O(n^2)$ and $O(n^2p)$ to rotate the phenotype and genotype matrices respectively, with p the number of genetic predictors. Even for moderately small cohorts, the number of predictors in GWAS is often greater than 1 million, such that the genotype matrix itself will require around one terabyte of space to be loaded in memory in a normal double-precision format.¹¹ In PQL regularized models, by minimizing the objective function with respect to the fixed effects vector only, we need not rotate the genotype matrix as we are conditioning on the random effects vector estimate.

Several authors have proposed to combine PQL estimation in presence of sparsity by inducing regularization to perform joint selection of fixed and/or random effects in multivariable GLMMs.^{12–14} However, these methods were not developed to specifically address selection of GEI effects. Although it is possible to perform naive selection of fixed and GEI effects by simply considering interaction terms as additional predictors, the aforementioned methods are not tailored to perform hierarchical selection, where interaction terms are only allowed to be selected if their corresponding main effects are active (i.e. non-zero) in the model.¹⁵ Hierarchical variable selection of GEI effects is appealing both for increasing statistical power¹⁶ and for enhancing model interpretability because interaction terms that have large main effects are more likely to be retained in the model.

Population structure and closer relatedness may also cause dependence between gene and environment, leading to selection of spurious GEI effects.¹⁷ In the context of GWAS, Sul et al.¹⁸ showed that under the polygenic model, ignoring this dependence may largely increase the false positive rate of GEI statistics. They proposed introducing an additional random effect that captures the similarity of individuals due to polygenic GEI effects to account for the fact that individuals who are genetically related and who share a common environmental exposure are more closely related. To our knowledge, the spurious selection of GEI effects in regularized models due to the dependence between gene and shared environmental exposure has not been explored yet. Thus, further work is needed to develop sparse regularized GLMMs for hierarchical selection of GEI effects in genetic association studies, while explicitly accounting for the complex correlation structure between individuals that arises from both genetic and environmental factors.

In this article, we develop a unified approach based on regularized PQL estimation to perform hierarchical selection of GEI effects in sparse regularized logistic mixed models. Similar to Sul et al.,¹⁸ we use a random effect that captures population structure and closer relatedness through a genetic kinship matrix, and shared environmental exposure through a GxE kinship matrix. We propose to use a composite absolute penalty (CAP) for hierarchical variable selection¹⁹ to seek a sparse subset of genetic and GEI effects that gives an adequate fit to the data. We derive a proximal Newton-type algorithm with block coordinate descent for PQL estimation with mixed lasso and group lasso penalties, relying on our previous work to address computational challenges associated with regularized PQL estimation in high-dimensional data.¹² We

compare the prediction and selection accuracy of our proposed model with existing methods through simulations under the presence of population structure and environmental exposure. Finally, we also apply our method to a real data application using the Orofacial Pain: Prospective Evaluation and Risk Assessment (OPPERA) study cohort²⁰ to study the sex-specific association between temporomandibular disorder (TMD) and genetic predictors.

2 Methodology

2.1 Model

We have the following GLMM

$$g(\mu_i) = \eta_i = \mathbf{Z}_i \boldsymbol{\theta} + D_i \alpha + \mathbf{G}_i \boldsymbol{\beta} + (D_i \mathbf{G}_i) \boldsymbol{\gamma} + b_i \quad (1)$$

for $i = 1, \dots, n$, where $\mu_i = \mathbb{E}(y_i | \mathbf{Z}_i, \mathbf{G}_i, D_i, b_i)$, \mathbf{Z}_i is a $1 \times m$ row vector of covariates for subject i , \mathbf{G}_i is a $1 \times p$ row vector of genotypes for subject i taking values $\{0, 1, 2\}$ as the number of copies of the minor allele, $(\boldsymbol{\theta}^\top, \boldsymbol{\beta}^\top)^\top$ is a $(m + p) \times 1$ column vector of fixed covariate and additive genotype effects including the intercept, D_i is the exposure of individual i to a binary or continuous environmental factor D with fixed effect α , and $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \dots, \gamma_p)^\top \in \mathbb{R}^p$ is the vector of fixed GEI effects. Thus, we have a total of $2p + m + 1$ coefficients. We assume that $\mathbf{b} = (b_1, \dots, b_n)^\top \sim \mathcal{N}(0, \tau_g \mathbf{K} + \tau_d \mathbf{K}^D)$ is an $n \times 1$ column vector of random effects, with $\boldsymbol{\tau} = (\tau_g, \tau_d)^\top$ the variance components that account for the relatedness between individuals. \mathbf{K} is a known GSM or kinship matrix and \mathbf{K}^D is an additional kinship matrix that describes how individuals are related both genetically and environmentally, because a pair of individuals who are genetically related and share the same environment exposure have a non-zero kinship coefficient. The kinship matrix \mathbf{K}^D corrects for the spurious association of GEI effects due to population structure and subjects relatedness, in the same way that the kinship matrix \mathbf{K} corrects for population structure and subjects relatedness on the main effects. Thus, the matrix \mathbf{K}^D can be interpreted as the covariance matrix between individuals that captures the residual variance explained by the sum of many small GEI effects across the genome. For a binary exposure, we define $K_{ij}^D = K_{ij}$ if $D_i = D_j$, and $K_{ij}^D = 0$ otherwise. For a continuous exposure, one possibility is to set $K_{ij}^D = K_{ij}(1 - d(D_i, D_j))$, where d is a metric with range $[0, 1]$. The phenotypes y_i 's are assumed to be conditionally independent and identically distributed given $(\mathbf{Z}_i, \mathbf{G}_i, D_i, \mathbf{b})$ and follow any exponential family distribution with canonical link function $g(\cdot)$, mean $\mathbb{E}(y_i | \mathbf{Z}_i, \mathbf{G}_i, D_i, \mathbf{b}) = \mu_i$ and variance $\text{Var}(y_i | \mathbf{Z}_i, \mathbf{G}_i, D_i, \mathbf{b}) = \phi a_i^{-1} v(\mu_i)$, where ϕ is a dispersion parameter, a_i are known weights, and $v(\cdot)$ is the variance function.

2.2 Regularized PQL estimation

In order to estimate the model parameters and perform variable selection, we use an approximation method to obtain an analytical closed form for the marginal likelihood of model (1). We propose to fit (1) using a PQL method,^{12,21} from where the log integrated quasi-likelihood function is equal to

$$\ell_{PQL}(\boldsymbol{\Theta}, \phi, \boldsymbol{\tau}; \tilde{\mathbf{b}}) = -\frac{1}{2} \log \left| (\tau_g \mathbf{K} + \tau_d \mathbf{K}^D) \mathbf{W} + \mathbf{I}_n \right| + \sum_{i=1}^n ql_i(\boldsymbol{\Theta}; \tilde{\mathbf{b}}) - \frac{1}{2} \tilde{\mathbf{b}}^\top (\tau_g \mathbf{K} + \tau_d \mathbf{K}^D)^{-1} \tilde{\mathbf{b}} \quad (2)$$

where $\boldsymbol{\Theta} = (\boldsymbol{\theta}^\top, \alpha, \boldsymbol{\beta}^\top, \boldsymbol{\gamma}^\top)^\top$, $\mathbf{W} = \phi^{-1} \boldsymbol{\Delta}^{-1} = \phi^{-1} \text{diag} \left\{ \frac{a_i}{v(\mu_i) [g'(\mu_i)]^2} \right\}$ is a diagonal matrix containing weights for each observation, $ql_i(\boldsymbol{\Theta}; \mathbf{b}) = \int_{y_i}^{\mu_i} \frac{a_i(y_i - \mu)}{\phi v(\mu)} d\mu$ is the quasi-likelihood for the i th individual given the random effects \mathbf{b} , and $\tilde{\mathbf{b}}$ is the solution which maximizes $\sum_{i=1}^n ql_i(\boldsymbol{\Theta}; \mathbf{b}) - \frac{1}{2} \mathbf{b}^\top (\tau_g \mathbf{K} + \tau_d \mathbf{K}^D)^{-1} \mathbf{b}$.

In typical genome-wide studies, the number of genetic predictors is much greater than the number of observations ($p > n$), and the fixed effects parameter vector $\boldsymbol{\Theta}$ becomes underdetermined when modeling p SNPs jointly. Moreover, we would like to induce a hierarchical structure, that is, a GEI effect can be present only if both exposure and genetic main effects are also included in the model. Thus, we propose to add a sparse group lasso penalty²² to the negative quasi-likelihood function in (2) to seek a sparse subset of genetic and GEI effects that gives an adequate fit to the data. Indeed, the sparse group lasso is part of the family of CAPs that can induce hierarchical variable selection.¹⁹ We define the following objective function Q_λ which we seek to minimize with respect to $(\boldsymbol{\Theta}, \phi, \boldsymbol{\tau})$:

$$Q_\lambda(\boldsymbol{\Theta}, \phi, \boldsymbol{\tau}; \tilde{\mathbf{b}}) := -\ell_{PQL}(\boldsymbol{\Theta}, \phi, \boldsymbol{\tau}; \tilde{\mathbf{b}}) + (1 - \rho) \lambda \sum_j \|(\beta_j, \gamma_j)\|_2 + \rho \lambda \sum_j |\gamma_j| \quad (3)$$

where $\lambda > 0$ controls the strength of the overall regularization and $\rho \in [0, 1)$ controls the relative sparsity of the GEI effects for each SNP. In our modeling approach, we do not penalize the environmental exposure fixed effect α . Thus, a value of

$\rho = 0$ is equivalent to a group lasso penalty where we only include a predictor in the model if both its main effect β_j and GEI effect γ_j are non-zero. A value of $0 < \rho < 1$ is equivalent to a sparse group lasso penalty where main effects can be selected without their corresponding GEI effects due to the different strengths of penalization, but a GEI effect is still only included in the model if the corresponding main effect is non-zero.

2.3 Estimation of variance components

Jointly estimating the variance components τ_g, τ_d and scale parameter ϕ with the regression effects vector Θ and random effects vector \mathbf{b} is a computationally challenging non-convex optimization problem. Updates for τ_g, τ_d and ϕ based on a majorization-minimization (MM) algorithm²³ would require inverting three different $n \times n$ matrices, with complexity $O(n^3)$, at each iteration. Thus, even for moderately small sample sizes, this is not practicable for genome-wide studies. Instead, we propose a two-step method where variance components and scale parameter are estimated only once under the null association of no genetic effect, that is assuming $\beta = \gamma = \mathbf{0}$, using the average information restricted maximum likelihood (AI-REML) algorithm.^{12,24}

2.4 Spectral decomposition of the random effects covariance matrix

Given $\hat{\tau}_g, \hat{\tau}_d$, and $\hat{\phi}$ estimated under the null, spectral decomposition of the random effects covariance matrix yields

$$\begin{aligned} (\hat{\tau}_g \mathbf{K} + \hat{\tau}_d \mathbf{K}^D)^{-1} &= (\mathbf{U} \mathbf{\Lambda} \mathbf{U}^T)^{-1} \\ &= \mathbf{U} \mathbf{\Lambda}^{-1} \mathbf{U}^T \end{aligned} \quad (4)$$

where \mathbf{U} is an orthonormal matrix of eigenvectors and $\mathbf{\Lambda}$ is a diagonal matrix of eigenvalues $\Lambda_1 \geq \Lambda_2 \geq \dots \geq \Lambda_n > 0$ when both \mathbf{K} and \mathbf{K}^D are positive definite. In practice, if \mathbf{K} is rank-deficient, one can replace it by $\mathbf{K} + \epsilon \mathbf{I}_n$ for $\epsilon > 0$ small, to ensure that both \mathbf{K} and \mathbf{K}^D are positive definite.

