Penalized generalized linear mixed models for longitudinal outcomes in genetic association studies

Julien St-Pierre[1], Sahir Rai Bhatnagar[1], Massimiliano Orri[1,2], Josée Dupuis[1], Karim Oualkacha[3].

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

[2]McGill Group for Suicide Studies, Douglas Mental Health University Institute, Department of Psychiatry, McGill University

[3]Département de Mathématiques, Université du Québec à Montréal

Abstract

This work is motivated by analyses of longitudinal data collected from participants in the Quebec Longitudinal Study of Child Development (QLSCD) and the Quebec Newborn Twin Study (QNTS) to identify important genetic predictors for emotional and behavioural difficulties in childhood and adolescence. We propose a lasso penalized mixed model for continuous and binary longitudinal traits that allows the inclusion of multiple random effects to account for random individual effects not attributable to the genetic similarity between individuals. Through simulation studies, we show that replacing the estimated genetic relatedness matrix (GRM) by a sparse matrix introduces bias in the variance components estimates, but that the obtained computational gain is major while the impact on the performance of the penalized model to retrieve important predictors is negligible. We compare the performance of the proposed penalized mixed model to a standard lasso and to a univariate mixed model association test and show that the proposed model always identifies causal predictors with greater precision. Finally, we show an application of the proposed methodology to predict three externalizing behavioural scores in the combined QLSCD and QNTS longitudinal cohorts.

5.1 Introduction

Our study of penalized generalized linear mixed models (GLMMs) for longitudinal traits was motivated by analyses of data collected from participants in the Quebec Longitudinal Study of Child Development (QLSCD) and the Quebec Newborn Twin Study (QNTS) to identify important genetic predictors for emotional and behavioral difficulties in childhood and adolescence, including externalizing (e.g., aggression) problems. Because of the longitudinal nature of the study, one needs to explicitly model the correlation between repeated measurements within an individual, one possibility being through the use of mixed-effects regression models. Moreover, genetic correlation between pairs of twins needs to be accounted for via a polygenic random effect (Yang et al. [2011]), otherwise the study may be prone to a loss of power and spurious associations (Yu et al. [2005], Price et al. [2010]). The generalized linear mixed model association test (GMMAT) proposed by Chen et al. [2016] allows to include a known kinship matrix when analysing family samples with known pedigree structures in a homogeneous population, or an empirical genetic relatedness matrix (GRM) to account for both population structure and cryptic relatedness, for genome-wide association studies (GWAS) of continuous and binary traits. In addition, the authors have implemented random intercept only models, and random intercept and random slope models to account for random individual effects not attributable to the similarity between individuals. Typically, between-individuals similarity can be caused by genetic relatedness, shared environmental exposure or study sampling design.

Given the relative low sample sizes of the QLSCD ($n = 721$) and QNTS cohort ($n = 636$), a GWAS may fail to discover significant genetic variants that are associated with emotional and behavioral difficulties in childhood and adolescence. Moreover, obtaining effect size estimates for a large number of individual predictors via logistic or linear mixed models is highly computationally intensive using the GMMAT model, given that a new mixed model needs to be fitted for each predictor. This complicates the calculation of a polygenic score

(PS) for longitudinal outcomes, in which variants effects across the genome are aggregated in order to predict complex traits (Dudbridge [2013], Choi et al. [2020]). Indeed, there is great clinical interest in being able to predict externalizing scores with precision as children following high-chronic trajectories of externalizing and internalizing behaviours have been shown to be at risk of negative long-term outcomes, including peer victimisation (van Lier et al. [2012], Oncioiu et al. [2020]), suicidal ideation and attempt (Orri et al. [2019], Forte et al. [2019]), and substance use (Lemyre et al. [2018], Navarro et al. [2020], Zdebik et al. [2019]).

Penalized models have been proposed as an alternative method to increase the power for identifying weaker genome-wide associations and interactions compared to univariable methods (Chu et al. [2020], Li et al. [2010], Zhou et al. [2010], Wu et al. [2009], St-Pierre et al. [2023]). In this paper, we propose a lasso penalized mixed model framework for continuous and binary traits that allows the inclusion of more than one random effect to account for random individual effects not attributable to the genetic similarity between individuals. We study the performance of the average information restricted maximum likelihood (AIREML (Gilmour et al. [1995])) algorithm when analyzing simulated data with both population structure and subjects relatedness for continuous and binary traits. In addition, we show that replacing the GRM by a sparse matrix greatly reduces the computational time required to fit the penalized model, while having little impact on the performance of the model in retrieving important predictors. Next, we compare the performance of our proposed model in retrieving important predictors for both continuous and binary traits, and demonstrate that it achieves better precision than the GMMAT model and that of a lasso penalized model without any random effect. Finally, we apply our proposed method for predicting externalizing scores in children from the combined QLSCD and QNTS cohorts and compare the performance of the lasso and adaptive lasso mixed models with respect to the predicted scores accuracy and models sparsity.

5.2 Methods

5.2.1 Model

Assume y_{ij}, $i = 1, ..., m$, $j = 1, ..., n_i$, is the measurement of a continuous or binary phenotype at time t_{ij} for subject i, where $n = \sum_{i=1}^{m} n_i$ is the total number of observations. Let $\mathbf{C}_{ij}$ be a $1 \times c$ row vector of possibly time-varying covariates for subject i, $\mathbf{G}_i$ a $1 \times p$ row vector of biallelic single nucleotide polymorphisms (SNPs) taking values $\{0, 1, 2\}$ as the number of copies of the minor allele, $(\boldsymbol{\theta}^\mathsf{T}, \boldsymbol{\beta}^\mathsf{T})^\mathsf{T}$ a $(c+p) \times 1$ column vector of fixed covariate and additive genotype effects including the intercept. We assume that $\mathbf{b}_0 = (b_{01}, ..., b_{0m})^\mathsf{T} \sim \mathcal{N}(0, \sum_{k=1}^{K} \tau_k \mathbf{V}_k)$ is an $m \times 1$ column vector of random intercepts, $\boldsymbol{\tau} = (\tau_1, ..., \tau_K)^\mathsf{T}$ are the variance component parameters that account for the relatedness between individuals, and $\mathbf{V}_1, ..., \mathbf{V}_K$ are known relatedness matrices. We typically define $\mathbf{V}_1$ as the GRM between individuals. Further, we assume that $\mathbf{b}_{1i} = (b_{11i}, b_{12i}, ..., b_{1ri})^\mathsf{T} \sim \mathcal{N}(0, \mathbf{D}(\psi))$ is the $r \times 1$ column vector of subject-specific random effects for $i = 1, ..., m$ to account for the correlation between repeated measurements, where $\mathbf{D}$ is an $r \times r$ covariance matrix and $\boldsymbol{\psi}$ contains the unique elements of $\mathbf{D}$. Let $\mathbf{Z}_{ij} = (Z_{ij1}, ..., Z_{ijr})$ be a $1 \times r$ covariate vector for subject-specific random effects $\mathbf{b}_{1i}$, possibly containing a non-polygenic random intercept and more than one random slope. The phenotypes y_{ij}'s are assumed to be conditionally independent and identically distributed given $(\mathbf{C}_{ij}, \mathbf{G}_i, \mathbf{Z}_{ij}, b_{0i}, \mathbf{b}_{1i})$ and follow any distribution with canonical link function $g(\cdot)$, mean $\mathbb{E}(y_{ij}|\mathbf{C}_{ij}, \mathbf{G}_i, \mathbf{Z}_{ij}, b_{0i}, \mathbf{b}_{1i}) = \mu_{ij}$ and variance $\mathrm{Var}(y_{ij}|\mathbf{C}_{ij}, \mathbf{G}_i, \mathbf{Z}_{ij}, b_{0i}, \mathbf{b}_{1i}) = \phi w_{ij}^{-1} \nu(\mu_{ij})$, where ϕ is a dispersion parameter, w_{ij} are known weights and $\nu(\cdot)$ is the variance function. We have the following GLMM for longitudinal data

$$g(\mu_{ij}) = \eta_{ij} = \mathbf{C}_{ij}\theta + \mathbf{G_i}\boldsymbol{\beta} + b_{0i} + \mathbf{Z}_{ij}\mathbf{b}_{1i}. \tag{5.1}$$

We assume that the random effects vector $\mathbf{b}_0$ is independent of $\mathbf{b}_1 = (\mathbf{b}_{11}^\mathsf{T}, ..., \mathbf{b}_{1r}^\mathsf{T})$, such that the stacked random individual effects vector is $\mathbf{b} = (\mathbf{b}_0^\mathsf{T}, \mathbf{b}_{11}^\mathsf{T}, ..., \mathbf{b}_{1r}^\mathsf{T})^\mathsf{T} \sim N(0, \mathrm{diag}\left\{\sum_{k=1}^{K} \tau_k \mathbf{V}_k, \mathbf{D} \otimes \mathbf{I}_m\right\}$

where $\otimes$ is the Kroneker product. Typically, when there are no time trends and observations for the same individual are assumed to be exchangeable, a model with two random intercepts is appropriate, where the first random intercept captures the correlation induced by genetic relatedness, and the second intercept captures the correlation between repeated measurements. In the case where we observe or suspect individual-specific time trends to vary substantially, we can add one or more random slope to the model.

For ease of presentation, let $\mathbf{y} = (y_{11}, ..., y_{1n_1}, ..., y_{m1}, ..., y_{mn_m})^{\intercal}$ be the stacked outcome vector and $\tilde{\mathbf{Z}}_k$ be an $n \times m$ block-diagonal matrix for the kth subject-specific random effect $\mathbf{b}_{1k}$ for $k = 1, ..., r$. For example, if $\mathbf{b}_{11}$ is a random intercept and $\mathbf{b}_{12}$ is a random slope, we would have for $m = 3$ and $n_1 = n_2 = n_3 = 2$, $\tilde{\mathbf{Z}}_1 = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}$ and $\tilde{\mathbf{Z}}_2 = \begin{bmatrix} Z_{112} & 0 & 0 \\ Z_{122} & 0 & 0 \\ 0 & Z_{212} & 0 \\ 0 & Z_{222} & 0 \\ 0 & 0 & Z_{312} \\ 0 & 0 & Z_{322} \end{bmatrix}$.

Next, let $\tilde{\mathbf{Z}} = \left[\tilde{\mathbf{Z}}_1, \tilde{\mathbf{Z}}_2, ..., \tilde{\mathbf{Z}}_r \right]$ be a $n \times mr$ block matrix, and $\mathbf{L}^{\intercal} = (\underbrace{\mathbf{L}_1^{\intercal}, ..., \mathbf{L}_1^{\intercal}}_{n_1 \; times}, ..., \underbrace{\mathbf{L}_m^{\intercal}, ..., \mathbf{L}_m^{\intercal}}_{n_m \; times})$ be an $m \times n$ matrix of indicators such that $b_{0i} = \mathbf{L}_i \mathbf{b}_0$. Thus, for a model with two random intercepts and one random slope, the random effects for all observations is given by

$$\mathbf{L}\mathbf{b}_0 + \tilde{\mathbf{Z}}\mathbf{b}_1 \sim N(0, \mathbf{L}\left(\sum_{k=1}^{K} \tau_k \mathbf{V}_k \right) \mathbf{L}^{\intercal} + \tilde{\mathbf{Z}}\{ \begin{pmatrix} \psi_1 & \psi_2 \\ \psi_2 & \psi_3 \end{pmatrix} \otimes \mathbf{I}_m \} \tilde{\mathbf{Z}}^{\intercal}).$$

Chen et al. [2016] proposed a different variance-covariance structure for the random effects for all observations in model (5.1). They assumed that $\mathbf{b}_0 \sim \mathcal{N}(0, \sum_{k=1}^{K} \tau_k \mathbf{V}_k + \tau_{K+1}\mathbf{I}_m)$, $Cov(\mathbf{b}_0, \mathbf{b}_1) = \sum_{k=1}^{K} \tau_{K+1+k}\mathbf{V}_k + \tau_{2K+2}\mathbf{I}_m$ and $\mathbf{b}_1 \sim \mathcal{N}(0, \sum_{k=1}^{K} \tau_{2K+2+k}\mathbf{V}_k + \tau_{3K+3}\mathbf{I}_m)$. Thus, in the model with one $(K = 1)$ similarity matrix, they assumed that the random effects for all observations is given by $\mathbf{L}\mathbf{b}_0 + \tilde{\mathbf{Z}}\mathbf{b}_1 \sim N(0, \tilde{\mathbf{Z}}\{ \begin{pmatrix} \tau_1 & \tau_3 \\ \tau_3 & \tau_5 \end{pmatrix} \otimes \mathbf{V}_1 + \begin{pmatrix} \tau_2 & \tau_4 \\ \tau_4 & \tau_6 \end{pmatrix} \otimes \mathbf{I}_m \} \tilde{\mathbf{Z}}^{\intercal})$.