Using (4) and assuming that the weights in \mathbf{W} vary slowly with the conditional mean,²⁵ minimizing (3) is now equivalent to

$$\begin{aligned} \hat{\Theta} &= \underset{\Theta}{\operatorname{argmin}} - \sum_{i=1}^n ql_i(\Theta; \tilde{\delta}) + \frac{1}{2} \tilde{\delta}^T \mathbf{\Lambda}^{-1} \tilde{\delta} + (1 - \rho) \lambda \sum_j \|(\beta_j, \gamma_j)\|_2 + \rho \lambda \sum_j |\gamma_j| \\ &= \underset{\Theta}{\operatorname{argmin}} f(\Theta; \tilde{\delta}) + g(\Theta) \end{aligned} \quad (5)$$

where $\tilde{\delta} = \mathbf{U}^T \tilde{\mathbf{b}}$ is the minimizer of $f(\Theta; \delta) := - \sum_{i=1}^n ql_i(\Theta; \delta) + \frac{1}{2} \delta^T \mathbf{\Lambda}^{-1} \delta$. Thus, iteratively solving (5) also requires updating the solution $\tilde{\delta}$ at each step until convergence. Conditioning on the previous solution for Θ , $\tilde{\delta}$ is obtained by minimizing a generalized ridge weighted least-squares (WLS) problem with $\mathbf{\Lambda}^{-1}$ as the regularization matrix. Then, conditioning on $\tilde{\delta}$, $\hat{\Theta}$ is found by minimizing a WLS problem with a sparse group lasso penalty. We present in Appendix A our proposed proximal Newton-type algorithm that cycles through updates of $\tilde{\delta}$ and Θ .

3 Simulation study

We first evaluated the performance of our proposed method, called *pglmm*, against that of a standard logistic lasso, using the Julia package *GLMNet* which wraps the Fortran code from the original R package *glmnet*.²⁶ Then, among logistic models that impose hierarchical interactions, we compared our method with the *glinternet*⁷ and *gesso*⁶ models which are both implemented in R packages. The *glinternet* method relies on overlapping group lasso, and even though it is optimized for selection of gene by gene interactions in high-dimensional data, it is applicable for selection of GEI effects. An advantage of the method is that it only requires tuning a single parameter value. On the other hand, *gesso* uses a CAP penalty with a group L_∞ norm (iCAP) to induce a hierarchical structure, and the default implementation fits solutions paths across a two-dimensional grid of tuning parameter values. For all methods, selection of the tuning parameters is performed by cross-validation. The default implementation for *glmnet*, *glinternet* and *pglmm* is to find the smallest value of the tuning parameter λ such that no predictor are selected in the model, and then to solve the penalized minimization problem over a grid of decreasing values of λ . For these three methods, we used a grid of 50 values of λ on the log10 scale with $\lambda_{\min} = 0.01 \lambda_{\max}$, where λ_{\max} is chosen such that no predictors are selected in the model. In addition, for *pglmm*, we solved the penalized minimization problem over a grid of 10 values of the tuning parameter ρ evenly spaced from 0 to 0.9,

Table 1. Number of samples by population for the high quality harmonized set of 4097 whole genomes from the Human Genome Diversity Project (HGDP) and the 1000 Genomes Project (1000 G).

Population	1000 genomes	HGDP	Total
African	879 (28%)	110 (12%)	989 (24%)
Admixed American	487 (15%)	62 (7%)	549 (13%)
Central/South Asian	599 (19%)	184 (20%)	783 (19%)
East Asian	583 (18%)	234 (25%)	817 (20%)
European	618 (20%)	153 (16%)	771 (19%)
Middle Eastern	0	158 (17%)	158 (4%)
Oceania	0	30 (3%)	30 (1%)
Total	3166	931	4097
Unrelated individuals	2520	880	3400

fitting a total of 500 models. The default implementation for *gesso* is to solve the minimization problem over a 20×20 two-dimensional grid of the tuning parameters values λ_1, λ_2 , starting from the smallest value such that all coefficients are zero, and setting $\lambda_{min} = 0.1\lambda_{max}$. Finally, for *glmnet*, *gesso*, and *glinternet*, population structure and environmental exposure is accounted for by adding the top 10 PCs of the kinship matrix as additional covariates.

3.1 Simulation model

We performed a total of 100 replications for each of our simulation scenarios, drawing anew genotypes and simulated traits, using real genotype data from a high quality harmonized set of 4097 whole genomes from the Human Genome Diversity Project (HGDP) and the 1000 Genomes Project (1000 G).²⁷ At each replication, we sampled 10,000 candidate SNPs from the chromosome 21 and randomly selected 100 (1%) to be causal. Let S be the set of candidate causal SNPs, with $|S| = 100$, then the causal SNPs fixed effects β_j were generated from a Gaussian distribution $\mathcal{N}(0, h_S^2 \sigma^2 / |S|)$, where h_S^2 is the fraction of variance on the logit scale that is due to total additive genetic fixed effects. Let S' be the set of candidate causal SNPs, not necessarily overlapping with S , that have a non-zero GEI effect, with $|S'| = 50$, then the GEI effects γ_j were generated from a Gaussian distribution $\mathcal{N}(0, h_{S'}^2 \sigma^2 / |S'|)$, where $h_{S'}^2$ is the fraction of variance on the logit scale that is due to total additive GEI fixed effects. Further, we simulated a random effect from a Gaussian distribution $\epsilon \sim \mathcal{N}(0, h_g^2 \sigma^2 \mathbf{K} + h_d^2 \sigma^2 \mathbf{K}^D)$, where h_g^2 and h_d^2 are the fractions of variance explained by the polygenic and polygenic by environment effects, respectively. The kinship matrices \mathbf{K} and \mathbf{K}^D were calculated using a set of 50,000 randomly sampled SNPs excluding the set of candidate SNPs, and PCs were obtained from the singular value decomposition of \mathbf{K} . We simulated a covariate for age using a Normal distribution and used the sex covariate provided with the data as a proxy for environmental exposure. Then, for $i = 1, \dots, 4097$, binary phenotypes were generated using the following model:

$$\text{logit}(\pi) = \text{logit}(\pi_{0k}) - \log(1.3) \times \text{Sex} + \log(1.05) \text{Age} / 10 + \sum_{j \in S} \beta_j \tilde{G}_j + \sum_{j \in S'} \gamma_j \cdot (\text{Sex} \times \tilde{G}_j) + \epsilon \quad (6)$$

where π_{0k} , for $k = 1, \dots, 7$, was simulated using a $U(0.1, 0.9)$ distribution to specify a different prevalence for each population in Table 1 under the null, and \tilde{G}_j is the j th column of the standardized genotype matrix $\tilde{g}_{ij} = (g_{ij} - 2p_i) / \sqrt{2p_i(1 - p_i)}$ and p_j is the minor allele frequency (MAF) for the j th predictor.

In all simulation scenarios, we set $h_S^2 = 0.2$ and $h_{S'}^2 = 0.1$ such that each of the main effects ($|S| = 100$) or GEI effects ($|S'| = 50$) explains 0.2% of the total variability on the logit scale. We compared the methods when $h_g^2 = 0.2$ and $h_d^2 = 0.1$ (i.e. low polygenic effects with $\sigma^2 = 9$), and when $h_g^2 = 0.4$ and $h_d^2 = 0.2$ (i.e. high polygenic effects with $\sigma^2 = 35$), respectively. In the first simulation scenario, we induced a hierarchical structure for the simulated data by imposing $\gamma_j \neq 0 \rightarrow \beta_j \neq 0$ for $j = 1, \dots, p$, such that the total number of causal SNPs is equal to 100, with half of them having non-zero GEI effects. In the second simulation scenario, we repeated the simulations from the first scenario, but without enforcing any hierarchical structure, such that the number of causal SNPs is equal to 150, with 100 of them having non-zero main effects, and 50 having non-zero GEI effects.

3.2 Metrics

To compare the performance of all methods in discovering important genetic predictors and estimating their main and interaction effects, we define in this section the performance metrics that will be used. First, we define the model size

as simply the number of non-zero coefficients estimated by a model, that is, $\sum_{j=1}^p I(\hat{\beta}_j \neq 0)$ for the main effects, and $\sum_{j=1}^p I(\hat{\gamma}_j \neq 0)$ for the GEI effects. The false positive rate (FPR) is defined as the number of non-causal predictors that are falsely identified as causal (false positives), divided by the total number of non-causal predictors. The true positive rate (TPR), also known as sensitivity or recall, is defined as the number of true causal predictors that are correctly identified (true positives), divided by the total number of causal predictors. The false discovery rate (FDR) is defined as the number of false positives divided by the total number of selected predictors in the model. Thus, while FPR and TPR measure the ability of a model to distinguish between causal and non-causal predictors, the FDR actually measures the proportion of predictors that are not causal among those declared significant. Moreover, in genetic association studies where the number of non-causal predictors is very high, we are more interested in controlling the FDR rather than the FPR. Alternatively, we can define the precision as 1 minus the FDR, which measures the proportion of causal predictors among those declared significant. The F_1 score is defined as the harmonic mean of the precision and TPR, and it can be used to take into account that methods with a large number of selected predictors will likely have a higher TPR, and inversely that methods with a lower number of selected predictors will likely have a higher precision. Finally, the area under the curve (AUC) is used as a measure of the predictive performance of all methods when predicting the binary status of individuals. It takes into account the TPR of all methods at various FPR values when making individual predictions. A higher AUC means that a method has a better capacity at distinguishing between cases and controls.

3.3 Results

We obtained solutions paths across a one dimensional (*glmnet*, *glinternet*) or two-dimensional grid of tuning parameter values (*gesso*, *pglmm*) for the hierarchical and non-hierarchical simulation scenarios and reported the mean precision, that is, the proportion of selected predictors that are causal, over 100 replications for the selection of GEI effects (Figure 1) and main genetic effects (Figure 2) respectively. We see from Figure 1 that in the hierarchical simulation scenario, *gesso* and *pglmm* retrieve important GEI effects with better precision than *glmnet* and *glinternet*. When we simulate a low random polygenic GEI effect, *gesso* slightly outperforms *pglmm*, but when we increase the heritability of the two random effects, both methods perform similarly. When we simulate data under no hierarchical assumption, precision for all hierarchical models fall drastically, although they still perform better than the standard lasso model. We note that *gesso* retrieves important GEI effects with equal or better precision than other methods in all simulation settings. This is explained by the fact that *gesso* is using a CAP penalty with L_∞ group norm which has been shown to perform better than the sparse group lasso for retrieving interaction effects.¹⁹ On the other hand, we see from Figure 2 that *pglmm* outperforms all methods for retrieving important main effects for both hierarchical and non-hierarchical simulation scenarios. When we simulate low polygenic effects, *pglmm* and *glmnet* perform comparably. We also note that *gesso* retrieves main effects with less precision than *glmnet* and *pglmm* in all scenarios. At last, the precision of *glinternet* is considerably lower than all other methods until the number of selected main genetic effects in the model is large.

In practice, we often do not have any a priori knowledge for the number of main effects and/or GEI effects that we want to include in the final model. Thus, instead of comparing methods at a fixed number of selected predictors along their regularization paths, we used cross-validation to compare how each method performs when having to select an optimal number of predictors in the model for the same two simulation scenarios that we previously described. We randomly split the data ($n = 4097$) into training and test subjects, using a 80/20 ratio, and fitted the full lasso solution path on the training set for 100 replications. We report the model size, FPR, TPR, FDR, and F_1 score on the training sets, and the area under the ROC curve (AUC) when making predictions on the independent test subjects. To assess the potential spurious association of both main and GEI effects due to shared environmental exposure, we compare our method when including only the kinship matrix \mathbf{K} (*plmm* (1 Random effect (RE))) and when including both \mathbf{K} and \mathbf{K}^D matrices (*pglmm* (2 REs)).

With respect to selection of the GEI effects (Table 2), the comparative performance of each method varies depending on the simulation scenario. As expected, we see that including an additional random effect reduces the FPR for all simulation scenarios for our proposed method. Unsurprisingly, *glinternet* and *glmnet* have the lowest FPRs of all methods since they always select the least number of GEI effects in the final models. Consequently, they have the smallest TPRs in all scenarios. By using the F_1 score to account for the trade-off between FDR and TPR, we have that *pglmm* performs the best in hierarchical simulation scenarios, while *gesso* performs better in the non-hierarchical scenarios.

With respect to the genetic main effects (Table 3), *pglmm* selects the lowest number of predictors in the model, and thus has the lowest FPR and FDR in all simulation scenarios. Again, adding an additional random effect reduces the FPR for *pglmm*, but to a lower extent than for selection of GEI effects. On the other hand, *glinternet* always selects the largest number of predictors in all scenarios, and hence has the highest TPR and FPR values. Using the F_1 score to balance FDR and TPR, we see that *pglmm* performs the best for retrieving the important main genetic effects in all simulation scenarios. Also, we see that *gesso* and *pglmm* perform similarly when the heritability of the polygenic random effects is

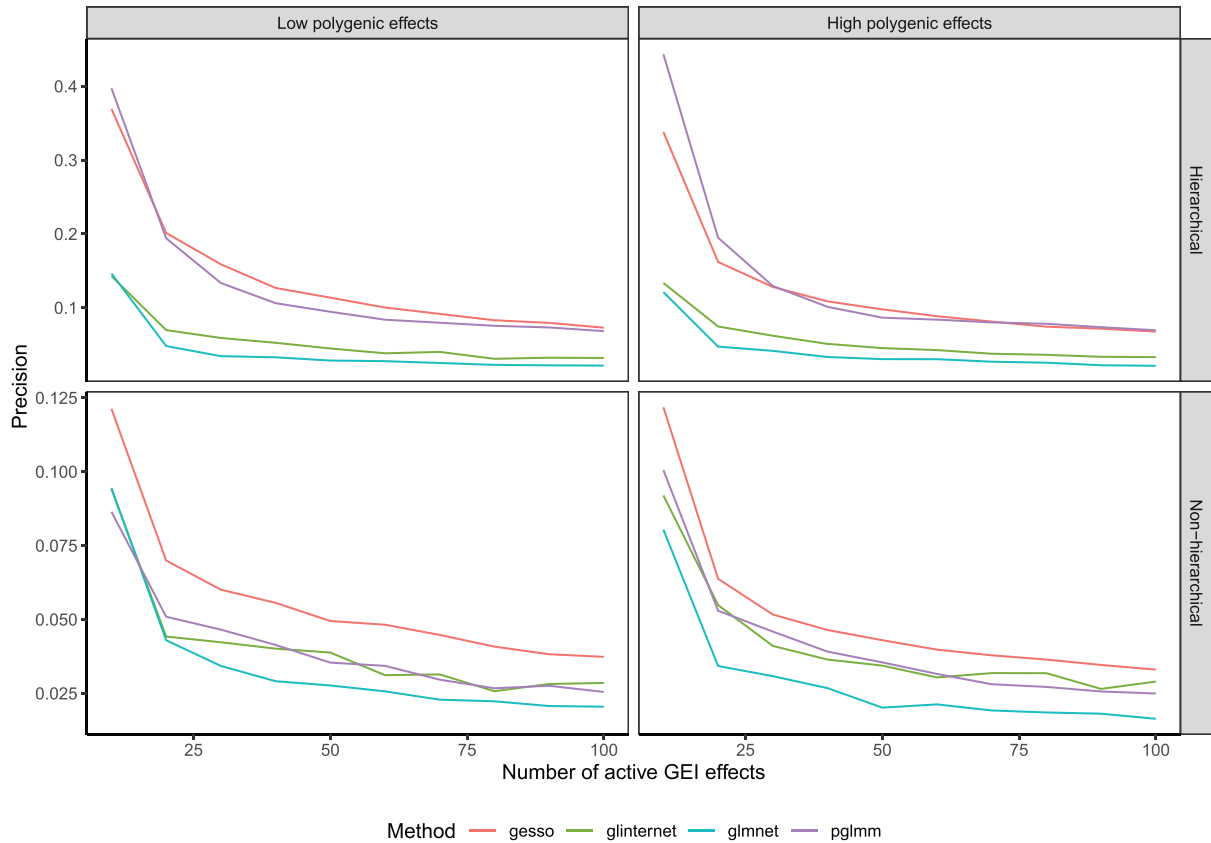


Figure 1. Precision of compared methods averaged over 100 replications as a function of the number of active gene-environment interaction (GEI) effects.

low, but when we increase the heritability, the FDR for *gesso* increases drastically, and the number of selected main effects becomes on average more than 1.5 times higher than for *pglmm*. Results for the accuracy of predicting binary outcomes in independent test sets are included in Table 4. We see that *pglmm* with two random effects outperforms all other methods for all simulation scenarios.

4 Discovering sex-specific genetic predictors of painful temporomandibular disorder

Significant associations between temporomandibular disorder (TMD), which is a painful disease of the jaw, and four distinct loci have been previously reported in combined or sex-segregated analyses on the OPPERA study cohort.²⁸ Moreover, TMD has much greater prevalence in females than in males and is believed to have some sex-specific pathophysiologic mechanisms.²⁹ In this analysis, we wanted to explore the comparative performance of our proposed method *pglmm* in selecting important sex-specific predictors of TMD and its performance predicting the risk of painful TMD in independent subjects from two replication cohorts, the OPPERA II Chronic TMD Replication case-control study, and the Complex Persistent Pain Conditions (CPPC): Unique and Shared Pathways of Vulnerability study, using the OPPERA cohort as discovery cohort. Sample sizes and distribution of sex, cases and ancestry for the three studies are shown in Table 5, and further details on study design, recruitment, subject characteristics, and phenotyping for each study are provided in the Supplemental Material of Smith et al.²⁸ (available at <http://links.lww.com/PAIN/A688>).

We used the imputed data described by Smith et al.²⁸ Genotypes were imputed to the 1000 Genomes Project phase 3 reference panel using the software packages SHAPEIT³⁰ for prephasing and IMPUTE version 2.³¹ For each cohort independently, we assessed imputation quality taking into account the number of minor alleles as well as the information score such that a SNP with rare MAF must pass a higher quality information threshold for inclusion. After merging all three cohorts, we tested for significant deviations of the Hardy-Weinberg equilibrium (HWE) separately in cases and controls, using a more strict *p*-value threshold for hypothesis testing among cases to avoid discarding disease-associated SNPs that are possibly under selection³² ($< 10^{-6}$ in controls, $< 10^{-11}$ in cases). We filtered using a SNP call rate $> 95\%$ on the

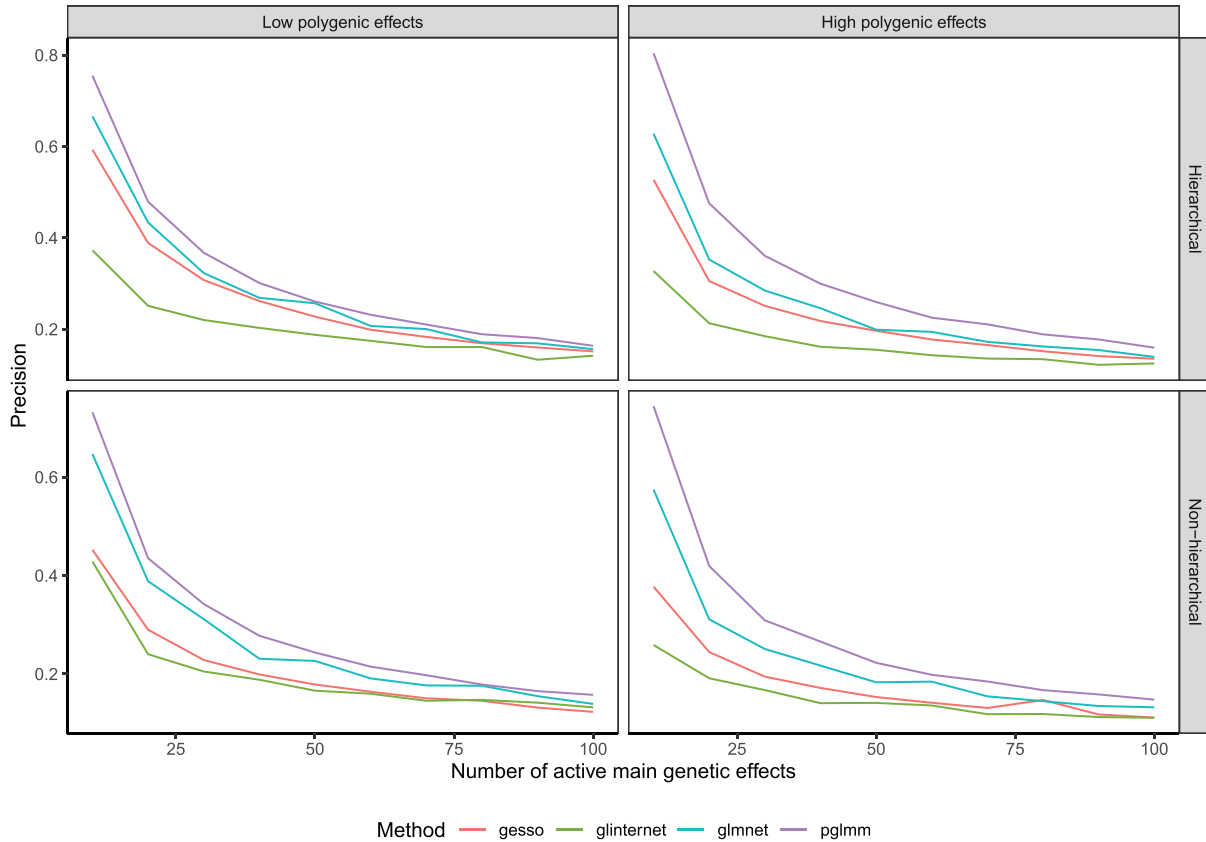


Figure 2. Precision of compared methods averaged over 100 replications as a function of the number of active main effects in the model.

combined dataset to retain imputed variants present in all cohorts, which resulted in a total of 4.8 million imputed SNPs. PCs and kinship matrices were calculated on the merged genotype data using the `-pca` and `-make-re1` flags in PLINK,³³ after using the same HWE p -value threshold and SNP call rate as for the imputed data. To reduce the number of candidate predictors in the regularized models, we performed a first screening by testing genome-wide association with TMD for subjects in the OPPERA discovery cohort using PLINK. We fitted a logistic regression for additive SNP effects, with age, sex, and enrollment site as covariates and the first 10 PCs to account for population stratification, and retained all SNPs with a p -value below 0.05, which resulted in a total of 243 thousand predictors.