Adding an additional random slope to model (5.1) to account for extra sources of variability

would increase the number of variance components to estimate from 6 to 12, compared to 7 variance components in our proposed variance-covariance structure, impacting not only the computational requirements to fit such model, but also the model interpretability. This is why we decided to adopt a different approach, where we assume that $\mathbf{b_0} \perp \mathbf{b_1}$ and that only the polygenic random intercept variance is proportional to the GRM.

5.2.2 Estimation

In order to estimate the model parameters $(\boldsymbol{\theta}^{\mathsf{T}}, \boldsymbol{\beta}^{\mathsf{T}}, \phi, \boldsymbol{\tau}, \boldsymbol{\psi})$ and perform variable selection, we use an approximation method to obtain an analytical closed form for the marginal likelihood of model (5.1). We propose to fit (5.1) using a penalized quasi-likelihood (PQL) method, where the log integrated quasi-likelihood function is equal to

$$
\ell_{PQL}(\boldsymbol{\Theta}, \phi, \boldsymbol{\tau}, \boldsymbol{\psi}; \tilde{\mathbf{b}}) = -\frac{1}{2}\log \left| (\tilde{\mathbf{Z}}(\mathbf{D} \otimes \mathbf{I}_m)\tilde{\mathbf{Z}}^{\mathsf{T}} + \mathbf{L}(\sum_{k=1}^{K} \tau_k \mathbf{V}_k)\mathbf{L}^{\mathsf{T}})\mathbf{W} + \mathbf{I}_n \right| + \sum_{i,j} ql_{ij}(\boldsymbol{\theta}; \tilde{\mathbf{b}})
$$
$$
- \frac{1}{2}\tilde{\mathbf{b}}^{\mathsf{T}} \left(\text{diag}\left\{ \sum_{k=1}^{K} \tau_k \mathbf{V}_k, \mathbf{D} \otimes \mathbf{I}_m \right\} \right)^{-1} \tilde{\mathbf{b}}, \tag{5.2}
$$

and $\boldsymbol{\Theta} = (\boldsymbol{\theta}^{\mathsf{T}}, \boldsymbol{\beta}^{\mathsf{T}})^{\mathsf{T}}$, $\mathbf{W} = \phi^{-1}\boldsymbol{\Delta}^{-1} = \phi^{-1}\text{diag}\left\{ \frac{a_{ij}}{\nu(\mu_{ij})[g'(\mu_{ij})^2]} \right\}$ is a diagonal matrix containing weights for each observation, a_{ij} are known weights, $ql_{ij}(\boldsymbol{\Theta}; \mathbf{b}) = \int_{y_{ij}}^{\mu_{ij}} \frac{a_{ij}(y_{ij}-\mu)}{\phi\nu(\mu)} d\mu$ is the quasi-likelihood for the jth observation from the ith individual given the random effects $\mathbf{b}$, and $\tilde{\mathbf{b}}$ is the solution which maximizes $\sum_{i,j} ql_{ij}(\boldsymbol{\Theta}, \mathbf{b}) - \frac{1}{2}\mathbf{b}^{\mathsf{T}}\left(\text{diag}\left\{ \sum_{k=1}^{K} \tau_k \mathbf{V}_k, \mathbf{D} \otimes \mathbf{I}_m \right\} \right)^{-1} \mathbf{b}$.

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 unidentifiable when modelling p SNPs jointly. Thus, we propose to add a lasso penalty (Tibshirani [1996]) to the negative quasi-likelihood function in (5.2) to seek a sparse subset of genetic effects that gives an adequate fit to the data. We define the following objective function Q_λ

which we seek to minimize with respect to $(\boldsymbol{\Theta}, \phi, \boldsymbol{\tau}, \boldsymbol{\psi})$:

$$Q_\lambda(\boldsymbol{\Theta}, \phi, \boldsymbol{\tau}, \boldsymbol{\psi}; \tilde{\mathbf{b}}) := -\ell_{PQL}(\boldsymbol{\Theta}, \phi, \boldsymbol{\tau}, \boldsymbol{\psi}; \tilde{\mathbf{b}}) + \lambda \sum_j \nu_j |\beta_j|, \qquad (5.3)$$

where $\lambda > 0$ controls the strength of the overall regularization and ν_j are penalization weights that allows incorporating a priori information about the SNP effects. Zou [2006] proposed the adaptive lasso where they defined the weights $\hat{\nu}_j = |\hat{\beta}_j|^{-\gamma}$ for $j = 1, ..., p$, and $\hat{\beta}_j$ is a root-n consistent estimator of β_j, for example the ordinary least squares (OLS) estimator, and $\gamma > 0$ is an additional tuning parameter.

5.2.3 Estimation of variance components

Jointly estimating the variance components and scale parameter vector $(\boldsymbol{\psi}, \boldsymbol{\tau}, \phi)$ with fixed effects parameters vector $\boldsymbol{\Theta}$ is a computationally challenging non-convex optimization problem. Thus, as detailed in St-Pierre et al. [2023], 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 $\boldsymbol{\beta} = 0$, using the AI-REML algorithm. Updates for $\boldsymbol{\tau}$, ϕ and $\boldsymbol{\psi}$ based on the AI-REML algorithm or a majorization-minimization algorithm (Zhou et al. [2019a]) requires iteratively inverting the $n \times n$ covariance matrix $\boldsymbol{\Sigma}$, with complexity $O(n^3)$. To reduce the computational cost of inverting the matrix $\boldsymbol{\Sigma}$, we can use the Woodbury matrix identity, and define the matrix $\mathbf{R} = \tilde{\mathbf{Z}}(\mathbf{D} \otimes \mathbf{I}_m)\tilde{\mathbf{Z}}^\intercal + \mathbf{W}^{-1}$, which yields

$$\boldsymbol{\Sigma}^{-1} = \mathbf{R}^{-1} - \mathbf{R}^{-1}\mathbf{L}(\mathbf{L}^\intercal\mathbf{R}^{-1}\mathbf{L} + (\sum \tau_k \mathbf{V}_k)^{-1})^{-1}\mathbf{L}^\intercal\mathbf{R}^{-1}.$$

Because $\mathbf{R}$ and $\mathbf{L}^\intercal\mathbf{R}^{-1}\mathbf{L}$ are respectively block-diagonal and diagonal matrices, the complexity of inverting the $n \times n$ matrix $\boldsymbol{\Sigma}$ is similar to that of inverting the $m \times m$ matrix $\sum \tau_k \mathbf{V}_k$. To further reduce the computational complexity of the AI-REML estimation procedure when $\sum \tau_k \mathbf{V}_k = \tau_1 \mathbf{V}_1$, that is when $K = 1$, we propose replacing $\mathbf{V_1}$ by a sparse GRM (Jiang

et al. [2019]), where pair-wise relatedness coefficients that are smaller than $2^{-9/2}$ are set to 0. This corresponds to a 3rd degree kinship threshold, meaning that anyone less related than first cousins are assumed to be unrelated. By rearranging the sparse GRM as a block-diagonal matrix, where each block consists of clusters of relatives, we show in the simulation study that it greatly reduces the computational time required to fit both the null and penalized models, and that the resulting bias in the variance components estimates has a limited impact on the performance of the penalized model to retrieve important predictors.

5.3 Simulation study

5.3.1 Simulation model

We performed simulation studies sampling real genotype data from a high quality harmonized set of 4,097 whole genomes from the Human Genome Diversity Project (HGDP) and the 1000 Genomes Project (1000G) (Koenig et al. [2023]), including both related and unrelated individuals from seven distinct population groups (Table 5.1). At each of the 50 replications, we sampled $10,000$ candidate SNPs from 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_s were generated from a Gaussian distribution $\mathcal{N}(0, h_S^2\sigma^2/|S|)$, where h_S^2 is the fraction of variance that is due to total additive genetic fixed effects and σ^2 is the total phenotypic variance.

We simulated a polygenic random intercept $\mathbf{b}_0 \sim \mathcal{N}(0, h_g^2\sigma^2\mathbf{V}_1)$ where h_g^2 is the fraction of variance explained by the polygenic random effect and $\mathbf{V}_1$ is the estimated GRM using the PC-Relate method (Conomos et al. [2016]). The polygenic random intercept $\mathbf{b}_0$ leverages the existing genetic relatedness in the sample due to familial or cryptic relatedness to simulate correlated phenotypes between individuals who share recent common ancestors. More specifically, pair-wise kinship coefficients were first estimated on a set of $13,750$ SNPs selected after LD pruning using the KING-Robust algorithm which is robust to population

Table 5.1: Number of samples by population for the high quality harmonized set of 4,097 whole genomes from the Human Genome Diversity Project (HGDP) and the 1000 Genomes Project (1000G).

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%)
Oceanian	0	30 (3%)	30 (1%)
Total	3,166	931	4,097
Unrelated individuals	2,520	880	3,400

structure (Manichaikul et al. [2010]). Then, PCs were estimated using the PC-AiR method (principal components analysis in related samples) that allows to identify a diverse subset of mutually unrelated individuals such that the top PCs are constructed to only reflect the ancestry and to be robust to both known or cryptic relatedness in the sample (Conomos et al. [2015]). Finally, the GRM was constructed from pair-wise kinship coefficients estimated using the residuals of a linear regression model after adjusting for the ancestry PCs calculated in the previous step. Hence, the PC-Relate method divides genetic correlations among sampled individuals into a component which represents familial relatedness, and another component which represents population structure (Conomos et al. [2016]). To induce additional confounding due to population stratification, we simulated different intercepts π_{0k}, $k = 1, .., 7$, for each population in Table 5.1 using a $U(0.1, 0.3)$ distribution.

For all individuals, we used the sex covariate available from the data set, and we simulated five measurements for age using a Normal distribution, after which 1 to 5 measurements were uniformly sampled to allow different number of observations per individual. To generate correlated observations for each individual, we simulated one random intercept, one random slope for the effect of age and one additional random slope representing the effect of a time-varying environmental exposure from a Gaussian distribution $\mathbf{b}_{1i} \sim \mathcal{N}(0, \mathbf{D})$ with the

covariance matrix $\mathbf{D}$ equal to $\begin{bmatrix} 0.4 & -0.2 & 0.1 \\ -0.2 & 0.5 & 0.2 \\ 0.1 & 0.2 & 0.3 \end{bmatrix}$. Then, for $i = 1, ..., 4097$, $j = 1, ..., n_i$, continuous phenotypes were generated using the following model

$$y_{ij} = \text{logit}(\pi_{0k}) - \log(1.3) \times Sex_i + \log(1.05) \times Age_{ij} + \sum_{s=1}^{100} \beta_s \cdot G_{is} + b_{0i} + \mathbf{Z}_{ij}\mathbf{b}_{1i} + \epsilon_{ij},$$

(5.4)

where G_{is} is the (standardized) number of alleles for the sth causal SNP, $\mathbf{Z}_{ij}$ is the covariate vector for subjects-specific random effects $\mathbf{b}_{1i}$, and ϵ_{ij} is an error term from a standard normal distribution $\mathcal{N}(0, \phi)$. To simulate binary traits, we used a cutoff value c determined such that $\text{P}(y_{ij}^{01} = 1) = \text{P}(y_{ij} > c) = 0.2$.