We present in Table 6 the estimated odds ratios (OR) by each method, *pglmm*, *gesso*, and *glmnet*, for the selected SNPs for both main and GEI effects. Of note, it was not possible to use the *glinetnet* package due to computational considerations, its memory requirement being too large for the joint analysis of the 243 thousand preselected predictors. All three methods selected the imputed insertion/deletion (indel) polymorphism on chromosome 4 at position 146,211,844 (rs5862730), which was the only reported SNP that reached genome-wide significance in the full OPPERA cohort ($OR = 1.4$, 95% confidence interval (CI): [1.26; 1.61], $P = 2.82 \times 10^{-8}$).²⁸ In a females-only analysis, rs5862730 was likewise associated with TMD ($OR = 1.54$, 95% CI: [1.33; 1.79], $P = 1.7 \times 10^{-8}$), and both *pglmm* and *gesso* selected the GEI term between rs5862730 and sex.

Moreover, we present in Table 7, the AUC in the training and test cohorts, the number of predictors selected in each model and the total computation time to fit each method. We see that *pglmm* has the highest AUC on the training data, as well as the best predictive performance on the CPPC cohort alone. On the other hand, *glmnet* and *gesso* both have a greater predictive performance in the OPPERA2 cohort compared to *pglmm*. When combining the predictions for OPPERA2 and CPPC cohorts, all three methods have similar predictive performance. In term of the number of predictors selected by each model, *glmnet* has selected two SNPs with important main effects and no GEI effects, while *gesso* has selected the highest number of predictors, that is a total of 13 SNPs with both main and GEI effects. On the other hand, our proposed method *pglmm* has selected a total of seven SNPs, among which three had a selected GEI effect with sex. Finally, we report for each method the computational time to fit the model on the training cohort using 10-folds cross-validation.

Table 2. Results for the GEI effects γ .

Metric	Method	Non-hierarchical model		Hierarchical model	
		Low ϵ	High ϵ	Low ϵ	High ϵ
Model size	<i>pglm</i> (1 RE)	84.6	99.2	95.5	103
	<i>pglm</i> (2 REs)	58.4	57.8	63.9	59.8
	<i>glmnet</i>	21.6	38.8	22.7	38.6
	<i>glinternet</i>	17.5	39.2	19.5	39.6
	<i>gesso</i>	64.3	102	78.6	110
FPR	<i>pglm</i> (1 RE)	8.31×10^{-3}	9.76×10^{-3}	8.96×10^{-3}	9.69×10^{-3}
	<i>pglm</i> (2 REs)	5.72×10^{-3}	5.65×10^{-3}	6.00×10^{-3}	5.55×10^{-3}
	<i>glmnet</i>	2.09×10^{-3}	3.80×10^{-3}	2.20×10^{-3}	3.76×10^{-3}
	<i>glinternet</i>	1.68×10^{-3}	3.79×10^{-3}	1.84×10^{-3}	3.79×10^{-3}
	<i>gesso</i>	6.20×10^{-3}	9.94×10^{-3}	7.38×10^{-3}	1.06×10^{-2}
TPR	<i>pglm</i> (1 RE)	0.039	0.043	0.126	0.124
	<i>pglm</i> (2 REs)	0.030	0.031	0.084	0.091
	<i>glmnet</i>	0.016	0.020	0.018	0.025
	<i>glinternet</i>	0.016	0.030	0.024	0.038
	<i>gesso</i>	0.052	0.068	0.104	0.103
FDR	<i>pglm</i> (1 RE)	0.966	0.965	0.926	0.908
	<i>pglm</i> (2 REs)	0.936	0.968	0.921	0.889
	<i>glmnet</i>	0.962	0.974	0.955	0.967
	<i>glinternet</i>	0.948	0.956	0.923	0.949
	<i>gesso</i>	0.945	0.948	0.912	0.930
F_1	<i>pglm</i> (1 RE)	0.035	0.033	0.091	0.084
	<i>pglm</i> (2 REs)	0.039	0.036	0.082	0.090
	<i>glmnet</i>	0.036	0.033	0.035	0.036
	<i>glinternet</i>	0.040	0.038	0.045	0.048
	<i>gesso</i>	0.052	0.048	0.083	0.067

Note: For each simulation scenario, we report the mean value over 100 replications when we simulate only one random effect with low heritability (low ϵ) and we simulate two random effects with high heritability (high ϵ). Bolded values indicate the method with the best performance according to each metric.

FPR: false positive rate; TPR: true positive rate; FDR: false discovery rate; RE: random effect.

Model size is defined as $\sum_{j=1}^p I(\hat{\gamma}_j \neq 0)$.

FPR is defined as $\sum_{j=1}^p I(\hat{\gamma}_j \neq 0 \cap \gamma_j = 0) / \sum_{j=1}^p I(\gamma_j = 0)$.

TPR is defined as $\sum_{j=1}^p I(\hat{\gamma}_j \neq 0 \cap \gamma_j \neq 0) / \sum_{j=1}^p I(\gamma_j \neq 0)$.

FDR is defined as $\sum_{j=1}^p I(\hat{\gamma}_j \neq 0 \cap \gamma_j = 0) / \sum_{j=1}^p I(\hat{\gamma}_j \neq 0)$.

F_1 is defined as $2 \times \left(\frac{1}{1-FDR} + \frac{1}{TPR} \right)^{-1}$.

While *glmnet* only took two hours to fit, it failed to retrieve any potentially important GEI effects between TMD and sex, albeit we note that it had a similar predictive performance than the hierarchical methods on the combined test sets. On the other hand, *pglm* had the highest computational time required to fit the model, because it requires iteratively estimating a random effects vector of size $n = 3030$, while both *glmnet* and *gesso* only require to estimate a vector of fixed effects of size 10 for the PCs. However, *pglm* had the highest AUC on the train set, and was able to retrieve potentially important GEI effects for some of the select SNPs in the model, while selecting half as many predictors than *gesso*.

5 Discussion

We have developed a unified approach based on regularized PQL estimation, for selecting important predictors and GEI effects in high-dimensional GWAS data, accounting for population structure, close relatedness, shared environmental exposure and binary nature of the trait. We proposed to combine PQL estimation with a CAP for hierarchical selection of main genetic and GEI effects, and derived a proximal Newton-type algorithm with block coordinate descent to find coordinate-wise updates. We showed that for all simulation scenarios, including and additional random effect to account for the shared environmental exposure reduced the FPR of our proposed method for selection of both GEI and main effects. Using the F_1

Table 3. Results for the genetic predictors main effects β .

Metric	Method	Non-hierarchical model		Hierarchical model	
		Low ϵ	High ϵ	Low ϵ	High ϵ
Model size	<i>pglm</i> (1 RE)	227	212	220	204
	<i>pglm</i> (2 REs)	206	190	214	179
	<i>glmnet</i>	278	444	286	450
	<i>glinternet</i>	299	481	312	480
	<i>gesso</i>	212	361	224	367
FPR	<i>pglm</i> (1 RE)	2.09×10^{-2}	1.96×10^{-2}	2.01×10^{-2}	1.87×10^{-2}
	<i>pglm</i> (2 REs)	1.89×10^{-2}	1.74×10^{-2}	1.95×10^{-2}	1.62×10^{-2}
	<i>glmnet</i>	2.59×10^{-2}	4.24×10^{-2}	2.66×10^{-2}	4.29×10^{-2}
	<i>glinternet</i>	2.80×10^{-2}	4.61×10^{-2}	2.91×10^{-2}	4.57×10^{-2}
	<i>gesso</i>	1.95×10^{-2}	3.42×10^{-2}	2.05×10^{-2}	3.47×10^{-2}
TPR	<i>pglm</i> (1 RE)	0.195	0.181	0.208	0.196
	<i>pglm</i> (2 REs)	0.188	0.177	0.206	0.188
	<i>glmnet</i>	0.215	0.244	0.226	0.257
	<i>glinternet</i>	0.216	0.246	0.237	0.271
	<i>gesso</i>	0.190	0.220	0.210	0.238
FDR	<i>pglm</i> (1 RE)	0.895	0.895	0.891	0.883
	<i>pglm</i> (2 REs)	0.888	0.885	0.885	0.870
	<i>glmnet</i>	0.920	0.944	0.917	0.942
	<i>glinternet</i>	0.925	0.948	0.922	0.943
	<i>gesso</i>	0.906	0.937	0.903	0.932
F_1	<i>pglm</i> (1 RE)	0.123	0.120	0.136	0.134
	<i>pglm</i> (2 REs)	0.126	0.124	0.138	0.140
	<i>glmnet</i>	0.115	0.091	0.119	0.094
	<i>glinternet</i>	0.109	0.085	0.116	0.094
	<i>gesso</i>	0.123	0.096	0.131	0.103

Note: For each simulation scenario, we report the mean value over 100 replications when we simulate two random effects with low heritability (low ϵ) and high heritability (high ϵ). Bolded values indicate the method with the best performance according to each metric.

FPR: false positive rate; TPR: true positive rate; FDR: false discovery rate; RE: random effect.

Model size is defined as $\sum_{j=1}^p I(\hat{\beta}_j \neq 0)$.

FPR is defined as $\sum_{j=1}^p I(\hat{\beta}_j \neq 0 \cap \beta_j = 0) / \sum_{j=1}^p I(\hat{\beta}_j \neq 0)$.

TPR is defined as $\sum_{j=1}^p I(\hat{\beta}_j \neq 0 \cap \beta_j \neq 0) / \sum_{j=1}^p I(\hat{\beta}_j \neq 0)$.

FDR is defined as $\sum_{j=1}^p I(\hat{\beta}_j \neq 0 \cap \beta_j = 0) / \sum_{j=1}^p I(\hat{\beta}_j \neq 0)$.

F_1 is defined as $2 \times \left(\frac{1}{1-FDR} + \frac{1}{TPR} \right)^{-1}$.

Table 4. Results for the prediction accuracy of a binary outcome on test sets.

Metric	Method	Non-hierarchical model		Hierarchical model	
		Low ϵ	High ϵ	Low ϵ	High ϵ
AUC	<i>pglm</i> (1 RE)	0.719	0.786	0.728	0.788
	<i>pglm</i> (2 REs)	0.723	0.790	0.730	0.792
	<i>glmnet</i>	0.688	0.753	0.695	0.751
	<i>glinternet</i>	0.702	0.760	0.710	0.761
	<i>gesso</i>	0.695	0.750	0.707	0.751

Note: For each simulation scenario, we report the mean AUC value over 100 replications when we simulate two random effects with low heritability (low ϵ) and high heritability (high ϵ). Bolded values indicate the method with the best performance according to each metric.

RE: random effect; AUC: area under the curve.

Table 5. Demographic data for the OPPERA training cohort, and for the OPPERA2 and CPPC test cohorts.

	Study name		
	OPPERA	OPPERA2	CPPC
N (% female)	3030 (64.6)	1342 (66.0)	390 (84.4)
Cases (%)	999 (33.0)	444 (33.0)	164 (42.0)
Ancestry (% white)	61	79	68

OPPERA: Orofacial Pain: Prospective Evaluation and Risk Assessment; CPPC: Complex Persistent Pain Conditions.

Table 6. Selected SNPs by each method with their estimated odds ratios (OR) for the main effects (β) and GEI effects (γ) from the TMD real data analysis.

Chromosome	Position	<i>pglmm</i>			<i>gesso</i>			<i>glmnet</i>
		OR $_{\beta}$	OR $_{\gamma}$	OR $_{\beta+\gamma}$	OR $_{\beta}$	OR $_{\gamma}$	OR $_{\beta+\gamma}$	OR $_{\beta}$
3	5,046,726	–	–	–	1.0042	1.0087	1.0129	–
3	153,536,154	1.0020	–	–	–	–	–	–
4	42,549,777	1.0068	1.0042	1.0110	1.0029	1.0060	1.0089	–
4	146,211,844	1.0252	1.0448	1.0712	1.0261	1.0553	1.0829	1.0312
11	17,086,381	1.0076	–	–	1.0014	1.0029	1.0042	–
11	132,309,606	0.9965	–	–	–	–	–	–
12	19,770,625	–	–	–	1.0045	1.0094	1.0140	–
12	47,866,802	1.0184	1.0001	1.0184	–	–	–	1.0140
12	47,870,741	–	–	–	1.0152	1.0320	1.0477	–
14	24,345,235	1.0013	–	–	–	–	–	–
16	81,155,867	–	–	–	1.0039	1.0082	1.0122	–
17	46,592,346	–	–	–	1.0025	1.0052	1.0077	–
17	52,888,414	–	–	–	1.0005	1.0011	1.0017	–
17	69,061,947	–	–	–	1.0021	1.0043	1.0064	–
18	36,210,549	–	–	–	1.0186	1.0392	1.0585	–
19	37,070,882	–	–	–	1.0020	1.0042	1.0062	–
21	32,760,615	–	–	–	1.0051	1.0107	1.0159	–

Note: All three methods selected the imputed insertion/deletion (indel) polymorphism on chromosome 4 at position 146,211,844 (rs5862730), which was the only reported SNP that reached genome-wide significance in the full OPPERA cohort.

SNPs: single nucleotide polymorphisms; GEI: gene-environment interaction; TMD: temporomandibular disorder; OPPERA: Orofacial Pain: Prospective Evaluation and Risk Assessment.

Table 7. Area under the ROC curve (AUC), model size and computational time for the analysis of TMD.

Method	AUC $_{train}$	AUC $_{test}$			Model size		
	OPPERA	OPPERA2	CPPC	OPPERA2+CPPC	Main effects	GEI effects	Computational time (hours)
<i>glmnet</i>	0.722	0.587	0.632	0.551	2	0	2
<i>gesso</i>	0.725	0.586	0.630	0.551	13	13	9
<i>pglmm</i>	0.867	0.512	0.652	0.550	7	3	47

CPPC: Complex Persistent Pain Conditions; GEI: gene-environment interaction; TMD: temporomandibular disorder; OPPERA: Orofacial Pain: Prospective Evaluation and Risk Assessment.

score as a balanced measure of the FDR and TPR, we showed that in the hierarchical simulation scenarios, *pglmm* outperformed all other methods for retrieving important GEI effects. Moreover, using real data from the OPPERA study to explore the comparative performance of our method in selecting important predictors of TMD, we found that our proposed method was able to retrieve a previously reported significant loci in a combined or sex-segregated GWAS.

A limitation of *pglmm* compared to a logistic lasso or group lasso with PC adjustment is the computational cost of performing multiple matrix calculations that comes from incorporating a GSM to account for population structure and relatedness between individuals. These computations become prohibitive when the sample size increases, and this may hinder the use of random effects in hierarchical selection of both genetic and GEI fixed effects in genetic association

studies. Solutions to explore in order to increase computation speed and decrease memory usage would be the use of conjugate gradient methods with a diagonal preconditioner matrix, as proposed by Zhou et al.,³⁴ and the use of sparse GSMs to adjust for the sample relatedness.³⁵

In this study, we focused solely on the sparse group lasso as a hierarchical regularization penalty. Although previous work has shown that using a CAP penalty with a group L_∞ norm (iCAP) might perform better than a sparse group lasso penalty for retrieving important interaction terms,¹⁹ substantive work is needed to develop an efficient algorithm to fit the iCAP penalty in the presence of random effects. It is also important to highlight that for selection of main effects, the sparse group lasso penalty might perform better than the iCAP penalty. Thus, the choice of which group penalty to use should reflect this trade off between improving the selection of main effects versus selection of important GEI effects. Moreover, it is known that estimated effects by lasso will have large biases because the resulting shrinkage is constant irrespective of the magnitude of the effects. Alternative regularizations like the Smoothly Clipped Absolute Deviation (SCAD)³⁶ and Minimax Concave Penalty (MCP)³⁷ could be explored, although we note that both SCAD and MCP require tuning an additional parameter which controls the relaxation rate of the penalty. Another alternative includes refitting the sparse group lasso penalty on the active set of predictors only, similarly to the relaxed lasso, which has shown to produce sparser models with equal or lower prediction loss than the regular lasso estimator for high-dimensional data.³⁸

Another interesting question to address in the context of high-dimensional GLMMs would be to assess the goodness of fit of the selected sparse model. In the context of high-dimensional GLMs, a recent methodology has been proposed to test for any signal left in the residuals after fitting a sparse model in order to assess whether a sparse non-linear model would be more appropriate.³⁹ Although there exist graphical and numerical methods for checking the adequacy of GLMMs,⁴⁰ to our knowledge no such procedure has been extended to high-dimensional mixed models. Finally, it would also be of interest to explore if joint selection of fixed and random effects could result in better selection and/or predictive performance. Future work includes tuning the generalized ridge regularization on the random effects,⁴¹ or replacing it by a lasso regularization to perform selection of individual random effects.^{14,42}

Acknowledgements

This study was enabled in part by support provided by Calcul Québec (<https://www.calculquebec.ca>) and Compute Canada (<https://www.computeCanada.ca>). The authors would like to recognize the contribution from SB Smith, L Diatchenko, and the analytical team at McGill University, in particular M Parisien, for providing support with the data from OPPERA, OPPERA II, and CPPC studies. OPPERA was supported by the National Institute of Dental and Craniofacial Research (NIDCR; <https://www.nidcr.nih.gov/>): grant number U01DE017018. The OPPERA program also acknowledges resources specifically provided for this project by the respective host universities: University at Buffalo, University of Florida, University of Maryland–Baltimore, and University of North Carolina–Chapel Hill. Funding for genotyping was provided by NIDCR through a contract to the Center for Inherited Disease Research at Johns Hopkins University (HHSN268201200008I). Data from the OPPERA study are available through the NIH dbGaP: phs000796.v1.p1 and phs000761.v1.p1. L Diatchenko and the analytical team at McGill University were supported by the Canadian Excellence Research Chairs (CERC) Program grant (<http://www.cerc.gc.ca/home-accueil-eng.aspx>, CERC09). The Complex Persistent Pain Conditions: Unique and Shared Pathways of Vulnerability Program Project were supported by NIH/National Institute of Neurological Disorders and Stroke (NINDS; <https://www.ninds.nih.gov>) grant NS045685 to the University of North Carolina at Chapel Hill, and genotyping was funded by the Canadian Excellence Research Chairs (CERC) Program (grant CERC09). The OPPERA II study was supported by the NIDCR under Award Number U01DE017018, and genotyping was funded by the Canadian Excellence Research Chairs (CERC) Program (grant CERC09).


Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Fonds de recherche Québec-Santé [267074 to K.O.]; and the Natural Sciences and Engineering Research Council of Canada [RGPIN-2019-06727 to K.O., RGPIN-2020-05133 to S.B.].

ORCID iD

Julien St-Pierre  <https://orcid.org/0000-0001-9627-576X>

Supplemental material

Supplemental material for this article is available online: Our Julia package PenalizedGLMM and codes for simulating data are available on <https://github.com/julstpierre/PenalizedGLMMgithub>. The data from the real data application that supports the findings of this study are available from the corresponding author upon reasonable request.

References

1. Visscher PM, Wray NR, Zhang Q, et al. 10 years of GWAS discovery: biology, function, and translation. *Am J Hum Genet* 2017; **101**: 5–22.
2. Manolio TA, Collins FS, Cox NJ, et al. Finding the missing heritability of complex diseases. *Nature* 2009; **461**: 747–753.
3. Shi M, O'Brien KM and Weinberg CR. Interactions between a polygenic risk score and non-genetic risk factors in young-onset breast cancer. *Sci Rep* 2020; **10**. DOI: 10.1038/s41598-020-60032-3.
4. Mukherjee B, Ahn J, Gruber SB, et al. Case-control studies of gene-environment interaction: Bayesian design and analysis. *Biometrics* 2009; **66**: 934–948.
5. Fang K, Li J, Zhang Q, et al. Pathological imaging-assisted cancer gene-environment interaction analysis. *Biometrics* 2023. DOI: 10.1111/biom.13873.
6. Zemlianskaia N, Gauderman WJ and Lewinger JP. A scalable hierarchical lasso for gene-environment interactions. *J Comput Graph Stat* 2022; 1–13. DOI: 10.1080/10618600.2022.2039161.
7. Lim M and Hastie T. Learning interactions via hierarchical group-lasso regularization. *J Comput Graph Stat* 2015; **24**: 627–654.
8. Yu J, Pressoir G, Briggs WH, et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat Genet* 2005; **38**: 203–208.
9. Hoffman GE. Correcting for population structure and kinship using the linear mixed model: theory and extensions. *PLoS ONE* 2013; **8**: e75707.
10. Price AL, Zaitlen NA, Reich D, et al. New approaches to population stratification in genome-wide association studies. *Nat Rev Genet* 2010; **11**: 459–463.
11. Qian J, Tanigawa Y, Du W, et al. A fast and scalable framework for large-scale and ultrahigh-dimensional sparse regression with application to the UK Biobank. *PLoS Genet* 2020; **16**: e1009141.
12. St-Pierre J, Oualkacha K and Bhatnagar SR. Efficient penalized generalized linear mixed models for variable selection and genetic risk prediction in high-dimensional data. *Bioinformatics* 2023; **39**. DOI: 10.1093/bioinformatics/btad063.
13. Hu L, Lu W, Zhou J, et al. MM algorithms for variance component estimation and selection in logistic linear mixed model. *Stat Sin* 2019. DOI: 10.5705/ss.202017.0220.
14. Hui FKC, Müller S and Welsh AH. Joint selection in mixed models using regularized PQL. *J Am Stat Assoc* 2017; **112**: 1323–1333.
15. Bien J, Taylor J and Tibshirani R. A lasso for hierarchical interactions. *Ann Stats* 2013; **41**. DOI: 10.1214/13-aos1096.
16. Cox DR. Interaction. *Int Stats Rev / Revue Internationale de Statistique* 1984; **52**: 1.
17. Dudbridge F and Fletcher O. Gene-environment dependence creates spurious gene-environment interaction. *Am J Hum Genet* 2014; **95**: 301–307.
18. Sul JH, Bilow M, Yang WY, et al. Accounting for population structure in gene-by-environment interactions in genome-wide association studies using mixed models. *PLoS Genet* 2016; **12**: e1005849.
19. Zhao P, Rocha G and Yu B. The composite absolute penalties family for grouped and hierarchical variable selection. *Ann Stats* 2009; **37**. DOI: 10.1214/07-aos584.
20. Maixner W, Diatchenko L, Dubner R, et al. Orofacial pain prospective evaluation and risk assessment study – the OPPERA study. *J Pain* 2011; **12**: T4–T11.e2.
21. Chen H, Wang C, Conomos MP, et al. Control for population structure and relatedness for binary traits in genetic association studies via logistic mixed models. *Am J Hum Genet* 2016; **98**: 653–666.
22. Simon N, Friedman J, Hastie T, et al. A sparse-group lasso. *J Comput Graph Stat* 2013; **22**: 231–245.
23. Zhou H, Hu L, Zhou J, et al. MM algorithms for variance components models. *J Comput Graph Stat* 2019; **28**: 350–361.
24. Gilmour AR, Thompson R and Cullis BR. Average information REML: an efficient algorithm for variance parameter estimation in linear mixed models. *Biometrics* 1995; **51**: 1440.
25. Breslow NE and Clayton DG. Approximate inference in generalized linear mixed models. *J Am Stat Assoc* 1993; **88**: 9–25.
26. Friedman J, Hastie T and Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw* 2010; **33**: 1–22. <https://www.jstatsoft.org/v33/i01/>.
27. Koenig Z, Yohannes MT, Nkambule LL, et al. A harmonized public resource of deeply sequenced diverse human genomes. 2023. DOI: 10.1101/2023.01.23.525248.
28. Smith SB, Parisien M, Bair E, et al. Genome-wide association reveals contribution of MRAS to painful temporomandibular disorder in males. *Pain* 2018; **160**: 579–591.
29. Bueno CH, Pereira DD, Pattussi MP, et al. Gender differences in temporomandibular disorders in adult populational studies: a systematic review and meta-analysis. *J Oral Rehabil* 2018; **45**: 720–729.
30. Delaneau O, Marchini J and Zagury JF. A linear complexity phasing method for thousands of genomes. *Nat Methods* 2011; **9**: 179–181.
31. Howie BN, Donnelly P and Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* 2009; **5**: e1000529.
32. Marees AT, de Kluiver H, Stringer S, et al. A tutorial on conducting genome-wide association studies: quality control and statistical analysis. *Int J Methods Psychiatr Res* 2018; **27**: e1608.
33. Chang CC, Chow CC, Tellier LC, et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* 2015; **4**. DOI: 10.1186/s13742-015-0047-8.