5.3.2 Results

Estimation of variance components

We first simulated continuous phenotypes under the null model of no genetic association to study the impact on the variance components estimation procedure of including or not the first 10 PCs to control for genetic ancestry, as well as the impact of using a sparse GRM versus using the full GRM. In the upper-right panel of Figure 5.1, we see that the median relative bias for all variance parameters is close to zero when fitting the model with the full GRM and the first 10 PCs, with the interquartile range (IQR) below $\pm 10\%$ for all parameters, except for the covariance parameter ψ_3. When replacing the full GRM by a sparse GRM (lower-right), the median relative bias for the variance of the non polygenic random intercept ψ_1 is around -10%, while the median relative bias for the variance component of the polygenic random intercept τ is around 10%. Thus, the non-polygenic random intercept is capturing some of the variability that is not captured by the polygenic random intercept due to the fact that we are using a sparse GRM. The impact of using no PC to adjust for population structure

113

in the sample is observed in the upper-left panel of Figure 5.1, where the median relative bias for the variance component of the polygenic random intercept τ is around 30%. This is explained by the fact that the PC-Relate estimator of the kinship coefficients is constructed after adjusting for PCs as predictors in linear regression models, and, thus, the residuals are orthogonal to the PCs. Genetic similarities due to more distant ancestry are therefore accounted for by using the PCs as covariates in the null model. When using only a sparse GRM without any PCs to adjust for both population structure and closer relatedness (lower-left), the median relative bias for τ and ψ_1 are quite important, as expected. Finally, in all four modelling strategies, the relative bias for the dispersion parameter ϕ which corresponds to the residual variance term of the errors ϵ_{ij} remains low. Adding the first 20 PCs to adjust for population structure decreased the mean relative bias for the variance component of the polygenic random intercept by a small amount compared to when using the first 10 PCs only (Supplementary Table C.1). Results for binary traits (Supplementary Table C.3 and Supplementary Figure C.1) are consistent to what is observed for continuous traits, that is adjusting for population structure with the first 10 PCs reduces the relative bias of variance components, and the use of a sparse GRM slightly increases the relative biases compared to using a full GRM. Compared to simulations of continuous phenotypes, relative biases for variance components are more important in simulations with binary traits.

We then simulated continuous phenotypes with 100 causal predictors explaining 2% of heritability to assess the impact on the estimation of variance components when the model is misspecified, since under our proposed model, we estimate all variance parameters under the null model of no genetic association. As can be seen in Figure 5.2, estimates of the variance parameters for the two random intercepts are upwardly biased even when using the full GRM with the first 10 PCs to account for population structure. When replacing the full GRM by a sparse one, the median relative bias for the variance of the non polygenic random intercept ψ_1 remains around -10%, which is similar to when the null model is the true model (Figure 5.1). However, we observe an increased relative bias when estimating the variance component

114

Figure 5.1: Relative bias of variance parameters estimated under the null model of no genetic association when simulating continuous phenotypes with no causal predictor. Top-left panel: Model with a full GRM and no PC to control for genetic ancestry. Top-right panel: Model with a full GRM and 10 PCs to control for genetic ancestry. Bottom-left panel: Model with a sparse GRM and no PC to control for genetic ancestry. Bottom-right panel: Model with a sparse GRM and 10 PCs to control for genetic ancestry.
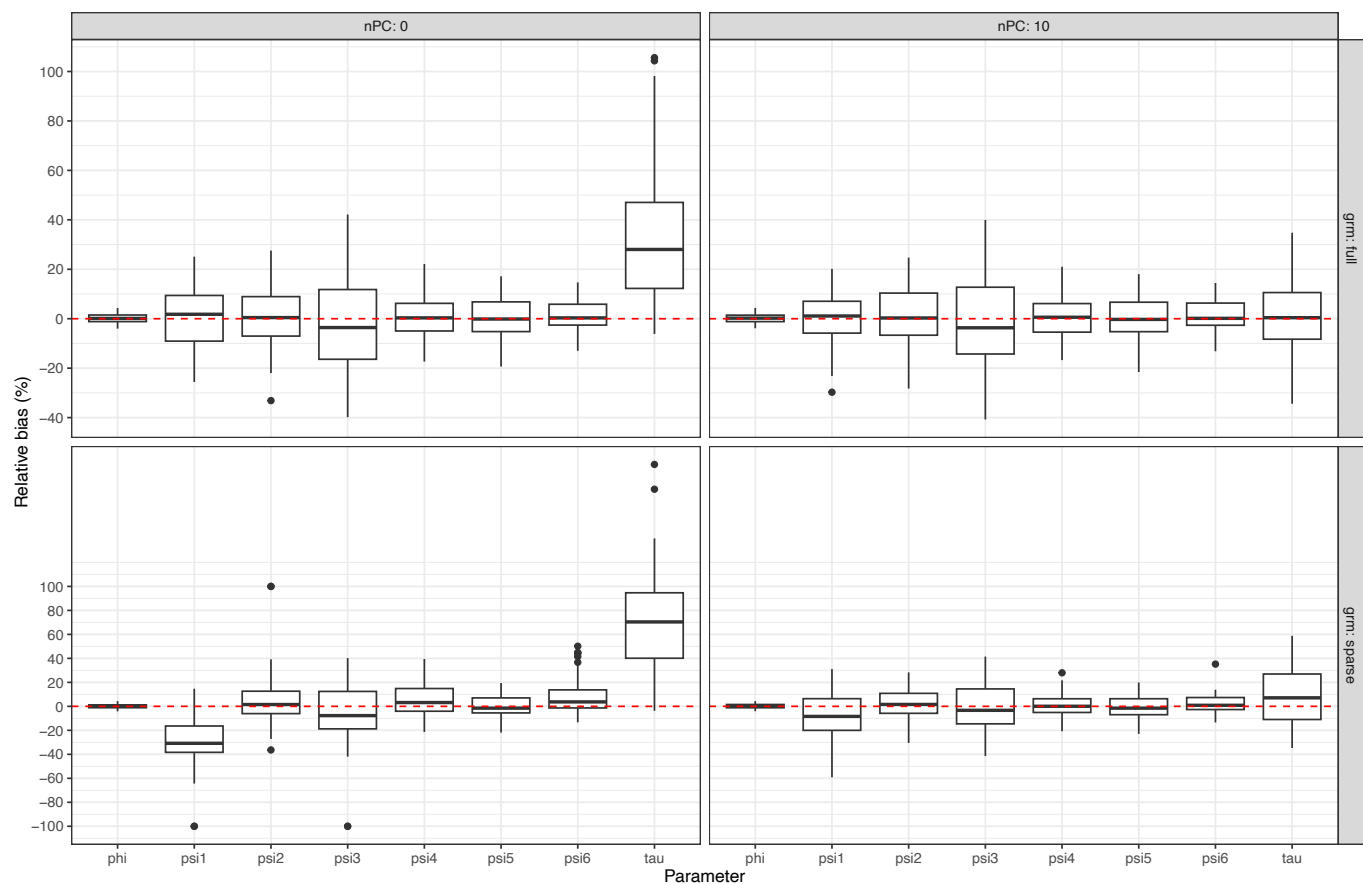
Table 5.2: Computation time in minutes to fit the AI-REML algorithm as a function of the modelling strategy for the simulation model with no causal predictor (0% heritability) and for the simulation model with 100 causal predictors explaining 2% of heritability for both continuous and binary phenotypes. We present the median value with IQR in brackets.

Number of PCs	GRM	Binary phenotypes		Continuous phenotypes	
		0%	2%	0%	2%
0	full	9.9 (7.0)	8.3 (7.0)	7.7 (0.4)	6.9 (1.0)
	sparse	2.7 (1.9)	2.5 (0.9)	1.3 (0.2)	1.4 (0.1)
10	full	7.3 (7.6)	8.3 (2.6)	7.4 (0.3)	6.7 (1.1)
	sparse	1.6 (1.0)	1.2 (1.2)	1.1 (0.2)	1.2 (0.1)
20	full	9.0 (8.1)	8.5 (4.4)	7.5 (1)	8.2 (1.2)
	sparse	1.6 (0.9)	1.7 (1.2)	1.3 (0.1)	1.7 (0.1)

of the polygenic random effect τ, with a median value lying above 30%, compared to a median relative bias of 10% when the model is not misspecified. Again, when fitting the model with either the full or sparse GRM but without any PC to control for population structure, estimates of variance parameters for the two random intercepts have large relative biases. Adding the first 20 PCs to adjust for population structure decreased the mean relative bias for the variance component of the polygenic random intercept by a small margin compared to when using the first 10 PCs only (Supplementary Table C.2). Results for binary traits (Supplementary Table C.4 and Supplementary Figure C.2) are again consistent to results for continuous traits, albeit with relative biases that are larger in magnitude.

We present in Table 5.2 the computation time in minutes to fit the AI-REML algorithm when using a full GRM versus a sparse GRM for the simulation model with no causal predictor and for the simulation model with 100 causal predictors explaining 2% of heritability. We see that although using a sparse GRM results in biased estimates of the variance parameters for the two random intercepts, the computation time is reduced by a factor of five for continuous phenotypes and by a factor of four for binary phenotypes.