34. Zhou W, Nielsen JB, Fritsche LG, et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat Genet* 2018; **50**: 1335–1341.
35. Jiang L, Zheng Z, Qi T, et al. A resource-efficient tool for mixed model association analysis of large-scale data. *Nat Genet* 2019; **51**: 1749–1755.
36. Fan J and Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J Am Stat Assoc* 2001; **96**: 1348–1360.
37. Zhang CH. Nearly unbiased variable selection under minimax concave penalty. *Ann Stats* 2010; **38**: 894–942.
38. Meinshausen N. Relaxed lasso. *Comput Stat Data Anal* 2007; **52**: 374–393.
39. Janková J, Shah RD, Bühlmann P, et al. Goodness-of-fit testing in high dimensional generalized linear models. *J R Stat Soc Ser B: Stat Methodol* 2020; **82**: 773–795.
40. Pan Z and Lin DY. Goodness-of-fit methods for generalized linear mixed models. *Biometrics* 2005; **61**: 1000–1009.
41. Shen X, Alam M, Fikse F, et al. A novel generalized ridge regression method for quantitative genetics. *Genetics* 2013; **193**: 1255–1268.
42. Bondell HD, Krishna A and Ghosh SK. Joint variable selection for fixed and random effects in linear mixed-effects models. *Biometrics* 2010; **66**: 1069–1077.
43. Ødegård J, Indahl U, Strandén I, et al. Large-scale genomic prediction using singular value decomposition of the genotype matrix. *Genet Sel Evol* 2018; **50**. DOI: 10.1186/s12711-018-0373-2.
44. Kooij A. *Prediction accuracy and stability of regression with optimal scaling transformations*. PhD Thesis, Faculty of Social and Behavioural Sciences, Leiden University, 2007.
45. Tibshirani R, Bien J, Friedman J, et al. Strong rules for discarding predictors in lasso-type problems. *J R Stat Soc: Ser B (Stat Methodol)* 2011; **74**: 245–266.
46. Liang X, Cohen A, Heinsfeld AS, et al. sparsegl: an R package for estimating sparse group lasso. *arXiv preprint arXiv:220802942* 2022.
47. Bhatnagar SR, Yang Y, Lu T, et al. Simultaneous SNP selection and adjustment for population structure in high dimensional prediction models. *PLoS Genet* 2020; **16**: e1008766.
48. Wu TT and Lange K. Coordinate descent algorithms for lasso penalized regression. *Ann Appl Stat* 2008; **2**. DOI: 10.1214/07-aos147.

A Appendix

A.1 Updates for $\tilde{\delta}$

The gradient and Hessian of $f(\Theta; \delta)$ are given by the following equations:

$$\begin{aligned}\nabla_{\delta} f(\Theta; \delta) &= -\hat{\phi}^{-1} U^{\top} (\mathbf{y} - \boldsymbol{\mu}) + \Lambda^{-1} \delta \\ \nabla_{\delta}^2 f(\Theta; \delta) &= \hat{\phi}^{-1} U^{\top} \Delta^{-1} U + \Lambda^{-1}\end{aligned}$$

This leads to the Newton updates

$$\begin{aligned}\tilde{\delta}^{(t+1)} &= \tilde{\delta}^{(t)} - [\nabla_{\delta}^2 f(\Theta | \tilde{\delta}^{(t)})]^{-1} \nabla_{\delta} f(\Theta | \tilde{\delta}^{(t)}) \\ &= \tilde{\delta}^{(t)} + [\hat{\phi}^{-1} U^{\top} \Delta^{-(t)} U + \Lambda^{-1}]^{-1} \left(\hat{\phi}^{-1} U^{\top} (\mathbf{y} - \boldsymbol{\mu}^{(t)}) - \Lambda^{-1} \tilde{\delta}^{(t)} \right) \\ &= [U^{\top} \Delta^{-(t)} U + \hat{\phi} \Lambda^{-1}]^{-1} U^{\top} \Delta^{-(t)} \left(\Delta^{(t)} (\mathbf{y} - \boldsymbol{\mu}^{(t)}) + U \tilde{\delta}^{(t)} \right)\end{aligned}\quad (7)$$

which requires repeatedly inverting the $n \times n$ matrix $\boldsymbol{\Sigma}^{(t)} := U^{\top} \Delta^{-(t)} U + \hat{\phi} \Lambda^{-1}$ with complexity $O(n^3)$ where n is the sample size. Defining the working vector $\tilde{\mathbf{Y}} = \mathbf{X}\Theta^{(t)} + U\tilde{\delta}^{(t)} + \Delta^{(t)}(\mathbf{y} - \boldsymbol{\mu}^{(t)})$, where $\mathbf{X}\Theta = \mathbf{Z}\theta + \mathbf{D}\alpha + \mathbf{G}\beta + (\mathbf{D} \odot \mathbf{G})\gamma$, the Newton updates in (7) can be rewritten as follows:

$$\tilde{\delta}^{(t+1)} = [U^{\top} \Delta^{-(t)} U + \hat{\phi} \Lambda^{-1}]^{-1} U^{\top} \Delta^{-(t)} (\tilde{\mathbf{Y}} - \mathbf{X}\Theta^{(t)})$$

which can be equivalently obtained as the solutions to the following generalized ridge WLS problem:

$$\tilde{\delta}^{(t+1)} = \underset{\delta}{\operatorname{argmin}} \hat{\phi}^{-1} (\tilde{\mathbf{Y}} - \mathbf{X}\Theta^{(t)} - U\delta)^{\top} \Delta^{-(t)} (\tilde{\mathbf{Y}} - \mathbf{X}\Theta^{(t)} - U\delta) + \delta^{\top} \Lambda^{-1} \delta \quad (8)$$

Equation (8) is analogous to the PC ridge regression (PCRR) model,⁴³ and demonstrates that PCA and MMs indeed share the same underlying model. At last, to solve (8) without repeatedly inverting the $n \times n$ matrix $\boldsymbol{\Sigma}^{(t)} := U^{\top} \Delta^{-(t)} U + \hat{\phi} \Lambda^{-1}$,

we propose using a coordinate descent algorithm,⁴⁴ for which each coordinate's updates are given, for $j = 1, \dots, n$, by

$$\tilde{\delta}_j \leftarrow \frac{\sum_{i=1}^n w_i U_{ij} \left(\tilde{Y}_i - \mathbf{X}_i \boldsymbol{\Theta}^{(t)} - \sum_{l \neq j} U_{il} \tilde{\delta}_l \right)}{\sum_{i=1}^n w_i U_{ij}^2 + \hat{\phi} \Lambda_j^{-1}} \quad (9)$$

where $w_i = \Delta_{ii}^{-{(t)}}$.

A.2 Updates for $\boldsymbol{\Theta}$

Since the objective function in (5) consists of a smooth convex function $f(\boldsymbol{\Theta}; \boldsymbol{\delta})$ and a non-smooth convex regularizer $g(\boldsymbol{\Theta})$, we propose a proximal Newton algorithm with cyclic coordinate descent to find PQL regularized estimates for $\boldsymbol{\Theta}$, in the spirit of the proposed algorithm by Friedman et al.²⁶ for estimation of GLMs with convex penalties. Let again $\mathbf{X}\boldsymbol{\Theta} = \mathbf{Z}\boldsymbol{\theta} + \mathbf{D}\boldsymbol{\alpha} + \mathbf{G}\boldsymbol{\beta} + (\mathbf{D} \odot \mathbf{G})\boldsymbol{\gamma}$ and $\boldsymbol{\Theta}^{(t)}$ be the current iterate, the iterative step reduces to

$$\begin{aligned} \boldsymbol{\Theta}^{(t+1)} &= \underset{\boldsymbol{\Theta}}{\operatorname{argmin}} \left\{ \frac{1}{2s_t} \left\| \boldsymbol{\Theta} - \left(\boldsymbol{\Theta}^{(t)} - s_t \left[\nabla_{\boldsymbol{\Theta}}^2 f(\boldsymbol{\Theta}^{(t)} | \tilde{\boldsymbol{\delta}}) \right]^{-1} \nabla_{\boldsymbol{\Theta}} f(\boldsymbol{\Theta}^{(t)} | \tilde{\boldsymbol{\delta}}) \right) \right\|_2^2 + g(\boldsymbol{\Theta}) \right\} \\ &= \underset{\boldsymbol{\Theta}}{\operatorname{argmin}} \left\{ \frac{1}{2s_t} \left\| \boldsymbol{\Theta} - [\mathbf{X}^T \boldsymbol{\Delta}^{-{(t)}} \mathbf{X}]^{-1} \mathbf{X}^T \boldsymbol{\Delta}^{-{(t)}} (\mathbf{X}\boldsymbol{\Theta}^{(t)} + s_t \boldsymbol{\Delta}^{(t)} (\mathbf{y} - \boldsymbol{\mu}^{(t)})) \right\|_2^2 + g(\boldsymbol{\Theta}) \right\} \end{aligned}$$

where s_t is a suitable step size. Defining the working vector $\tilde{\mathbf{Y}} = \mathbf{X}\boldsymbol{\Theta}^{(t)} + \mathbf{U}\tilde{\boldsymbol{\delta}}^{(t+1)} + s_t \boldsymbol{\Delta}^{(t)} (\mathbf{y} - \boldsymbol{\mu}^{(t)})$, we can again rewrite the minimization problem as a WLS problem where

$$\begin{aligned} \boldsymbol{\Theta}^{(t+1)} &= \underset{\boldsymbol{\Theta}}{\operatorname{argmin}} \left\{ \frac{1}{2s_t} \left\| \boldsymbol{\Theta} - [\mathbf{X}^T \boldsymbol{\Delta}^{-{(t)}} \mathbf{X}]^{-1} \mathbf{X}^T \boldsymbol{\Delta}^{-{(t)}} (\tilde{\mathbf{Y}} - \mathbf{U}\tilde{\boldsymbol{\delta}}^{(t+1)}) \right\|_2^2 + g(\boldsymbol{\Theta}) \right\} \\ &= \underset{\boldsymbol{\Theta}}{\operatorname{argmin}} \left\{ \frac{1}{2s_t} \sum_{i=1}^n w_i \left(\tilde{Y}_i - \mathbf{X}_i \boldsymbol{\Theta} - \mathbf{U}_i \tilde{\boldsymbol{\delta}}^{(t+1)} \right)^2 + (1 - \rho) \lambda \sum_j \|\beta_j, \gamma_j\|_2 + \rho \lambda \sum_j |\gamma_j| \right\} \quad (10) \end{aligned}$$

where $w_i = \Delta_{ii}^{-{(t)}}$. We use block coordinate descent and minimize (10) with respect to each component of $\boldsymbol{\Theta} = (\boldsymbol{\theta}^T, \boldsymbol{\alpha}^T, \boldsymbol{\beta}^T, \boldsymbol{\gamma}^T)^T$. In practice, we set $s_t = 1$ and do not perform step-size optimization. We present in Appendix 5 the detailed derivations and our block coordinate descent algorithm to obtain PQL regularized estimates for $\boldsymbol{\Theta}$.

A.3 Strong rule

In modern genome-wide studies, the number of genetic predictors is often very large, and assuming that most of the predictors effects are equal to 0, it would be desirable to discard them from the coordinate descent steps to speed up the optimization procedure. Tibshirani et al.⁴⁵ derived sequential strong rules that can be used when solving the lasso and lasso-type problems over a grid of tuning parameter values $\lambda_1 \geq \lambda_2 \geq \lambda_m$, and more details about the derivation of the sequential strong rule for the sparse group lasso can be found by Liang.⁴⁶ Therefore, having already computed the solution $\hat{\boldsymbol{\Theta}}_{k-1}$ at λ_{k-1} , the sequential strong rule discards the j^{th} genetic predictor from the optimization problem at λ_k if

$$\sqrt{\left(G_j^T (\mathbf{y} - \boldsymbol{\mu}(\hat{\boldsymbol{\Theta}}_{k-1})) \right)^2 + \left(S_{\rho \lambda_{k-1}} \left((\mathbf{D} \odot G_j)^T (\mathbf{y} - \boldsymbol{\mu}(\hat{\boldsymbol{\Theta}}_{k-1})) \right) \right)^2} \leq (1 - \rho)(2\lambda_k - \lambda_{k-1})$$

where $S_\lambda(\cdot)$ is the soft-thresholding function defined as follows:

$$S_\lambda(a) = \begin{cases} a - \lambda & \text{if } a > \lambda \\ 0 & \text{if } |a| \leq \lambda \\ a + \lambda & \text{if } a < -\lambda \end{cases}$$

A.4 Prediction

Our proposed method to calculate prediction scores in individuals that were not used in training the models is presented in this section. In sparse regularized PQL estimation, we iteratively fit on a training set of size n the working LMM

$$\tilde{Y} = X\hat{\Theta} + \tilde{b} + \epsilon$$

where $\hat{\Theta} = \{\hat{\Theta}_k \neq 0 | 1 \leq k \leq 2p + m + 1\}$ is the set of non-null predictors, and $\epsilon = g'(\boldsymbol{\mu})(\mathbf{y} - \boldsymbol{\mu}) \sim \mathcal{N}(0, \mathbf{W}^{-1})$, with $\mathbf{W} = \phi^{-1} \text{diag} \left\{ \frac{a_i}{v(\mu_i) |g'(\mu_i)|^2} \right\}$ the diagonal matrix containing weights for each observation. Let \tilde{Y}_s be the latent working vector in a testing set of n_s individuals with predictor set X_s . Similar to Bhatnagar et al.,⁴⁷ we assume that the marginal joint distribution of \tilde{Y}_s and \tilde{Y} is multivariate Normal :

$$\begin{bmatrix} \tilde{Y}_s \\ \tilde{Y} \end{bmatrix} \sim \mathcal{N} \left(\begin{bmatrix} X_s \hat{\Theta} \\ X \hat{\Theta} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix} \right)$$

where $\boldsymbol{\Sigma}_{12} = \text{Cov}(\tilde{Y}_s, \tilde{Y}) = \hat{\tau}_g \mathbf{K}_{12} + \hat{\tau}_d \mathbf{K}_{12}^D$ is the sum of the $n_s \times n$ GSMs between the testing and training individuals, and $\boldsymbol{\Sigma}_{22} = \text{Var}(\tilde{Y}) = \mathbf{W}^{-1} + \hat{\tau}_g \mathbf{K}_{22} + \hat{\tau}_d \mathbf{K}_{22}^D$. It follows from standard normal theory that

$$\tilde{Y}_s | \tilde{Y}, \hat{\phi}, \hat{\tau}, \hat{\Theta}, X, X_s \sim \mathcal{N} \left(X_s \hat{\Theta} + \boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_{22}^{-1} (\tilde{Y} - X \hat{\Theta}), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\Sigma}_{21} \right)$$

The predictions are based on the conditional expectation $\mathbb{E}[\tilde{Y}_s | \tilde{Y}, \hat{\phi}, \hat{\tau}, \hat{\Theta}, X, X_s]$, that is,

$$\begin{aligned} \hat{\mu}_s &= g^{-1} \left(\mathbb{E}[\tilde{Y}_s | \tilde{Y}, \hat{\phi}, \hat{\tau}, \hat{\Theta}, X, X_s] \right) \\ &= g^{-1} \left(X_s \hat{\Theta} + \boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_{22}^{-1} (\tilde{Y} - X \hat{\Theta}) \right) \\ &= g^{-1} \left(X_s \hat{\Theta} + \boldsymbol{\Sigma}_{12} (\mathbf{W}^{-1} + \mathbf{U} \boldsymbol{\Lambda} \mathbf{U}^T)^{-1} (\tilde{Y} - X \hat{\Theta}) \right) \end{aligned}$$

where $g(\cdot)$ is the link function and $\mathbf{U} \boldsymbol{\Lambda} \mathbf{U}^T$ is the spectral decomposition of the GSM for training subjects, with \mathbf{U} the $n \times n$ matrix of eigenvectors.

A.5 Proximal Newton method

Defining the working vector $\tilde{Y} = X\boldsymbol{\Theta}^{(t)} + U\tilde{\boldsymbol{\delta}}^{(t+1)} + s_t \boldsymbol{\Delta}^{(t)}(\mathbf{y} - \boldsymbol{\mu}^{(t)})$ with suitable step size s_t , we can again rewrite the minimization problem as a WLS problem where

$$\begin{aligned} \boldsymbol{\Theta}^{(t+1)} &= \underset{\boldsymbol{\Theta}}{\text{argmin}} \left\{ \frac{1}{2s_t} \left\| \boldsymbol{\Theta} - [X^T \boldsymbol{\Delta}^{-(t)} X]^{-1} X^T \boldsymbol{\Delta}^{-(t)} (\tilde{Y}^{(t)} - U\tilde{\boldsymbol{\delta}}^{(t+1)}) \right\|_2^2 + g(\boldsymbol{\Theta}) \right\} \\ &= \underset{\boldsymbol{\Theta}}{\text{argmin}} \left\{ \frac{1}{2s_t} \sum_{i=1}^n w_i (\tilde{Y}_i - X_i \boldsymbol{\Theta} - U_i \tilde{\boldsymbol{\delta}}^{(t+1)})^2 + (1 - \rho)\lambda \sum_j \|(\beta_j, \gamma_j)\|_2 + \rho\lambda \sum_j |\gamma_j| \right\} \end{aligned} \quad (11)$$

with $w_i = \Delta_{ii}^{-(l)}$. We use block coordinate descent and minimize (11) with respect to each component of $\Theta = (\theta^T, \alpha^T, \beta^T, \gamma^T)^T$. Suppose we have estimates $\tilde{\theta}_l$ for $l \neq j$, $\tilde{\beta}$, $\tilde{\gamma}$, and $\tilde{\delta}$, it is straightforward to show that the updates for θ_j and α are given by the following equations:

$$\begin{aligned}\tilde{\theta}_j &\leftarrow \frac{\sum_{i=1}^n w_i Z_{ij} \left(\tilde{Y}_i - \sum_{l \neq j} Z_{il} \tilde{\theta}_l - D_i \tilde{\alpha} - G_i \tilde{\beta} - (D_i \odot G_i) \tilde{\gamma} - U_i \tilde{\delta} \right)}{\sum_{i=1}^n w_i Z_{ij}^2} \\ \tilde{\alpha} &\leftarrow \frac{\sum_{i=1}^n w_i D_i \left(\tilde{Y}_i - Z_i \tilde{\theta} - G_i \tilde{\beta} - (D_i \odot G_i) \tilde{\gamma} - U_i \tilde{\delta} \right)}{\sum_{i=1}^n w_i D_i^2}\end{aligned}$$

Denote the residual $r_{i;-j} = \tilde{Y}_i - Z_i \tilde{\theta} - D_i \tilde{\alpha} - \sum_{l \neq j} G_{il} \tilde{\beta}_l - \sum_{l \neq j} (D_i \odot G_{il}) \tilde{\gamma}_l - U_i \tilde{\delta}$. The subgradient equations for β_j and γ_j are equal to

$$0 \in \begin{bmatrix} -\sum_{i=1}^n w_i G_{ij} (r_{i;-j} - G_{ij} \tilde{\beta}_j - (D_i \odot G_{ij}) \tilde{\gamma}_j) \\ -\sum_{i=1}^n w_i (D_i \odot G_{ij}) (r_{i;-j} - G_{ij} \tilde{\beta}_j - (D_i \odot G_{ij}) \tilde{\gamma}_j) + \rho \lambda s_t \partial \|\tilde{\gamma}_j\|_1 \end{bmatrix} + (1 - \rho) \lambda s_t \partial \|\tilde{\beta}_j, \tilde{\gamma}_j\|_2$$

where we define the subgradients

$$u \in \partial \|\tilde{\gamma}_j\|_1 = \begin{cases} [-1, 1] & \text{if } \tilde{\gamma}_j = 0 \\ \text{sign}(\tilde{\gamma}_j) & \text{if } \tilde{\gamma}_j \neq 0 \end{cases}; \quad v \in \partial \|\tilde{\beta}_j, \tilde{\gamma}_j\|_2 = \begin{cases} \{v \mid \|v\|_2 \leq 1\} & \text{if } \tilde{\beta}_j = \tilde{\gamma}_j = 0 \\ \frac{1}{\|\tilde{\beta}_j, \tilde{\gamma}_j\|_2} \begin{bmatrix} \tilde{\beta}_j \\ \tilde{\gamma}_j \end{bmatrix} & \text{otherwise} \end{cases}$$