Figure 5.2: Relative bias of variance parameters estimated under the null model of no genetic association when simulating continuous phenotypes with 100 causal predictors explaining 2% of heritability. Top-left panel: Model with a full GRM and no PC to control for genetic ancestry. Top-right panel: Model with a full GRM and 10 PCs to control for genetic ancestry. Bottom-left panel: Model with a sparse GRM and no PC to control for genetic ancestry. Bottom-right panel: Model with a sparse GRM and 10 PCs to control for genetic ancestry.
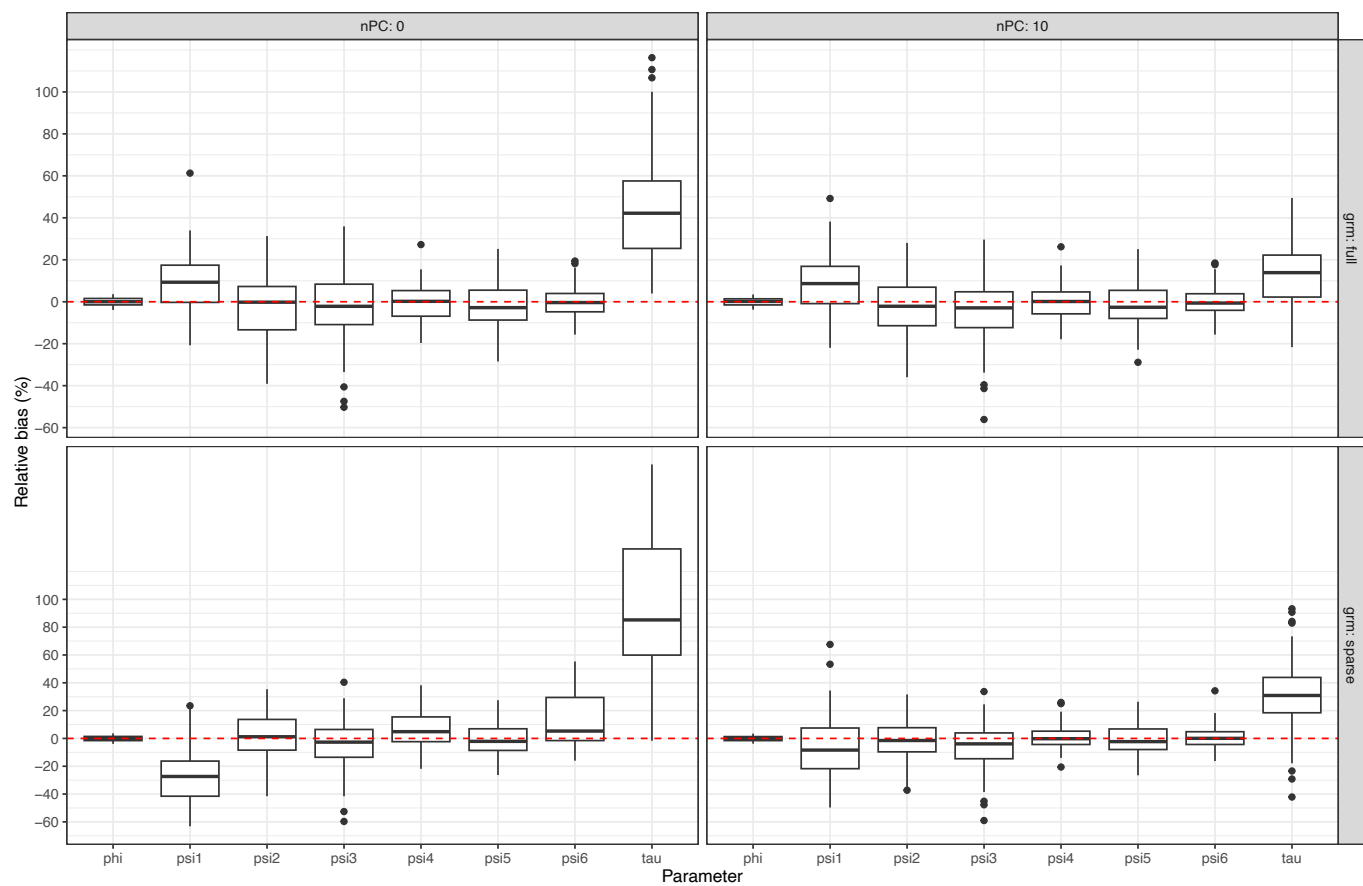
Selection of genetic predictors

We present in Figure 5.3 the Precision-Recall (PR) curve to illustrate the performance of our proposed method in retrieving causal genetic predictors as a function of the modelling strategy for the simulation model with continuous phenotypes and 100 causal predictors explaining 2% and 10% of heritability respectively. The PR curve displays the precision of a method, i.e. the proportion of selected predictors that are truly causal, as a function of the true positive rate (TPR), that is the percentage of true causal predictors that are selected by the model. Because the number of non-causal predictors is significantly greater than the number of causal predictors in genetic association studies, the PR curve is a more robust tool than the receiving operator characteristic (ROC) curve which is sensitive to the imbalance between the number of causal and non-causal predictors, since the false positive rate (FPR) is naturally weighted down due to the very large number of true negatives (Saito and Rehmsmeier [2015]). We see that including the 10 PCs as covariates to adjust for population structure in the penalized model greatly increases the ability to retrieve causal predictors. On the other hand, using a sparse GRM in place of the full GRM to adjust for relatedness in both the null and penalized models has little impact on the performance of the model, which is encouraging. Results for binary phenotypes, presented in Supplementary Figure C.3, are consistent with results obtained when simulating continuous phenotypes. Finally, median computation times to fit the lasso regularization path for our proposed penalized mixed model is presented in Table 5.3. Using a sparse GRM in place of the full GRM reduces the median computation time by at least a factor of two in all simulations.

Next we compared the performance of our method versus that of a standard lasso, using the `Julia` package `GLMNet` which wraps the `Fortran` code from the original `R` package `glmnet` (Friedman et al. [2010b]). The default implementation for `glmnet` and `pglmm` is to find the smallest value of the tuning parameter λ such that no predictors are selected in the model, and then to solve the penalized minimization problem over a grid of decreas-

Figure 5.3: Precision-recall curve for selection of genetic predictors for our proposed method as a function of the modelling strategy. The left and right panels illustrate the average performance of the method over 50 replications for the simulation model with continuous phenotypes and 100 causal predictors explaining 2% and 10% of heritability respectively.
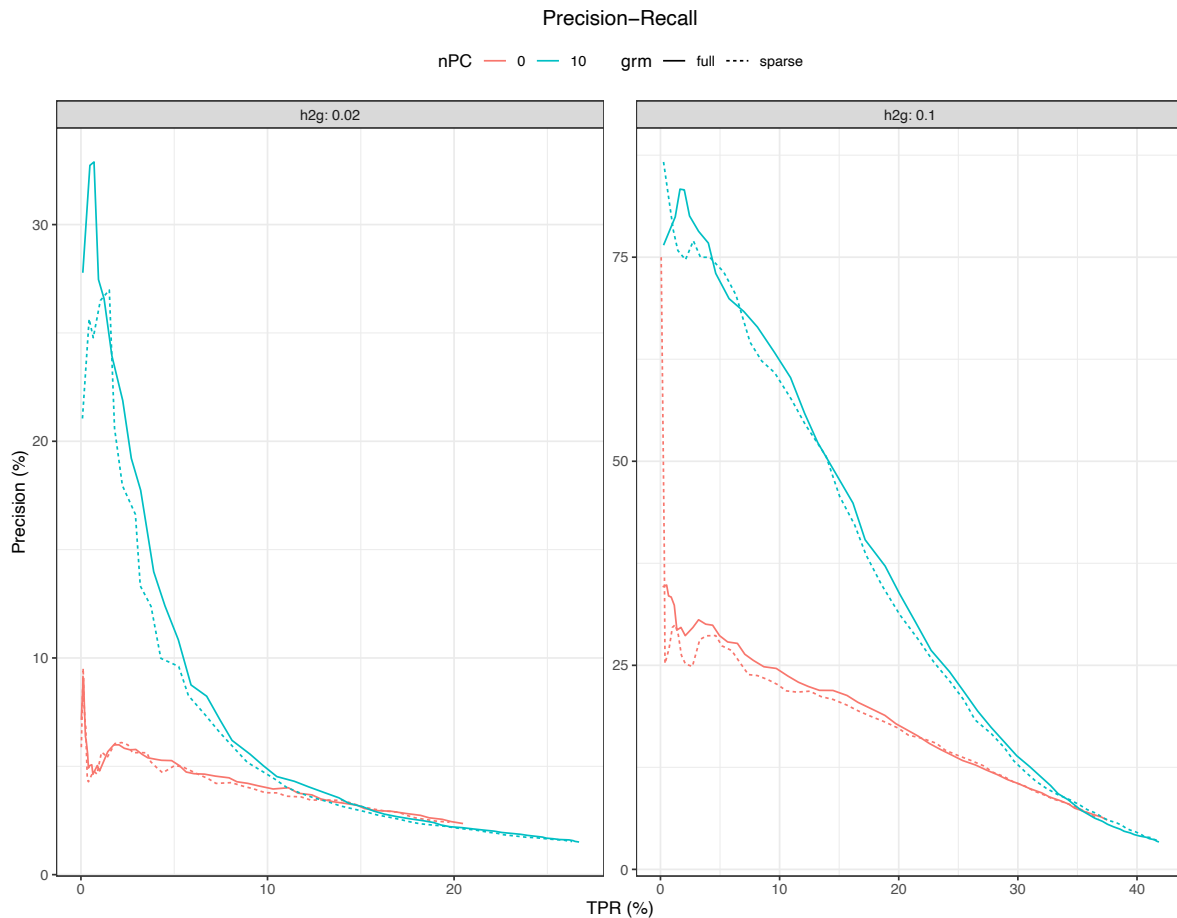

ing values of λ. For these two methods, we used a grid of 100 values of λ on the log10 scale with $\lambda_{min} = 0.01\lambda_{max}$, where λ_{max} is chosen such that no predictors are selected in the model. We also compared the performance of our method with that of GMMAT, using the freely available R package. Variable selection for GMMAT was performed by varying the cutoff value p_α such that all predictors with a p-value smaller than p_α were retained in the model. For GMMAT and our proposed method, we use a sparse GRM to account for relatedness between individuals. For all three methods, population structure was accounted for by adding the first 10 PCs as additional covariates. As can be seen on Figure 5.4, our proposed method's ability to retrieve causal predictors is uniformly superior to that of both

Table 5.3: Computation time in minutes to fit the lasso regularization path for each modelling strategy for the simulation model with no causal predictor (0% heritability) and for the simulation model with 100 causal predictors explaining 2% of heritability for both continuous and binary phenotypes. We present the median value with IQR in brackets.

Number of PCs	GRM	Binary phenotypes		Continuous phenotypes	
		0%	2%	0%	2%
0	full	82.3 (17.6)	88.0 (19.8)	60.8 (18.6)	66.8 (26.4)
	sparse	34.2 (10.5)	35.1 (6.9)	25.0 (15.2)	24.2 (9.5)
10	full	96.0 (14.8)	97.2 (15.3)	113 (29.4)	81.5 (19.7)
	sparse	45.2 (8.6)	40.0 (8.5)	57.5 (16.3)	35.1 (8.1)

glmnet and GMMAT. Results for binary phenotypes, presented in Supplementary Figure C.4, are again consistent with those obtained when simulating continuous phenotypes.

Figure 5.4: Precision-recall curve for selection of genetic predictors for the three compared methods. The left and right panels illustrate the average performance with 95% confidence interval of the methods over 50 replications for the simulation model with continuous phenotypes and 100 causal predictors explaining 2% and 10% of heritability respectively.
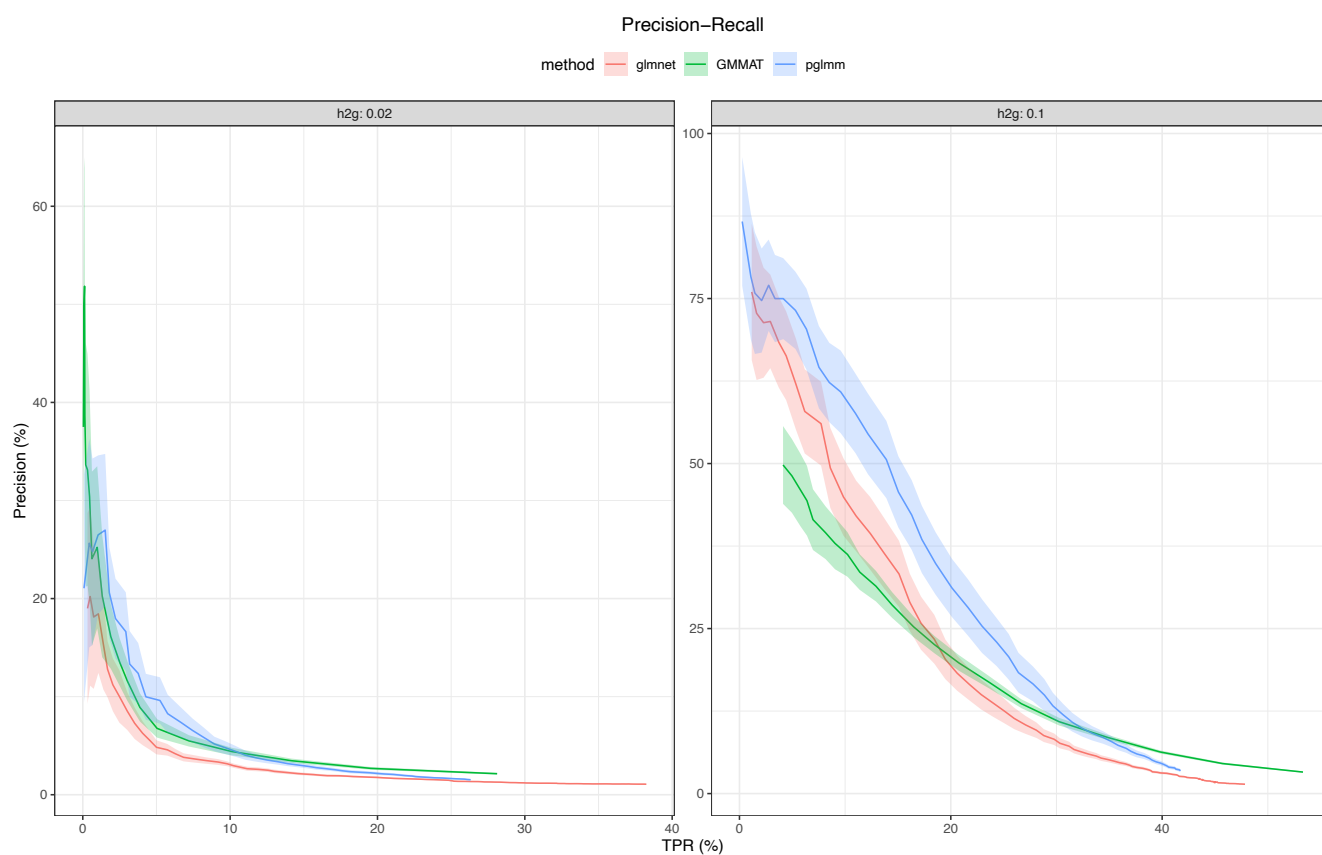
5.4   Identification of genetic predictors for emotional and behavioral difficulties in childhood and adolescence

Emotional and behavioral difficulties in childhood and adolescence are associated with significant impairment in various domains of functioning and are a major public health concern (Ogundele [2018]). The etiology of such problems is complex and involves both genetic and environmental determinants that are likely to interact and change over the course of development. Understanding the genetic and environmental risk factors underlying these difficulties is crucial for the development of effective prevention and intervention strategies. However, most of the research in genetic epidemiology is conducted in samples of adults, and then generalized to children and adolescences. This is mainly due to the fact that samples to conduct GWAS in children and adolescent are of relatively small size, which is not adapted to the standard GWAS approaches.

The objective of this study is to identify genetic predictors for emotional and behavioral difficulties in childhood and adolescents from the the Quebec Longitudinal Study of Child Development (QLSCD) and the Quebec Newborn Twin Study (QNTS). The QLSCD was designed to examine the long-term associations of preschool physical, cognitive, social, and emotional development with biopsychosocial development across childhood, adolescence, and young adulthood (Orri et al. [2021]). Data have been collected annually or every 2 years from children born in 1997/1998 in the Canadian province of Quebec and followed up from ages 5 months to 25 years. Data were collected, in either English or French, by trained interviewers from the person most knowledgeable about the child (mother in $> 98\%$ of the cases). The QNTS is a population-based cohort initially including 1324 twins (i.e., 662 pairs) born between April 1st, 1995 and December 31st, 1998 in the greater Montreal area of the province of Québec, Canada (Boivin et al. [2012]). Zygosity was initially assessed via questionnaire and confirmed with DNA tests on a subsample of $n = 123$ same-sex pairs (96% correspondence; (Forget-Dubois et al. [2003])).

5.4.1 Outcomes and covariates

When children were 6, 7, 8, 9, 10, and 12 years of age, school teachers rated their social and emotional behavior in the past 6 months using validated questionnaires. For this study, we focused on three externalizing behaviors: aggression, hyperactivity and opposition. Hyperactivity, aggression and opposition externalizing behaviors scores were rated on a 3-point Likert scale (0=never/1=sometimes/2=often) using respectively six, ten and four validated items from the Social Behavior Questionnaire (Collet et al. [2022]), and then averaged at each age (range 0-2, higher scores indicating higher propensity to display such behavior). Children's behavior was assessed by a different teacher each year, thus reducing risks of rater bias. Socio-economic status was derived using household income, education level and prestige of the profession of both parents, and then standardized with zero mean and unit standard deviation (SD) (Institut de la statistique du Québec, Direction des enquêtes longitudinales et sociales [2016]). We dichotomized the variable (Low vs High) using 0 as the cutoff value.

5.4.2 Genotype data

Genome-wide genotype data was available from $n = 721$ participants in the QLSCD cohort and $n = 641$ participants in the QNTS cohort. Biological samples from the QNTS participants were genotyped in two batches, the first in 2016 and the second in 2019. QLSCD participants were genotyped in 2016. For all batches, Illumina PsychArrayv1.X was used with assembly b37. Data were exported in FWD/REV format using Illumia's GenomeStudio software. Quality control (QC) was conducted with PLINK v1.90b6.20 and R v4.0, and all steps are detailed in Appendix C.3. For the calculation of ancestry components used to determine genetic outliers and as covariates in the analyses, pre-imputation genotype data were used. Genetic outliers were defined as individuals with a distance from the mean of > 4 SD in the first eight multidimensional scaling (MDS) ancestry components. Additional variant filtering steps for calculation of ancestry components were the removal of variants

with a MAF < 0.05 or Hardy-Weinberg equilibrium (HWE) exact test p-value $< 10^{-3}$ in sets of unrelated individuals, removal of variants mapping to the extended MHC region (chromosome 6, 25-35 Mbp) or to a typical inversion site on chromosome 8 (7-13 Mbp), and linkage disequilibrium (LD) pruning. Next, the pairwise identity-by-state (IBS) matrix of all individuals was calculated using filtered genotype data. MDS analysis was performed on the IBS matrix using the eigendecomposition-based algorithm in PLINK v1.90b6.7 (QLSCD) and PLINK v1.90b5.2 (QNTS). Imputation was conducted using SHAPEIT v2 (r837) (Delaneau et al. [2013]), IMPUTE2 v2.3.2 (Howie et al. [2009]), and the 1000 Genomes Phase 3 reference panel. After imputation, variants with a MAF $< 1\%$, an HWE exact test p-value $< 1 \times 10^{-6}$, and an INFO metric < 0.8 were removed. Finally, variants that were not SNPs or that were strand-ambiguous were removed, leaving about four millions imputed SNPs for the analysis. Imputed SNPs were converted into PLINK binary format.

5.4.3 Methods

Imputation of missing data

Socio-economic status was missing for 3 (0.42%) participants from the QLSCD cohort and 110 (17%) participants from the QNTS cohort. Thus, we imputed the missing covariate using the average in each cohort independently. A significant number of children whose genotype data were available had missing data for all time points (1% of participants in the QLSCD cohort; 12% of the participants in the QNTS cohort). In order to avoid discarding them from GWAS which may reduce the power to find any significant SNPs, we imputed their externalizing scores using the average score stratified by cohort, sex, binary socio-economic status and age. For children who were lost to follow-up, externalizing scores were imputed using the last observation carried forward (LOCF) procedure. In the following analyses, we compare the results between the complete case analysis (CCA, N = 1217) and the single imputation (SI, N = 1357) approach.

GWAS

We ran a GWAS using the GMMAT package in R using age, age^2, cohort, binary socio-economic status, sex and the first 10 PCs as fixed-effects covariates in the model. To account for the correlations between repeated measurements, we added a random intercept and a random slope for age. To account for the genetic correlation between individuals, we used a sparse GRM that was calculated using the PC-Relate method (Conomos et al. [2016]). The GWAS analysis was performed on just over 4 millions imputed SNPs that passed QC.

Prediction model

We used our proposed lasso penalized mixed model for selecting important predictors and for building a prediction model for the externalizing behaviors. We used the same fixed-effects covariates in the model as for the GWAS. To account for the correlations between repeated measurements, we again added a random intercept and a random slope for age, and we used a sparse GRM to account for the genetic correlation between individuals. We fitted the model removing the last observation for each participant (train set), and used the last observation for each participant (test set) to assess the accuracy of the predicted externalizing scores. We compared the performance of the lasso penalized mixed model with an adaptive lasso mixed model, using the `bigstatsr` package in R to fit an elastic-net penalty model on the training data. More specifically, we used the estimated predictors coefficients from the elastic-net model to define the adaptive lasso weights to be used in our proposed penalized mixed model, such that for the jth predictor, the weight was defined as $\hat{w}_j = |\hat{\beta}_j^{enet}|^{-0.25}$.

To assess the calibration and predictive performance of the prediction models, we report a mean squared prediction error (MSPE)-based definition of the R^2 coefficient of determina-

tion (Staerk et al. [2024]) defined as

$$R^2_{MSPE} = 1 - \frac{\sum_{i=1}^{m}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{m}(y_i - \bar{y})^2},$$

where $\hat{y}_i$ is the predicted score for the ith participant, for $i = 1, ..., m$, and $\bar{y}$ is the mean score in the test set. We note that if R^2_{MSPE} is negative, this means that a simple intercept model on the test data performs better than the polygenic model in predicting externalizing scores. Compared to the squared correlation between predicted and observed values which only focuses on the discriminative performance of the model, R^2_{MSPE} takes both discrimination and calibration of the prediction model into account. To reduce the number of predictors to incorporate in the penalized mixed models, we merged the list of imputed SNPs that passed QC ($\sim$ 4 millions SNPs) with the list of SNPs from the third phase of the International HapMap Project (The International HapMap Consortium [2003], Pain [2023]), containing a total of 1.2 millions of SNPs that are generally well imputed in most studies. After merging both list of SNPs, the final analysis included a total of 735K SNPs to be included in our penalized regressions mixed models.

5.4.4 Results

Characteristics of the QLSCD and QNTS cohorts participants are presented in Table 5.4. For both cohorts, externalizing scores are higher on average in males compared to females. Aggression mean scores are similar between the two cohorts, as opposed to hyperactivity and opposition mean score values that are higher in the QNTS cohort. We present in Table 5.5 the list of SNPs with a p-value smaller than 5×10^{-8} for either the CCA or SI model, which is generally considered as the genome-wide significance threshold (Pe'er et al. [2008]). P-values were obtained by fitting the GMMAT model on all imputed SNPs, using a score test statistics that was computed using variance components estimated under the null model of no genetic association. When two significant SNPs were in LD, defined as a squared correlation

based on genotypic allele counts greater than 0.2, only the most significant was retained. A total of 9 SNPs were significant in the CCA model for the hyperactivity externalizing score, but none were significant in the SI model. Similarly, a total of 9 SNPs were significant in the CCA model for the aggression externalizing score, but none were significant in the SI model. Finally, 2 SNPs were significant in the GWAS of the opposition externalizing score for both analyses. We note that there was no overlap between the SNPs that were found to be significant among the three externalizing scores, which may be surprising given that the correlation coefficient between the three scores ranges from 0.57 to 0.70.

The predictive performance of the lasso and adaptive lasso penalized mixed models as a function of the number of selected genetic predictors is presented in Supplementary Figures C.5 and C.6 respectively. For the lasso penalized mixed model, the best predictions for both imputation methods are obtained for the opposition externalizing score (CCA $R^2_{MSPE} = 0.310$ and SI $R^2_{MSPE} = 0.441$), followed by hyperactivity (CCA $R^2_{MSPE} = 0.304$ and SI $R^2_{MSPE} = 0.420$) and aggression (CCA $R^2_{MSPE} = 0.261$ and SI $R^2_{MSPE} = 0.403$) as presented in Table 5.6. The number of selected genetic predictors in all models is relatively high, ranging from 601 to 851 for the CCA, and from 965 to 1163 for the models using SI to impute missing values.