(1) The case $\tilde{\beta}_j = \tilde{\gamma}_j = 0$ implies

$$\begin{bmatrix} \sum_{i=1}^n w_i G_{ij} r_{i;-j} \\ \sum_{i=1}^n w_i (D_i \odot G_{ij}) r_{i;-j} - \rho \lambda s_t u \end{bmatrix} = (1 - \rho) \lambda s_t v$$

Since $\|v\|_2 \leq 1$, equality of the constraint holds as long as

$$\left(\sum_{i=1}^n w_i G_{ij} r_{i;-j} \right)^2 + \left(\sum_{i=1}^n w_i (D_i \odot G_{ij}) r_{i;-j} - \rho \lambda s_t u \right)^2 \leq ((1 - \rho) \lambda s_t)^2$$

Since $u \in [-1, 1]$, a necessary and sufficient condition for $\tilde{\beta}_j = \tilde{\gamma}_j = 0$ being a solution is

$$\left(\sum_{i=1}^n w_i G_{ij} r_{i;-j} \right)^2 + \left(S_{\rho \lambda s_t} \left(\sum_{i=1}^n w_i (D_i \odot G_{ij}) r_{i;-j} \right) \right)^2 \leq ((1 - \rho) \lambda s_t)^2 \quad (12)$$

where $S_\lambda(\cdot)$ is the soft-thresholding function defined as follows:

$$S_\lambda(a) = \begin{cases} a - \lambda & \text{if } a > \lambda \\ 0 & \text{if } |a| \leq \lambda \\ a + \lambda & \text{if } a < -\lambda \end{cases}$$

(2) The case $(\tilde{\beta}_j, \tilde{\gamma}_j)^\top \neq \mathbf{0}$ implies

$$\begin{bmatrix} \sum_{i=1}^n w_i G_{ij} (r_{i;-j} - D_i G_{ij} \tilde{\gamma}_j) \\ \sum_{i=1}^n w_i (D_i \odot G_{ij}) (r_{i;-j} - G_{ij} \tilde{\beta}_j) - \rho \lambda s_t u \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n w_i G_{ij}^2 & 0 \\ 0 & \sum_{i=1}^n w_i (D_i \odot G_{ij})^2 \end{bmatrix} + \frac{(1-\rho)\lambda s_t}{\sqrt{\tilde{\beta}_j^2 + \tilde{\gamma}_j^2}} \mathbf{I}_2 \begin{bmatrix} \tilde{\beta}_j \\ \tilde{\gamma}_j \end{bmatrix} \quad (13)$$

We have that $\tilde{\gamma}_j = 0$ if $|\sum_{i=1}^n w_i (D_i \odot G_{ij}) (r_{i;-j} - G_{ij} \tilde{\beta}_j)| \leq \rho \lambda s_t$ since $u \in [-1, 1]$. This implies that

$$\sum_{i=1}^n w_i G_{ij} r_{i;-j} = \left(\sum_{i=1}^n w_i G_{ij}^2 + (1-\rho) \frac{\lambda s_t}{|\tilde{\beta}_j|} \right) \tilde{\beta}_j$$

with the solution being equal to

$$\tilde{\beta}_j = \frac{S_{(1-\rho)\lambda s_t} \left(\sum_{i=1}^n w_i G_{ij} r_{i;-j} \right)}{\sum_{i=1}^n w_i G_{ij}^2}$$

There is no closed-form solution for (13) if both $\tilde{\gamma}_j$ and $\tilde{\beta}_j$ are non-null. In this case, we can replace (11) by a surrogate objective function using a MM algorithm.⁴⁸ From the concavity of the ℓ_2 norm $\|\beta_j, \gamma_j\|_2 = \sqrt{\beta_j^2 + \gamma_j^2}$, we have the following inequality:

$$\|\beta_j, \gamma_j\|_2 \leq \|\beta_j^{(t)}, \gamma_j^{(t)}\|_2 + \frac{1}{2\|\beta_j^{(t)}, \gamma_j^{(t)}\|_2} \left(\|\beta_j, \gamma_j\|_2^2 - \|\beta_j^{(t)}, \gamma_j^{(t)}\|_2^2 \right)$$

from where we derive the MM iterative step

$$\Theta^{(t+1)} = \underset{\Theta}{\operatorname{argmin}} \left\{ \frac{1}{2s_t} \sum_{i=1}^n w_i \left(\tilde{Y}_i^{(t)} - \mathbf{X}_i \Theta - \mathbf{U}_i \delta^{(t+1)} \right)^2 + (1-\rho)\lambda \sum_j \frac{\|\beta_j, \gamma_j\|_2^2}{2\|\beta_j^{(t)}, \gamma_j^{(t)}\|_2} + \rho\lambda \sum_j |\gamma_j| \right\}$$

Using cyclic coordinate descent, the updates for β_j and γ_j are given by the following equations:

$$\begin{aligned} \tilde{\beta}_j &\leftarrow \frac{\sum_{i=1}^n w_i G_{ij} (r_{i;-j} - D_i G_{ij} \tilde{\gamma}_j)}{\sum_{i=1}^n w_i G_{ij}^2 + (1-\rho)\lambda \tilde{s}_t} \\ \tilde{\gamma}_j &\leftarrow \frac{S_{\rho\lambda s_t} \left(\sum_{i=1}^n w_i D_i G_{ij} (r_{i;-j} - G_{ij} \tilde{\beta}_j) \right)}{\sum_{i=1}^n w_i (D_i G_{ij})^2 + (1-\rho)\lambda \tilde{s}_t} \end{aligned}$$

where we defined $\tilde{s}_t = s_t / \|\beta_j^{(t)}, \gamma_j^{(t)}\|_2$. Algorithm 1 summarizes our block coordinate descent (BCD) procedure to obtain regularized estimates for the fixed effects vector $\Theta = (\theta^\top, \alpha^\top, \beta^\top, \gamma^\top)^\top$.

Algorithm 1: BCD algorithm to minimize the PQL loss function of the GEI model (5) with mixed lasso and group lasso penalties for GLMMs.

Input: $y, X = [Z D G (D \odot G)]$

Output: $\hat{\theta}, \hat{\alpha}, \hat{\beta}, \hat{\gamma}$

Estimate τ_g, τ_d and ϕ under the null model (i.e. $\beta = \gamma = \mathbf{0}$) using the AI-REML algorithm; Given $\hat{\tau}_g, \hat{\tau}_d$ and $\hat{\phi}$, perform spectral decomposition of the random effects covariance matrix $\hat{\tau}_g \mathbf{K} + \hat{\tau}_d \mathbf{K}^D = \mathbf{U} \mathbf{\Lambda} \mathbf{U}^T$; Initialize

$\Theta^{(0)} = (\theta^{(0)T}, \alpha^{(0)T}, \beta^{(0)T}, \gamma^{(0)T})^T$ and $\tilde{\delta}^{(0)}$; **for** $\lambda = \lambda_1, \lambda_2, \dots$ **do**

for $t=0, 1, \dots$ **until convergence do**

Select a suitable step size s_t ; Update $\mu^{(t)} \leftarrow g^{-1}(\mathbf{X}\Theta^{(t)} + \mathbf{U}\tilde{\delta}^{(t)})$, $\mathbf{\Delta}^{(t)} \leftarrow \text{diag}(g'(\mu^{(t)}))$ and $w_i \leftarrow \mathbf{\Delta}_{ii}^{-1(t)}$ for $i = 1, \dots, n$; Update $\tilde{Y} \leftarrow \mathbf{X}\Theta^{(t)} + \mathbf{U}\tilde{\delta}^{(t)} + s_t \mathbf{\Delta}^{(t)}(y - \mu^{(t)})$; **/* Inner loop to estimate $\tilde{\delta}$ for $j=1, \dots, n$ until convergence do**

$$\tilde{\delta}_j^{(t+1)} \leftarrow \frac{\sum_{i=1}^n w_i U_{ij} (\tilde{Y}_i - X_i \Theta^{(t)} - \sum_{l \neq j} U_{il} \tilde{\delta}_l)}{\sum_{i=1}^n w_i U_{ij}^2 + \hat{\phi} \Lambda_j^{-1}}$$

Update $\mu^{(t)} \leftarrow g^{-1}(\mathbf{X}\Theta^{(t)} + \mathbf{U}\tilde{\delta}^{(t+1)})$; Update $\tilde{Y} \leftarrow \mathbf{X}\Theta^{(t)} + \mathbf{U}\tilde{\delta}^{(t+1)} + s_t \mathbf{\Delta}^{(t)}(y - \mu^{(t)})$;

/* Inner loop to estimate $\Theta^{(t+1)}$ for $k=1, \dots, m$ until convergence do

$$\tilde{\theta}_k \leftarrow \frac{\sum_{i=1}^n w_i Z_{ik} (\tilde{Y}_i - \sum_{l \neq k} Z_{il} \tilde{\theta}_l - D_i \tilde{\alpha} - G_i \tilde{\beta} - (D_i \odot G_i) \tilde{\gamma} - U_i \tilde{\delta})}{\sum_{i=1}^n w_i Z_{ik}^2}$$

$$\tilde{\alpha} \leftarrow \frac{\sum_{i=1}^n w_i D_i (\tilde{Y}_i - Z_i \tilde{\theta} - G_i \tilde{\beta} - (D_i \odot G_i) \tilde{\gamma} - U_i \tilde{\delta})}{\sum_{i=1}^n w_i D_i^2}$$

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

Compute $r_{i;-j} = \tilde{Y}_i - Z_i \tilde{\theta} - D_i \tilde{\alpha} - \sum_{l \neq j} G_{il} \tilde{\beta}_l - \sum_{l \neq j} (D_i \odot G_{il}) \tilde{\gamma}_l - U_i \tilde{\delta}$; If $|\sum_{i=1}^n w_i (D_i \odot G_{ij})(r_{i;-j} - G_{ij} \tilde{\beta}_j)| \leq \lambda s_t$ then set

$$\tilde{\gamma}_j \leftarrow 0 \text{ and } \tilde{\beta}_j \leftarrow \frac{S_{\lambda s_t} (\sum_{i=1}^n w_i G_{ij} r_{i;-j})}{\sum_{i=1}^n w_i G_{ij}^2}$$

Else then set

$$\tilde{\beta}_j \leftarrow \frac{\sum_{i=1}^n w_i G_{ij} (r_{i;-j} - D_i G_{ij} \tilde{\gamma}_j)}{\sum_{i=1}^n w_i G_{ij}^2 + \lambda \tilde{s}_t}$$

$$\tilde{\gamma}_j \leftarrow \frac{S_{\lambda \tilde{s}_t} (\sum_{i=1}^n w_i D_i G_{ij} (r_{i;-j} - G_{ij} \tilde{\beta}_j))}{\sum_{i=1}^n w_i (D_i G_{ij})^2 + \lambda \tilde{s}_t}$$

where $\tilde{s}_t = s_t / \|\beta_j^{(t)}, \gamma_j^{(t)}\|_2$.