For the adaptive lasso penalized mixed model, the best predictions for both imputation methods are again obtained for the opposition externalizing score (CCA $R^2_{MSPE} = 0.294$ and SI $R^2_{MSPE} = 0.439$), followed by hyperactivity (CCA $R^2_{MSPE} = 0.301$ and SI $R^2_{MSPE} = 0.411$) and aggression (CCA $R^2_{MSPE} = 0.235$ and SI $R^2_{MSPE} = 0.376$). Thus, the proportion of variance explained by the estimated polygenic score is slightly lower in the adaptive lasso compared to the lasso regularized prediction model. On the the other hand, the number of selected genetic predictors in all models is comparatively lower in the adaptive lasso mixed model, ranging from 419 to 628 for the CCA, and from 744 to 966 for the models using SI to replace missing values.

The set of overlapping SNPs, i.e. SNPs that were selected by the penalized mixed model for all three externalizing scores included 3 SNPs in the complete case analysis and 16 SNPs in the single imputation analysis as presented in Supplementary Table C.6. A total of 2 SNPs in the complete case analysis and 17 SNPs in the single imputation analysis were selected by the adaptive penalized mixed model for all three externalizing scores as presented in Supplementary Table C.7. We report for each SNP the gene where they are located when applicable, and if they are intronic, missense or non coding variants.

Finally, we examined the impact of imputing missing data on the variance component estimates under our proposed model. As shown in Supplementary Table C.5, the estimates for the polygenic variance component ($\hat{\tau}$) were consistent in both the CCA and the SI analysis. However, imputing missing data led to lower estimates for the residual variance parameter ($\hat{\phi}$) compared to CCA, while the estimates for the variance of the non-polygenic random intercept (ϕ_1) were higher in the SI analysis. This result aligns with the imputation method we used, which involved imputing missing externalizing scores using stratified sample averages, thereby reducing variability between individuals. The increase in within-individual variability after imputing missing data suggests that differences between individuals play a significant role relative to within-individual fluctuations in explaining trajectories of externalizing behaviors.

Table 5.4: Characteristics of the Quebec Longitudinal Study of Child Development (QLSCD) and the Quebec Newborn Twin Study (QNTS) participants.

	Cohort					
	QLSCD			QNTS		
	Females	Males	Total	Females	Males	Total
n	398 (55%)	323 (45%)	721	319 (50%)	317 (50%)	636
Socio-economic status (%)						
High	206 (52%)	165 (51%)	371 (51%)	133 (42%)	139 (44%)	272 (43%)
Aggression score, Mean (SD)						
6 years old	0.16 (0.28)	0.33 (0.45)	0.23 (0.37)	0.18 (0.36)	0.38 (0.54)	0.28 (0.47)
7 years old	0.15 (0.30)	0.34 (0.40)	0.23 (0.36)	0.11 (0.29)	0.43 (0.54)	0.27 (0.46)
8-9 years old	0.14 (0.29)	0.31 (0.40)	0.22 (0.35)	0.10 (0.30)	0.37 (0.54)	0.23 (0.45)
10 years old	0.11 (0.23)	0.32 (0.45)	0.20 (0.36)	0.11 (0.31)	0.35 (0.50)	0.23 (0.43)
12 years old	0.09 (0.20)	0.26 (0.39)	0.16 (0.31)	0.02 (0.10)	0.25 (0.42)	0.13 (0.32)
Hyperactivity score, Mean (SD)						
6 years old	0.27 (0.39)	0.56 (0.56)	0.39 (0.49)	0.41 (0.50)	0.70 (0.65)	0.55 (0.60)
7 years old	0.28 (0.40)	0.57 (0.56)	0.41 (0.50)	0.36 (0.48)	0.65 (0.61)	0.51 (0.57)
8 years old	0.28 (0.40)	0.53 (0.52)	0.39 (0.47)	0.57 (0.44)	0.67 (0.44)	0.62 (0.44)
9 years old	-	-	-	0.30 (0.46)	0.60 (0.62)	0.45 (0.56)
10 years old	0.17 (0.34)	0.50 (0.54)	0.32 (0.47)	0.25 (0.41)	0.56 (0.55)	0.41 (0.51)
12 years old	0.12 (0.24)	0.42 (0.48)	0.25 (0.40)	0.21 (0.38)	0.46 (0.52)	0.33 (0.47)
Opposition score, Mean (SD)						
6 years old	0.19 (0.33)	0.38 (0.49)	0.27 (0.42)	0.27 (0.45)	0.45 (0.56)	0.36 (0.51)
7 years old	0.17 (0.34)	0.40 (0.51)	0.27 (0.44)	0.21 (0.38)	0.46 (0.57)	0.34 (0.50)
8-9 years old	0.20 (0.40)	0.41 (0.49)	0.29 (0.45)	0.23 (0.44)	0.44 (0.56)	0.33 (0.51)
10 years old	0.18 (0.36)	0.41 (0.55)	0.28 (0.47)	0.22 (0.41)	0.43 (0.53)	0.33 (0.49)
12 years old	0.15 (0.33)	0.40 (0.51)	0.26 (0.44)	0.16 (0.28)	0.41 (0.52)	0.28 (0.43)

Table 5.5: Top SNPs identified from GWAS of three externalizing scores in the combined QLSCD and QNTS cohortsf.

Ext. score	SNP	Chr	Position	A1 / A2	MAF	Gene : Consequence	N CCA	N SI	p-value CCA	p-value SI	$\hat{\beta}$ CCA	$\hat{\beta}$ SI
Hyp	rs144702376	5	119325132	T / C	0.014	None	1191	1329	4.28×10^{-8}	1.06×10^{-5}	2.8×10^{-4}	$3.9 \times 10^{-}$
	rs115921653	6	10089651	A / G	0.017	OFCC1 : Intronic	1210	1348	2.38×10^{-8}	9.62×10^{-6}	2.8×10^{-4}	$3.9 \times 10^{-}$
	rs149436805	6	97566274	T / C	0.015	KLHL32 : Intronic	1211	1350	2.90×10^{-8}	1.20×10^{-4}	1.9×10^{-4}	$3.9 \times 10^{-}$
	rs77406243	8	89633141	A / G	0.010	LOC105375630 : Intronic	1211	1350	4.09×10^{-9}	-	2.4×10^{-4}	$3.9 \times 10^{-}$
	rs80133571	12	108969583	C / T	0.010	None	1204	1338	3.78×10^{-8}	1.52×10^{-5}	2.6×10^{-4}	$3.9 \times 10^{-}$
	rs112641948	15	76967976	T / C	0.016	SCAPER : Intronic	1208	1347	4.98×10^{-8}	5.43×10^{-7}	2.6×10^{-4}	$3.9 \times 10^{-}$
	rs8053367	16	53875484	T / G	0.468	FTO : Intronic	1216	1356	4.62×10^{-8}	4.53×10^{-4}	2.5×10^{-4}	$3.9 \times 10^{-}$
	rs17231282	18	67692722	T / C	0.018	RTTN : Intronic	1194	1331	3.01×10^{-8}	1.76×10^{-3}	2.5×10^{-4}	$3.9 \times 10^{-}$
	rs12971894	19	29985790	A / G	0.123	VSTM2B-DT : Intronic	1197	1336	3.58×10^{-8}	3.28×10^{-4}	2.4×10^{-4}	$3.9 \times 10^{-}$
Aggr	rs13393674	2	158403762	T / C	0.052	ACVR1C : Intronic	1193	1330	8.69×10^{-8}	4.38×10^{-9}	2.3×10^{-4}	$8.1 \times 10^{-}$
	rs10812324	9	25933433	C / T	0.047	None	1200	1338	4.06×10^{-8}	1.08×10^{-4}	2.6×10^{-4}	$8.0 \times 10^{-}$
	rs7911165	10	131626650	C / T	0.359	LOC107984281 : NC	1216	1357	4.09×10^{-8}	2.34×10^{-5}	2.2×10^{-4}	$7.8 \times 10^{-}$
	rs113352461	12	33390548	G / A	0.010	None	1199	1337	8.56×10^{-9}	7.92×10^{-5}	2.3×10^{-4}	$7.9 \times 10^{-}$
	rs202096	13	30820199	A / G	0.350	KATNAL1 : Intronic	1206	1346	1.19×10^{-8}	1.42×10^{-5}	2.1×10^{-4}	$8.0 \times 10^{-}$
	rs144975159	15	52666187	T / G	0.012	MYO5A : Intronic	1214	1355	4.26×10^{-8}	2.59×10^{-6}	2.3×10^{-4}	$7.9 \times 10^{-}$
	rs4792882	17	43869277	G / T	0.018	CRHR1 : Intronic	1210	1347	1.09×10^{-8}	2.08×10^{-6}	2.3×10^{-4}	$8.1 \times 10^{-}$
	rs187443053	18	67161236	A / G	0.012	DOK6 : Intronic	1212	1353	3.29×10^{-8}	3.35×10^{-6}	2.5×10^{-4}	$8.0 \times 10^{-}$
	rs77702389	21	32712872	T / G	0.049	TIAM1 : Intronic	1216	1357	2.97×10^{-12}	6.48×10^{-7}	2.2×10^{-4}	$8.0 \times 10^{-}$
Opp	rs116541675	2	18419626	G / A	0.011	None	1213	1354	3.42×10^{-8}	2.00×10^{-9}	2.9×10^{-4}	$6.7 \times 10^{-}$
	rs145226968	3	153021039	C / T	0.011	None	1195	1333	2.87×10^{-7}	1.74×10^{-8}	3.0×10^{-4}	$6.8 \times 10^{-}$

MAF = Minor allele frequency. N = Number of non-missing observations. CCA = Complete case analysis. SI = Single imputation.
Hyp = Hyperactivity. Aggr = Aggression. Opp = Opposition.
NC = Non-coding.
Note 1: MAF was reported for the effect allele A1.
Note 2: p-values were obtained using the GMMAT score test. Effect size estimates were obtained using the GMMAT Wald test.

Table 5.6: Performance of the best prediction models on the test set for each externalizing score, imputation model and regularization procedure.

Externalizing score	Analysis	Lasso mixed model Number of selected genetic predictors	MSPE	R^2_{MSPE}	Adaptive lasso mixed model Number of selected genetic predictors	MSPE	R^2_{MSPE}
Aggression	Complete case analysis	601	0.083	0.261	419	0.086	0.235
	Single imputation	965	0.066	0.403	744	0.069	0.376
Hyperactivity	Complete case analysis	771	0.147	0.304	533	0.147	0.301
	Single imputation	1080	0.119	0.420	905	0.121	0.411
Opposition	Complete case analysis	851	0.135	0.310	628	0.138	0.294
	Single imputation	1163	0.109	0.441	966	0.109	0.439

MSPE = Mean squared prediction error.

5.5 Discussion

We proposed a methodology for fitting penalized longitudinal mixed models with more than a single random effect to account for random individual effects not attributable to the genetic similarity between individuals. Our proposed model is based on regularized PQL estimation, which does not require making any assumption about the distribution of the outcome, but only the mean-variance relationship. We studied the performance of the AIREML algorithm when simulating population structure and subjects relatedness for continuous and binary traits. We showed that using PC-AiR to calculate PCs that account for genetic correlations due to distant common ancestry and PC-Relate to estimate kinship due to the sharing of more recent ancestors was effective in controlling the relative bias of variance components estimates under the null model of no genetic association. In addition, we showed that the use of a sparse GRM greatly reduced the computational burden of estimating the variance components and fitting the penalized mixed model, while having little impact on the performance of the lasso penalized mixed model in retrieving important predictors. In simulation studies for both continuous and binary longitudinal traits using real genotype data, we demonstrated that our proposed model achieved better precision (lower FDR) than an univariate association test (GMMAT) and that of a lasso penalized model without any random effect.

We further showed that omitting to add the top PCs as covariates in the model to adjust for population stratification led to an increase in the relative bias of variance components and in the false discovery rate (FDR) of the lasso penalized mixed model. This is due to PC-Relate only measuring the genetic relatedness due to alleles shared identically by descent (IBD) from recent common ancestors, because genetic relatedness due to more distant ancestry have been adjusted for by previously regressing the genotype values on the top PCs. Using a different approach to estimate the kinship coefficients between individuals, such as the REAP (Thornton et al. [2012]) and RelateAdmix (Moltke and Albrechtsen [2013]) methods may circumvent the need to add the top PCs as covariates in the model to adjust

for population structure (Chen et al. [2020]). However, the advantage of PC-Relate compared to the aforementioned methods is that it does not require model-based estimates of individual ancestry and population-specific allele frequencies nor using external reference population panels.

Albeit consistent estimation of variance components in mixed models is often overlooked in genetic association studies, we studied the impact of different modelling strategies on the relative bias of the estimates, and showed that increased relative biases were associated with an increase of the FDRs in the penalized model. Since the total phenotypic variance is usually expressed as the sum of polygenic and residual variances, when the polygenic variance is overestimated, the model fails to capture the correct extent of the genetic influences on the phenotype. Furthermore, the variance attributable to the error terms or to within-individual fluctuations in longitudinal studies will be underestimated, leading to potentially inflated type I error rates. In our approach, as we do not test for individual significance of the predictors, this would be translated in observing an increase of the FDR, as was shown in the simulation studies. Additionally, we showed in St-Pierre et al. [2023] that the inverse polygenic variance component τ^{-1} can be seen as a ridge regularization parameter for penalization of the individual polygenic random intercepts. Thus, overestimating the polygenic variance component τ results in overfitting of the contribution of the PCs obtained from the spectral decomposition of the GRM in explaining the observed phenotypic variance. In other words, it means that the model is overestimating the heterogeneity of the individual random intercepts. From a bias-variance tradeoff point of view, overestimation of the polygenic variance component results in increasing the variability of the random effects estimation, which will be reflected in higher error rates in independent data sets.

In a real case study for identifying important genetic predictors of aggression, hyperactivity and opposition externalizing behaviors in childhood and adolescents from the QLSCD and QNTS cohorts, we fitted the GMMAT model and our proposed penalized mixed model and

showed that the two methods identified different sets of potentially important SNPs. We performed an analysis based on single imputation method to handle missing data due to non-response or loss to follow-up, and compared the results with a complete case analysis. We found two SNPs (rs116541675, rs145226968) that were genome-wide significant (p-value $< 5 \times 10^{-8}$) for the opposition externalizing behaviour in both the single imputation and complete case analyses. For the hyperactivity and aggression behaviours, we found nine mutually exclusive SNPs that were significant in the complete case analysis only. We further demonstrated the utility of our proposed methodology in predicting externalizing behaviours scores in children from the combined QLSCD and QNTS cohorts, and showed using an MSPE-based definitions of the R^2 coefficient of determination that the obtained predictions were well-calibrated, and that genetic effects improved the accuracy of the predicted scores compared to a model without any genetic contribution. In addition, we fitted an adaptive lasso penalized mixed model with weights inversely proportional to effect sizes estimates obtained via an elastic-net regularized regression. We found that the adaptive lasso mixed model resulted in sparser models than the lasso mixed model while the predicted scores accuracy was comparable.

In this study, we have not considered gene-environment interaction (GEI) effects. Conducting genome-wide GEI analyses would contribute to identify important genetic variants whose effects are modified by environmental factors, and could improve the accuracy of a prediction model as predictors of externalizing behaviours are likely heterogeneous among different environmental exposures, such as socio-economic status. To our knowledge, there exists no GEI univariate association test that allows to account for both genetic similarity and correlation between repeated measurements within an individual for both continuous and binary traits. Further work is needed to assess the computational efficiency and evaluate the estimation and selection accuracy of a penalized hierarchichal variable selection method based on regularized PQL for longitudinal outcomes. An interesting idea to explore is the use of adaptive weights combined with a sparse group lasso penalty in order to perform

hierarchical selection of main genetic and GEI effects in longitudinal studies (Mendez-Civieta et al. [2020]).

One limitation of the proposed methodology is that PQL estimation results in biased estimators of both regression coefficients and variance components in GLMMs (Jang and Lim [2009]). It would be interesting to explore if first-order and second-order correction procedures that were proposed in the literature to correct for this bias (Breslow and Lin [1995], Lin and Breslow [1996]) would increase the performance of the penalized mixed model in retrieving important predictors when between and within individuals correlations are important. 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) (Fan and Li [2001]) and Minimax Concave Penalty (MCP) (Zhang [2010]) could be explored. Another limitation of our proposed method is that we cannot directly analyse imputed SNPs in the dosage format as we rely on the `SnpArrays` package (Zhou et al. [2019b]) developed in the `julia` programming language which provides computationally efficient routines for reading and manipulating compressed storage of biallelic SNP data. Finally, when pairwise correlations between SNPs in blocks of LD are important, it is known that the lasso has a tendency to only select one variant among a group of correlated variants. In low-dimensional settings, Freijeiro-González et al. [2021] showed through a simulation study that the adaptive lasso with weights based on elastic-net regression estimates performed well to retrieve causal predictors in different dependence structures. This is the approach we followed in the real case study. Another direction to explore would be replacing the lasso regularization by an elastic-net penalty in our proposed penalized longitudinal mixed model.

References

Michel Boivin, Mara Brendgen, Ginette Dionne, Lise Dubois, Daniel Pérusse, Philippe Robaey, Richard E. Tremblay, and Frank Vitaro. The Quebec Newborn Twin Study Into Adolescence: 15 Years Later. *Twin Research and Human Genetics*, 16(1):64–69, December 2012. ISSN 1839-2628. 10.1017/thg.2012.129. URL http://dx.doi.org/10.1017/thg.2012.129.

Norman E. Breslow and Xihong Lin. Bias correction in generalised linear mixed models with a single component of dispersion. *Biometrika*, 82(1):81–91, March 1995. ISSN 1464-3510. 10.1093/biomet/82.1.81. URL http://dx.doi.org/10.1093/biomet/82.1.81.

Han Chen, Chaolong Wang, Matthew P. Conomos, Adrienne M. Stilp, Zilin Li, Tamar Sofer, Adam A. Szpiro, Wei Chen, John M. Brehm, Juan C. Celedón, Susan Redline, George J. Papanicolaou, Timothy A. Thornton, Cathy C. Laurie, Kenneth Rice, and Xihong Lin. Control for population structure and relatedness for binary traits in genetic association studies via logistic mixed models. *The American Journal of Human Genetics*, 98(4):653–666, April 2016. ISSN 0002-9297. 10.1016/j.ajhg.2016.02.012. URL http://dx.doi.org/10.1016/j.ajhg.2016.02.012.

Yuning Chen, Gina M. Peloso, Ching-Ti Liu, Anita L. DeStefano, and Josée Dupuis. Evaluation of population stratification adjustment using genome-wide or exonic variants. *Genetic Epidemiology*, 44(7):702–716, June 2020. ISSN 1098-2272. 10.1002/gepi.22332. URL http://dx.doi.org/10.1002/gepi.22332.

Shing Wan Choi, Timothy Shin-Heng Mak, and Paul F. O'Reilly. Tutorial: a guide to performing polygenic risk score analyses. *Nature Protocols*, 15(9):2759–2772, July 2020. 10.1038/s41596-020-0353-1. URL https://doi.org/10.1038/s41596-020-0353-1.

Benjamin B Chu, Kevin L Keys, Christopher A German, Hua Zhou, Jin J Zhou, Eric M Sobel, Janet S Sinsheimer, and Kenneth Lange. Iterative hard thresholding in genome-wide association studies: Generalized linear models, prior weights, and double sparsity. *GigaScience*, 9(6), June 2020. ISSN 2047-217X. 10.1093/gigascience/giaa044. URL http://dx.doi.org/10.1093/gigascience/giaa044.

Ophélie A. Collet, Massimiliano Orri, Richard E. Tremblay, Michel Boivin, and Sylvana M. Côté. Psychometric properties of the Social Behavior Questionnaire (SBQ) in a longitudinal population-based sample. *International Journal of Behavioral Development*, 47(2):180–189, August 2022. ISSN 1464-0651. 10.1177/01650254221113472. URL http://dx.doi.org/10.1177/01650254221113472.

Matthew P. Conomos, Michael B. Miller, and Timothy A. Thornton. Robust inference of population structure for ancestry prediction and correction of stratification in the presence of relatedness. *Genetic Epidemiology*, 39(4):276–293, March 2015. ISSN 1098-2272. 10.1002/gepi.21896. URL http://dx.doi.org/10.1002/gepi.21896.

Matthew P. Conomos, Alexander P. Reiner, Bruce S. Weir, and Timothy A. Thornton. Model-free estimation of recent genetic relatedness. *The American Journal of Human Genetics*, 98(1):127–148, January 2016. ISSN 0002-9297. 10.1016/j.ajhg.2015.11.022. URL http://dx.doi.org/10.1016/j.ajhg.2015.11.022.

Olivier Delaneau, Jean-Francois Zagury, and Jonathan Marchini. Improved whole-chromosome phasing for disease and population genetic studies. *Nature Methods*, 10(1):5–6, January 2013. ISSN 1548-7105. 10.1038/nmeth.2307. URL http://dx.doi.org/10.1038/nmeth.2307.

Frank Dudbridge. Power and predictive accuracy of polygenic risk scores. *PLoS Genetics*, 9 (3):e1003348, March 2013. ISSN 1553-7404. 10.1371/journal.pgen.1003348. URL http://dx.doi.org/10.1371/journal.pgen.1003348.

Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, December 2001. ISSN 1537-274X. 10.1198/016214501753382273. URL http://dx.doi.org/10.1198/016214501753382273.

Nadine Forget-Dubois, Daniel Pérusse, Gustavo Turecki, Alain Girard, Jean-Michel Billette, Guy Rouleau, Michel Boivin, Jocelyn Malo, and Richard E. Tremblay. Diagnosing zygosity in infant twins: Physical similarity, genotyping, and chorionicity. *Twin Research*, 6(6): 479–485, December 2003. ISSN 1369-0523. 10.1375/136905203322686464. URL http://dx.doi.org/10.1375/136905203322686464.

Alberto Forte, Massimiliano Orri, Cédric Galera, Maurizio Pompili, Gustavo Turecki, Michel Boivin, Richard E. Tremblay, and Sylvana M. Côté. Developmental trajectories of childhood symptoms of hyperactivity/inattention and suicidal behavior during adolescence. *European Child & Adolescent Psychiatry*, 29(2):145–151, April 2019. ISSN 1435-165X. 10.1007/s00787-019-01338-0. URL http://dx.doi.org/10.1007/s00787-019-01338-0.

Laura Freijeiro-González, Manuel Febrero-Bande, and Wenceslao González-Manteiga. A critical review of lasso and its derivatives for variable selection under dependence among covariates. *International Statistical Review*, 90(1):118–145, August 2021. ISSN 1751-5823. 10.1111/insr.12469. URL http://dx.doi.org/10.1111/insr.12469.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010. URL https://www.jstatsoft.org/v33/i01/.

Arthur R. Gilmour, Robin Thompson, and Brian R. Cullis. Average information REML: An

efficient algorithm for variance parameter estimation in linear mixed models. *Biometrics*, 51(4):1440, December 1995. 10.2307/2533274. URL https://doi.org/10.2307/2533274.

Institut de la statistique du Québec, Direction des enquêtes longitudinales et sociales. Étude longitudinale du développement des enfants du Québec – (ELDEQ 1998-2015), sep 2016. https://www.jesuisjeserai.stat.gouv.qc.ca/informations_chercheurs/documentation_technique/E18_Variables_Derivees_A.pdf.

Woncheol Jang and Johan Lim. A numerical study of pql estimation biases in generalized linear mixed models under heterogeneity of random effects. *Communications in Statistics - Simulation and Computation*, 38(4):692–702, February 2009. ISSN 1532-4141. 10.1080/03610910802627055. URL http://dx.doi.org/10.1080/03610910802627055.

Longda Jiang, Zhili Zheng, Ting Qi, Kathryn E Kemper, Naomi R Wray, Peter M Visscher, and Jian Yang. A resource-efficient tool for mixed model association analysis of large-scale data. *Nature genetics*, 51(12):1749–1755, 2019.

Zan Koenig, Mary T. Yohannes, Lethukuthula L. Nkambule, Julia K. Goodrich, Heesu Ally Kim, Xuefang Zhao, Michael W. Wilson, Grace Tiao, Stephanie P. Hao, Nareh Sahakian, Katherine R. Chao, Michael E. Talkowski, Mark J. Daly, Harrison Brand, Konrad J. Karczewski, Elizabeth G. Atkinson, and Alicia R. Martin and. A harmonized public resource of deeply sequenced diverse human genomes. January 2023. 10.1101/2023.01.23.525248. URL https://doi.org/10.1101/2023.01.23.525248.

Alexandre Lemyre, Natalia Poliakova, Frank Vitaro, Richard E. Tremblay, Michel Boivin, and Richard E. Bélanger. Does shyness interact with peer group affiliation in predicting substance use in adolescence? *Psychology of Addictive Behaviors*, 32(1):132–139, February 2018. ISSN 0893-164X. 10.1037/adb0000328. URL http://dx.doi.org/10.1037/adb0000328.

Jiahan Li, Kiranmoy Das, Guifang Fu, Runze Li, and Rongling Wu. The Bayesian lasso for

genome-wide association studies. *Bioinformatics*, 27(4):516–523, December 2010. 10.1093/ bioinformatics/btq688. URL https://doi.org/10.1093/bioinformatics/btq688.

Xihong Lin and Norman E. Breslow. Bias correction in generalized linear mixed models with multiple components of dispersion. *Journal of the American Statistical Association*, 91(435):1007–1016, September 1996. ISSN 1537-274X. 10.1080/01621459.1996.10476971. URL http://dx.doi.org/10.1080/01621459.1996.10476971.

Ani Manichaikul, Josyf C. Mychaleckyj, Stephen S. Rich, Kathy Daly, Michèle Sale, and Wei-Min Chen. Robust relationship inference in genome-wide association studies. *Bioinformatics*, 26(22):2867–2873, October 2010. ISSN 1367-4803. 10.1093/bioinformatics/btq559. URL http://dx.doi.org/10.1093/bioinformatics/btq559.

Alvaro Mendez-Civieta, M. Carmen Aguilera-Morillo, and Rosa E. Lillo. Adaptive sparse group lasso in quantile regression. *Advances in Data Analysis and Classification*, 15(3): 547–573, July 2020. ISSN 1862-5355. 10.1007/s11634-020-00413-8. URL http://dx.doi .org/10.1007/s11634-020-00413-8.

Ida Moltke and Anders Albrechtsen. RelateAdmix: a software tool for estimating relatedness between admixed individuals. *Bioinformatics*, 30(7):1027–1028, November 2013. ISSN 1367-4803. 10.1093/bioinformatics/btt652. URL http://dx.doi.org/10.1093/ bioinformatics/btt652.

Marie C. Navarro, Massimiliano Orri, Daniel Nagin, Richard E. Tremblay, Sînziana I. Oncioiu, Marilyn N. Ahun, Maria Melchior, Judith van der Waerden, Cédric Galéra, and Sylvana M. Côté. Adolescent internalizing symptoms: The importance of multi-informant assessments in childhood. *Journal of Affective Disorders*, 266:702–709, April 2020. ISSN 0165-0327. 10.1016/j.jad.2020.01.106. URL http://dx.doi.org/10.1016/ j.jad.2020.01.106.

Michael O Ogundele. Behavioural and emotional disorders in childhood: A brief overview

for paediatricians. *World Journal of Clinical Pediatrics*, 7(1):9–26, February 2018. ISSN 2219-2808. 10.5409/wjcp.v7.i1.9. URL http://dx.doi.org/10.5409/wjcp.v7.i1.9.

Sînziana I. Oncioiu, Massimiliano Orri, Michel Boivin, Marie-Claude Geoffroy, Louise Arseneault, Mara Brendgen, Frank Vitaro, Marie C. Navarro, Cédric Galéra, Richard E. Tremblay, and Sylvana M. Côté. Early Childhood Factors Associated With Peer Victimization Trajectories From 6 to 17 Years of Age. *Pediatrics*, 145(5), May 2020. ISSN 1098-4275. 10.1542/peds.2019-2654. URL http://dx.doi.org/10.1542/peds.2019-2654.

Massimiliano Orri, Cedric Galera, Gustavo Turecki, Michel Boivin, Richard E. Tremblay, Marie-Claude Geoffroy, and Sylvana M. Côté. Pathways of Association Between Childhood Irritability and Adolescent Suicidality. *Journal of the American Academy of Child & Adolescent Psychiatry*, 58(1):99–107.e3, January 2019. ISSN 0890-8567. 10.1016/j.jaac.2018.06.034. URL http://dx.doi.org/10.1016/j.jaac.2018.06.034.

Massimiliano Orri, Michel Boivin, Chelsea Chen, Marilyn N Ahun, Marie-Claude Geoffroy, Isabelle Ouellet-Morin, Richard E Tremblay, and Sylvana M Côté. Cohort profile: Quebec Longitudinal Study of Child Development (QLSCD). *Soc. Psychiatry Psychiatr. Epidemiol.*, 56(5):883–894, May 2021.

Oliver Pain. HapMap3 SNP-list, 2023. https://zenodo.org/record/7773502.

Itsik Pe'er, Roman Yelensky, David Altshuler, and Mark J. Daly. Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genetic Epidemiology*, 32(4):381–385, March 2008. ISSN 1098-2272. 10.1002/gepi.20303. URL http://dx.doi.org/10.1002/gepi.20303.

Alkes L. Price, Noah A. Zaitlen, David Reich, and Nick Patterson. New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics*, 11 (7):459–463, June 2010. 10.1038/nrg2813. URL https://doi.org/10.1038/nrg2813.

Takaya Saito and Marc Rehmsmeier. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, 10(3): e0118432, March 2015. ISSN 1932-6203. 10.1371/journal.pone.0118432. URL http://dx.doi.org/10.1371/journal.pone.0118432.

Julien St-Pierre, Karim Oualkacha, and Sahir Rai Bhatnagar. Efficient penalized generalized linear mixed models for variable selection and genetic risk prediction in high-dimensional data. *Bioinformatics*, 39(2), January 2023. ISSN 1367-4811. 10.1093/bioinformatics/btad063. URL http://dx.doi.org/10.1093/bioinformatics/btad063.

Christian Staerk, Hannah Klinkhammer, Tobias Wistuba, Carlo Maj, and Andreas Mayr. Generalizability of polygenic prediction models: how is the r^2 defined on test data? *BMC Medical Genomics*, 17(1), May 2024. ISSN 1755-8794. 10.1186/s12920-024-01905-8. URL http://dx.doi.org/10.1186/s12920-024-01905-8.

The International HapMap Consortium. The International HapMap Project. *Nature*, 426 (6968):789–796, December 2003. ISSN 1476-4687. 10.1038/nature02168. URL http://dx.doi.org/10.1038/nature02168.

Timothy Thornton, Hua Tang, Thomas J. Hoffmann, Heather M. Ochs-Balcom, Bette J. Caan, and Neil Risch. Estimating kinship in admixed populations. *The American Journal of Human Genetics*, 91(1):122–138, July 2012. ISSN 0002-9297. 10.1016/j.ajhg.2012.05.024. URL http://dx.doi.org/10.1016/j.ajhg.2012.05.024.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996. ISSN 00359246. URL http://www.jstor.org/stable/2346178.

Pol A. C. van Lier, Frank Vitaro, Edward D. Barker, Mara Brendgen, Richard E. Tremblay, and Michel Boivin. Peer victimization, poor academic achievement, and the link between childhood externalizing and internalizing problems. *Child Development*, 83(5):1775–1788,

June 2012. ISSN 1467-8624. 10.1111/j.1467-8624.2012.01802.x. URL http://dx.doi .org/10.1111/j.1467-8624.2012.01802.x.

Tong Tong Wu, Yi Fang Chen, Trevor Hastie, Eric Sobel, and Kenneth Lange. Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics*, 25(6):714–721, March 2009.

Jian Yang, S. Hong Lee, Michael E. Goddard, and Peter M. Visscher. GCTA: A tool for genome-wide complex trait analysis. *The American Journal of Human Genetics*, 88(1): 76–82, January 2011. ISSN 0002-9297. 10.1016/j.ajhg.2010.11.011. URL http://dx.doi .org/10.1016/j.ajhg.2010.11.011.

Jianming Yu, Gael Pressoir, William H Briggs, Irie Vroh Bi, Masanori Yamasaki, John F Doebley, Michael D McMullen, Brandon S Gaut, Dahlia M Nielsen, James B Holland, Stephen Kresovich, and Edward S Buckler. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics*, 38(2):203–208, December 2005. 10.1038/ng1702. URL https://doi.org/10.1038/ng1702.

Magdalena A. Zdebik, Michel Boivin, Marco Battaglia, Richard E. Tremblay, Bruno Falissard, and Sylvana M. Côté. Childhood multi-trajectories of shyness, anxiety and depression: Associations with adolescent internalizing problems. *Journal of Applied Developmental Psychology*, 64:101050, July 2019. ISSN 0193-3973. 10.1016/j.appdev.2019.101050. URL http://dx.doi.org/10.1016/j.appdev.2019.101050.

Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894 – 942, 2010. 10.1214/09-AOS729. URL https://doi.org/ 10.1214/09-AOS729.

Hua Zhou, Mary E. Sehl, Janet S. Sinsheimer, and Kenneth Lange. Association screening of common and rare genetic variants by penalized regression. *Bioinformatics*, 26(19):

2375–2382, August 2010. ISSN 1367-4803. 10.1093/bioinformatics/btq448. URL http://dx.doi.org/10.1093/bioinformatics/btq448.

Hua Zhou, Liuyi Hu, Jin Zhou, and Kenneth Lange. MM algorithms for variance components models. *Journal of Computational and Graphical Statistics*, 28(2):350–361, March 2019a. 10.1080/10618600.2018.1529601. URL https://doi.org/10.1080/10618600.2018.1529601.

Hua Zhou, Janet S. Sinsheimer, Douglas M. Bates, Benjamin B. Chu, Christopher A. German, Sarah S. Ji, Kevin L. Keys, Juhyun Kim, Seyoon Ko, Gordon D. Mosher, Jeanette C. Papp, Eric M. Sobel, Jing Zhai, Jin J. Zhou, and Kenneth Lange. Open-Mendel: a cooperative programming project for statistical genetics. *Human Genetics*, 139(1):61–71, March 2019b. ISSN 1432-1203. 10.1007/s00439-019-02001-z. URL http://dx.doi.org/10.1007/s00439-019-02001-z.

Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, December 2006. ISSN 1537-274X. 10.1198/016214506000000735. URL http://dx.doi.org/10.1198/016214506000000735.